Machine learning reveals limited contribution of trans-only encoded variants to the HLA-DQ immunopeptidome by accurate and comprehensive HLA-DQ antigen presentation prediction

Jonas Birkelund Nilsson[1]*, Saghar Kaabinejadian[2,5]*, Hooman Yari[5], Bjoern Peters[4], Carolina Barra[1], Loren Gragert[3], William Hildebrand[5] and Morten Nielsen[1]

* shared first authorship

[1] Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark

[2] Pure MHC, LLC., Oklahoma City, OK, United States

[3] Department of Pathology and Laboratory Medicine, Tulane University School of Medicine, New Orleans, LA 70112, USA.

[4] Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, La Jolla, CA 92037, California, USA

[5] Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

Abstract

HLA class II antigen presentation is key for controlling and triggering T cell immune responses. Characterizing this event is hence of critical importance for our understanding of immune system reactions. HLA-DQ is believed to play a major role in autoimmune diseases, although so far limited progress has been made for predicting HLA-DQ antigen presentation. HLA-DQ molecules are heterodimers that can be formed as both cis and trans variants depending on whether the two chains are encoded on the same (cis) or opposite (trans) chromosomes. Still, the contribution of trans-only variants (i.e. variants not observed as cis) in shaping the HLA-DQ immunopeptidome remains largely unresolved. Here, we seek to address these issues by integrating state-of-the-art immunoinformatics data mining models with large volumes of high quality HLA-DQ specific MS-immunopeptidomics data. The analysis demonstrated a highly improved predictive power and molecular coverage for models trained including these novel HLA-DQ data. More importantly, investigating the role of trans-only HLA-DQ variants revealed a limited to no contribution to the overall HLA-DQ immunopeptidome. In conclusion, this study has furthered our understanding of HLA-DQ specificities, and has for the first time cast light on the relative role of cis versus trans-only HLA-DQ variants in the HLA class II antigen presentation space. The developed method, NetMHCIIpanDQ, is available at https://services.healthtech.dtu.dk/services/NetMHCIIpanDQ-1.0.

Introduction

Major histocompatibility complex class II molecules (MHC class II) are expressed on the surface of professional antigen presenting cells such as B cells, dendritic cells (DCs), and monocytes/macrophages (Rocha and Neefjes, 2008). These molecules, which are designed to bind and present fragments of the exogenous proteins to T-helper cells, are heterodimers consisting of α and β chains which together form the peptide-binding cleft.

In humans, HLA (human leukocyte antigen) class II is encoded by three different loci (HLA-DR, -DQ, and -DP). These HLA genes have numerous allelic variants with polymorphisms that are mainly clustered around the peptide-binding groove, resulting in a wide range of distinct peptide-binding specificities (Reynisson et al., 2020a). In many autoimmune diseases, HLA class II genes are major genetic susceptibility factors (Arango et al., 2017; Rocha and Neefjes, 2008) that play a central role in the pathogenesis of these conditions by presenting antigenic peptides to CD4+ T cells.

Several studies have explored the importance of HLA-DR and DQ at haplotype and genotype levels among type 1 diabetes (T1D) patients (Arango et al., 2017). These genetic and functional studies have indicated that both HLA-DR and DQ alleles are associated with the risk of T1D (Arango et al., 2017; Erlich et al., 2008). In addition, the associated DR-DQ haplotypes demonstrate a risk hierarchy, ranging from highly predisposing to highly protective (Erlich et al., 2008). Interestingly, more recently it was demonstrated that HLA-DR, which generally plays the primary role in autoimmune diseases, has an important but secondary role to the HLA-DQ locus in T1D (Hu et al., 2015).

Autoimmune disorders like T1D in addition to other conditions such as Celiac disease, where a direct and exceptionally strong association for HLA-DQ has been established (Stepniak et al., 2008), thus necessitate a more thorough and systematic characterization of antigen presentation by HLA-DQ molecules to enable study of their function. So far, peptide binding motifs of only a few HLA-DQ molecules have been exhaustively studied (Bergseng et al., 2014; Sidney et al., 2010; Vartdal et al., 1996). One reason for this is that HLA-DQ molecules are more complex to study experimentally. For instance, because of the monomorphic nature of the α-chain in HLA-DR, the polymorphic variations are only provided by the β-chain (Tollefsen et al., 2012). In HLA-DQ, both α- and β-chains contribute to polymorphic variations. However, evidence suggests that not every α- and β-chain pairing will result in a stable heterodimer due to key structural requirements on the α and β dimerization interface (Kwok et al., 1993; Tollefsen et al., 2012). For example, the DQA1*02, 03, 04, 05, and 06 alleles form stable heterodimers only with the DQB1*02, 03, and 04 alleles. Likewise, DQA1*01 has only been detected to form stable heterodimers with DQB1*05 and 06 (Creary et al., 2021; Kwok et al., 1993; Petersdorf et al., 2022).

In addition, studying the function of HLA-DQ alleles is challenging because of the extensive linkage disequilibrium between HLA-DR and HLA-DQ within the HLA class II region, making it difficult to differentiate the role of individual HLA-DQ alleles from the associated HLA-DR molecules (Arango et al., 2017; Tollefsen et al., 2012).

Finally, unique *cis* and *trans* encoded DQ molecules can occur where α- and β-chains that pair to form the heterodimer are encoded by the same (*cis*) or different (*trans*) chromosomes making the study of these molecules even further complicated. While the majority of the current knowledge on HLA-DQ molecules comes from cis encoded variants, the surface expression and function of a small number of trans encoded DQ variants have been confirmed (Lundin et al., 1990; Tollefsen et al., 2012). Here, it is important to emphasize that these functional trans molecules have also been observed to be functional as the corresponding cis-encoded variant. Therefore, it is generally believed that alleles of DQα- and DQβ-chains pair up primarily in cis rather than in trans variants (Kwok and Nepom, 1991; McFarland and Beeson, 2002). Hereafter, we refer to any DQα- and β-chain pairing other than what was mentioned above as “trans-only”, which includes all combinations that have not been detected or reported as cis encoded.

In recent years, the information related to *cis*-encoded HLA-DQ variants has been greatly expanded due to large volumes of HLA sequence data becoming available (Creary et al., 2021). Here, the assumption is that all observed DQ haplotypes, by natural selection, are able to form stable and functional cis-encoded molecules. However, the contribution of trans-only encoded variants in antigen presentation and their role compared to cis-variants in shaping and complementing the HLA-DQ immunopeptidome has remained largely unresolved.

Given the critical role of HLA class II antigen presentation in the control and shaping of the adaptive immune response, great efforts have been dedicated to the development of prediction models capable of predicting this event (reviewed in Nielsen et al., 2020). Current state-of-the-art prediction methods include NetMHCIIpan (Reynisson et al., 2020b), a pan-specific method allowing for prediction of antigen presentation for any HLA class II molecule with known protein sequence. For HLA-DQ and DP heterodimers, this means that sequence information about both the α and β chains is required in order to make predictions.

Originally, in vitro peptide-HLA binding affinity (BA) assays have been used to generate data to characterize the motifs of HLA class II molecules (Reynisson et al., 2020a), and development of different machine-learning prediction models to identify the rules of peptide–HLA binding (Gfeller and Bassani-Sternberg, 2018; Nielsen et al., 2010). However, experimental results indicate binding affinity (BA) to be a relatively weak correlate of antigen processing and presentation by HLA molecules (Bassani-Sternberg et al., 2016). In addition, multiple studies have demonstrated that the performance of the HLA-class II peptide-binding prediction models improve significantly when trained with immunopeptidome data acquired by liquid chromatography coupled with mass spectrometry (LC-MS/MS) (Alvarez et al., 2018; Gfeller and Bassani-Sternberg, 2018; Kaabinejadian et al., 2022; Reynisson et al., 2020a). Generally, in an HLA class II immunopeptidome eluted ligand (EL) assay, HLA molecules are affinity purified from lysed antigen presenting cells (APCs) using HLA specific monoclonal antibodies. The HLA molecules are next denatured and peptide ligands are isolated and sequenced via LC-MS/MS (Caron et al., 2015; Purcell et al., 2019). The result of such an assay is a list of peptide sequences restricted to at least one of the HLA class II molecules expressed by the interrogated cell line. EL data has a major advantage over BA data as they contain signals from different steps of HLA class II antigen presentation, such as antigen digestion, HLA loading of ligands, and transport to the cell surface (Barra et al., 2018; Paul et al., 2018; Racle et al., 2019).

HLA class II binding predictions have been widely used to identify epitope candidates in infectious, cancer and autoimmune diseases (Wang et al., 2010). The majority of prediction algorithms for HLA class II have so far been focused on HLA-DR molecules due to the large data availability for those. However, in the context of HLA-DQ, both pairing of synthetic alpha and beta chains in order to perform binding affinity experiments, and the use of MS-immunopeptidomics have proven to be challenging. The latter mostly due to a lack of HLA-DQ specific antibodies resulting in limited yield in the immunoprecipitation step.

In recent years, proteomics and peptide analysis by mass spectrometry (MS) has made huge progress, due to cutting edge technology and increased sensitivity of the instruments along with advanced software platforms and algorithms that support peptide identification and quantification. These advancements, along with the use of a highly specific HLA-DQ antibody, have enabled us to characterize, in a single assay, thousands of peptides which naturally bind the HLA-DQ molecules and generate stable peptide-HLA complexes that are transported to the cell surface to be presented to immune cells. Here, we have applied this setup to generate a large set of peptides presented by a set of HLA-DQ molecules frequent in the worldwide population from a panel of homozygous B lymphoblastoid cell lines. These large data sets were directly submitted to bioinformatic motif identification and machine learning pipelines to define the motifs and uncover the rules governing the processing and presentation of peptides in a biological context. Further, this study allowed us to move towards resolving the challenge of cis versus trans formation of functional HLA-DQ heterodimers and determine the role of trans-only variants in shaping the HLA-DQ immunopeptidome. The extensive insight into the peptide-binding characteristics of the investigated HLA-DQ molecules provided by this study will facilitate better understanding of HLA-DQ disease association and discovery of novel therapeutic targets.

Materials and methods

Cell lines and antibody

Homozygous B lymphoblastoid cell lines (BLCL) were obtained from the International Histocompatibility Working Group (IHWG) Cell and DNA bank housed at the Fred Hutchinson Cancer Research Center, Seattle, WA (http://www.ihwg.org). A group of 16 cell lines expressing the high frequency HLA-DQ alleles were selected for the study (supplementary table 1). To guarantee intact class II processing and presentation machinery and to ensure that the total HLA-DQ expression represents the physiological level, use of engineered cells was avoided.

The cells were grown in high density cultures in roller bottles in complete RPMI medium (Gibco) supplemented with 15% fetal bovine serum (FBS; Gibco/Invitrogen Corp) and 1% 100 mM sodium pyruvate (Gibco). Cells were harvested from the suspension, washed with PBS and spun down at 4C for 10 minutes. The cell pellets were immediately frozen in LN2 and stored at -80 until downstream processing (Kaabinejadian et al., 2022). All cell lines were subjected to high-resolution HLA typing (HLA-A, -B, -C, DRB1,3, 4, 5, DP and DQ) immediately upon receipt and growth in our laboratory, for authentication prior to large scale culture and data collection.

The anti-human HLA-DQ specific monoclonal antibody was produced in house from a hybridoma cell line (clone SPVL3) and used for affinity purification of total HLA DQ from the BLCLs.

Isolation and purification of HLA-DQ bound peptides

HLA-DQ molecules were purified from the cells by affinity chromatography using the anti-human HLA-DQ specific antibody (clone SPVL3) coupled to CNBr-activated Sepharose 4 Fast Flow (Amersham Pharmacia Biotech, Orsay, France) as described previously (Kaabinejadian et al., 2022). Briefly, frozen cell pellets were pulverized using Retsch Mixer Mill MM400, resuspended in lysis buffer comprised of Tris pH 8.0 (50 mM), Igepal, 0.5%, NaCl (150 mM) and complete protease inhibitor cocktail (Roche, Mannheim, Germany). Lysates were centrifuged in an Optima XPN-80 ultracentrifuge (Beckman Coulter, IN, USA) and filtered supernatants were loaded on immunoaffinity columns overnight at 4C. Columns were washed sequentially with a series of wash buffers (Purcell et al., 2019) and were eluted with 0.2 N acetic acid. The HLA was denatured, and the peptides were isolated by adding glacial acetic acid and heat. The mixture of peptides and HLA-DQ was subjected to reverse phase high performance liquid chromatography (RP-HPLC).

MS

Fractionation of the HLA/Peptide Mixture by RP-HPLC

RP-HPLC was used to reduce the complexity of the peptide mixture eluted from the affinity column. First, the eluate was dried under vacuum using a CentriVap concentrator (Labconco, Kansas City, Missouri, USA). The solid residue was dissolved in 10% acetic acid and fractionated using a Paradigm MG4 instrument (Michrom BioResources, Auburn, California, USA). An acetonitrile (ACN) gradient was run at pH 2 using a two-solvent system. Solvent A contained 2% ACN in water, and solvent B contained 5% water in ACN. Both solvent A and Solvent B contained 0.1% trifluoroacetic acid (TFA). The column was pre-equilibrated at 2% solvent B. Then the sample was loaded at a flow rate of 120 µl/min and a two-segment gradient was run at 160 µl/min flow rate as described in detail in (Kaabinejadian et al., 2022). Fractions were collected in 2 min intervals using a Gilson FC 203B fraction collector (Gilson, Middleton, Wisconsin, USA), and the ultra-violet (UV) absorption profile of the eluate was recorded at 215 nm wavelength.

Nano LC-MS/MS Analysis

Peptide-containing HPLC fractions were dried and resuspended in a solvent composed of 10% acetic acid, 2% ACN and iRT peptides (Biognosys, Schlieren, Switzerland) as internal standards. Fractions were applied individually to an Eksigent nanoLC 415 nanoscale RP-HPLC (AB Sciex, Framingham, Massachusetts, USA), including a 5-mm long, 350 µm internal diameter Chrom XP C18 trap column with 3 µm particles and 120 Å pores, and a 15-cm-long ChromXP C18 separation column (75 µm internal diameter) packed with the same medium (AB Sciex, Framingham, Massachusetts, USA). An ACN gradient was run at pH 2.5 using a two-

solvent system. Solvent A was 0.1% formic acid in water, and solvent B was 0.1% formic acid in 95% ACN in water. The column was pre-equilibrated at 2% solvent B. Samples were loaded at 5 µL/min flow rate onto the trap column and run through the separation column at 300 nL/min with two linear gradients: 10% to 40% B for 70 minutes, followed by 40% to 80% B for 7 minutes.

The column effluent was ionized using the nanospray III ion source of an AB Sciex TripleTOF 5600 quadruple time-of-flight mass spectrometer (AB Sciex, Framingham, MA, USA) with the source voltage set to 2,400 V. Information-dependent analysis (IDA) method was used for data acquisition (Kaabinejadian et al., 2022). PeakView Software version 1.2.0.3 (AB Sciex, Framingham, MA, USA) was used for data visualization.

Peptide Data Analysis

Peptide sequences were identified using PEAKS Studio 10.5 software (Bioinformatics Solutions, Waterloo, Canada). A database composed of SwissProt Homo sapiens (taxon identifier 9606) and iRT peptide sequences was used as the reference for database search. Variable post-translational modifications (PTM) including acetylation, deamination, pyroglutamate formation, oxidation, sodium adducts, phosphorylation, and cysteinylation were included in database search. Identified peptides were further filtered at a false discovery rate (FDR) of 1% using PEAKS decoy-fusion algorithm.

Immunopeptidome Data

The immunopeptidome data consist of MS-eluted ligand (EL) and binding affinity (BA) data from the earlier NetMHCIIpan-4.1 combined with the EL data generated specifically for this study (see above). The novel MS-immunopeptidome data set covers 14 different HLA-DQ molecules obtained from 16 homozygous BLCLs summarized in table 1 below. The data was filtered as described earlier (Kaabinejadian et al., 2022) to exclude potential HLA class I and other co-immunoprecipitated contaminants, resulting in a list of peptides of length 12-21.

IHW ID	DQA1	DQB1	Raw	Final
721	DQA1*01:01	DQB1*05:01	1433	1157
9004	DQA1*01:01	DQB1*05:01	4230	2386
9013	DQA1*01:02	DQB1*06:02	2561	2046
9023	DQA1*05:01	DQB1*02:01	2784	1788
9031	DQA1*03:01	DQB1*03:02	3816	2410
9043	DQA1*05:05	DQB1*03:01	2782	2268
9052	DQA1*02:01	DQB1*03:03	3600	2331
9060	DQA1*01:03	DQB1*06:03	4353	2530
9061	DQA1*01:04	DQB1*05:03	3333	2275
9063	DQA1*01:02	DQB1*06:04	4361	2935
9072	DQA1*04:01	DQB1*04:02	4291	2747
9075	DQA1*03:02	DQB1*03:03	3601	2285
9084	DQA1*01:02	DQB1*05:02	3862	2474
9090	DQA1*03:03	DQB1*03:01	6407	4672
9093	DQA1*02:01	DQB1*02:02	3218	1919
9100	DQA1*04:01	DQB1*04:02	4370	3111
Total			59002	39334

Table 1: Summary of MS data generated for this study. IHW ID gives the cell line id of each sample. DQA1,DQB1 gives the DQ alleles expressed in the given cell line. Raw gives the total number of identified peptides, and Final gives the final number of identified peptides filtered to only include peptides of length 12-21 aa excluding predicted class I binders.

The EL data were mapped to the human reference source proteome to define source protein context. Peptides with no identical reference match were excluded, resulting in ~4% of peptides being discarded. Finally, the EL data were enriched in a per sample-id manner with random natural peptides assigned as negatives. This enrichment was done by sampling peptides of 12-21 amino acids in length in a uniform manner in an amount equal to five times the number of peptides for the most prevalent length in the positive data for the given sample.

The final novel data set consists of 39,334 positive and 369,313 negative peptides covering 14 unique HLA-DQ molecules. This data set is available in supplementary table 2. Merging the novel EL data with the earlier NetMHCIIpan-4.1 data (expanded to include peptides 12 amino acids in length), the complete EL data consists of 480,845 positive and 4,910,165 negative data from 177 samples/cell lines, and the BA data consist of 129,110 data points.

For two of the NetMHCIIpan-4.1 datasets from (Abelin et al., 2019), the original DQ typing was listed to have only two DQ molecules. However, we believe this to be an erroneous HLA typing, as the cell lines should express four different DQ molecules. As such, the HLA typing for these cell lines were extended to include all four DQ haplotypes (see supplementary table 1).

The data was partitioned into 5 subsets for cross-validated method training and evaluation using the common-motif approach described earlier (Nielsen et al., 2007) merging EL and BA data ensuring that peptides sharing an identical overlap of 9 or more consecutive amino acids were placed in the same subset.

Methods

Model training

Models were trained using the NNAlign_MA machine learning framework (Alvarez et al., 2019) in a manner similar to that described earlier for NetMHCIIpan-4.0 (Reynisson et al., 2020a). That is, the complete model consists of an ensemble of 100 neural networks of two different architectures both with one hidden layer and either 40 or 60 hidden neurons, with 10 random weight initializations for each of the 5 cross-validation folds (2 architectures, 10 seeds, and 5 folds). All models were trained using back-propagation with stochastic gradient descent, for 300 epochs, without early stopping, and a constant learning rate of 0.05. Only single allele (SA) data were included in training for a burn-in period of 20 epochs. Subsequent training cycles included multi-allele (MA) data. Two main models were trained, one including the original NetMHCIIpan-4.1 data and one including the novel HLA-DQ data. Furthermore, an additional model was trained with the novel data using peptide context encoding. Here, context is defined as a stretch of six amino acids on either side of the peptide within the source protein sequence, which are concatenated into a 12-mer stretch of amino acids.

Performance evaluation and MHC restriction deconvolution

HLA annotation for MA datasets is based on which HLA molecule expressed in a given cell line has the highest prediction score for a given ligand. To balance the differences in the prediction score distributions between HLAs, percentile normalized prediction scores for each were generated by ranking against a distribution of prediction scores of random natural peptides as described earlier (Nielsen and Andreatta, 2016; Reynisson et al., 2020b).

Performance was evaluated on the concatenated cross-validation test set predictions using three separate metrics, namely AUC (Area Under the ROC Curve), AUC 0.1 (Area Under the ROC Curve integrated up to a False Positive Rate of 10%) and Positive Predictive Value (PPV). Here, PPV was calculated as the fraction of true positives in the top N predictions, where N is the number of ligands assigned to a given HLA molecule. Each metric was calculated in a per-HLA manner from the “raw” prediction scores after HLA annotation. For the per-HLA performance evaluation, only HLA molecules with at least 10 positive peptides in both models were included in the performance evaluation, to ensure a level of certainty in the calculated performance metrics.

Statistical tests

Statistical tests in the performance evaluation were performed using one-tailed binomial tests excluding ties, with a significance level of 0.05. The alternative hypothesis in these tests is thus

that the model trained with the novel data is more likely to perform better on a given HLA molecule than the model trained without.

Consistency correlation matrix analysis

In order to assess the novel DQ data's impact on NNAlign_MA's motif deconvolution, a consistency correlation matrix analysis was performed. To avoid potential MS co-immunoprecipitated contaminant peptides biasing this analysis, the union of identified trash peptides (i.e. positive peptides given a percentile rank greater than 20 in either of the two models) were removed. A position-specific scoring matrix (PSSM) was next generated for each molecule in each cell line based on the predicted peptide binding cores. Here, a minimum of 20 positive peptides was required in order for a PSSM to be generated. Then, for each pair of cell lines sharing a given molecule, the Pearson Correlation Coefficient (PCC) between the molecule's PSSMs was calculated. The mean consistency value for a given molecule was then given as the average PCC over each unique cell line pair (excluding self-correlations). This metric thus indicates how consistent the identified binding motifs are across different datasets for each HLA class II molecule.

Similarity distance measure

Distance between two HLA class II molecules was estimated as described earlier (Hoof et al., 2009) from the pseudo-distance of the two molecules, i.e.

$$d = 1 - \frac{s(A,B)}{\sqrt{s(A,A)s(B,B)}}$$

where s(X,Y) is the summed BLOSUM 50 similarity between the pseudo-sequences of molecule X and Y. Here, the pseudo-sequences are defined as described earlier from the set of 34 polymorphic residues within the HLA sequence concatenated into a continuous sequence, of which 15 and 19 residues derive from the α and β chain, respectively (Karosiene et al., 2013).

Estimation of prevalent stable HLA-DQ molecules

A list of HLA-DQ α and β chains forming prevalent stable HLA-DQ heterodimers were constructed by first obtaining lists of DQA1 and DQB1 alleles with annotated worldwide allele frequencies. This was done by querying the allelefrequencies.net database (Gonzalez-Galarza et al., 2011) for high resolution alleles in populations of size 100 and above. Next, worldwide allele frequencies were obtained as population size weighted averages capping the maximum population size to 1000. Finally, a list of prevalent HLA-DQ molecules was constructed by pairing all α and β combinations following the restrictions outlined in table 2, only including molecules with a combined allele frequency greater than 0.00005. This resulted in a list of 154 HLA-DQ molecules.

DQA1* - DQB1*	DQA1* - DQB1*
DQA1*01-DQB1*05	DQA1*04-DQB1*02
DQA1*01-DQB1*06	DQA1*04-DQB1*03
DQA1*02-DQB1*02	DQA1*04-DQB1*04
DQA1*02-DQB1*03	DQA1*05-DQB1*02
DQA1*02-DQB1*04	DQA1*05-DQB1*03
DQA1*03-DQB1*02	DQA1*05-DQB1*04
DQA1*03-DQB1*03	DQA1*06-DQB1*02
DQA1*03-DQB1*04	DQA1*06-DQB1*03
	DQA1*06-DQB1*04

Table 2: List of DQA1-DQB1 haplotypes extracted from (Creary et al., 2021; Petersdorf et al., 2022). Due to relatively small sample size for some populations included in study by Creary et al., this data may not reflect all haplotypes observed in entire populations. Therefore, only low resolution (2 digit) information related to the observed DQA1-DQB1 haplotypes was included here to define "observed" cis DQA1, DQB1 combinations.

Estimation of worldwide haplotype frequencies

Worldwide HLA-DQ haplotype frequencies were estimated by querying the allelefrequencies.net database (Gonzalez-Galarza et al., 2011) for high resolution DQ haplotype in populations of size 100 and above, and average across population as described above for HLA-DQ frequencies.

HLA-DQ specificity tree

An HLA-DQ specificity tree was constructed by first reducing the list of 154 prevalent HLA-DQ molecules to the set of unique pseudo-sequences among the molecules. Then, each unique pseudo-sequence was mapped to a representative HLA-DQ molecule name. By default, a DQ molecule in the list of molecules covered by the training data was used to represent a pseudo-sequence when possible. Furthermore, all 14 DQ molecules in the novel data were used to represent their given pseudo-sequences. In other cases of multiple options for a given pseudo-sequence, the most prevalent DQ molecule in terms of global allelic frequency was chosen. The specificity tree was then calculated using the MHCCluster method (Thomsen et al., 2013), and visualized using the Iroki phylogenetic tree viewer (Moore et al., 2020).

Results

For the study, immunopeptidome data from 14 different HLA-DQ molecules was obtained from 16 homozygous BLCLs cell lines using LC-MS/MS as described in materials and methods. From this data, three essential issues related to HLA-DQ were investigated and addressed namely i) the relatively low predictive power of current prediction models for DQ molecules, ii) the contribution of trans-only encoded DQ variants to the overall DQ immunopeptidome, and iii) the overall coverage of the DQ specificity space of the current experimental data and developed in-silico prediction models.

Impact of novel DQ data on predictive performance

To investigate the impact on the predictive power by integration of the novel DQ data, we employed the NNAlign_MA algorithm (Alvarez et al., 2019) which is a highly powerful machine learning method for deconvoluting MS immunopeptidomics data. Two peptide antigen presentation prediction models were trained, one including the novel DQ affinity purified data (termed w_Saghar_DQ) and one without (termed wo_Saghar_DQ. These models were then evaluated using cross-validation on a per-molecule basis within four different subsets of all the HLA class II molecules included in the data. These subsets are non-DQ molecules (NotDQ), all DQ molecules (DQ), DQ molecules present in the novel data (DQ_Saghar) and DQ molecules not present in the novel data (DQ_NotSaghar).

Figure 1 displays the result of this experiment and demonstrates that incorporation of the novel DQ data resulted in a significant performance gain for DQ ($p=0.011$ for all metrics). As expected, a large part of this overall performance gain for DQ can be attributed to the DQ molecules present in the novel data. However, a significant performance gain ($p<0.05$) in terms of AUC and AUC 0.1 was also observed for DQ molecules not present in the novel data, indicating that the novel data has boosted performance across the entire DQ space. In addition, the two models performed similarly for most non-DQ molecules, however a significant performance gain ($p=0.034$) in terms of PPV was observed for the NotDQ subset.

Even though we observe an overall improvement for DQ by inclusion of the novel data, the DQ peptide binding prediction performance remains lower compared to that of non-DQ molecules. Here one must first note that the reported performance for the DQ_Saghar molecules in figure 1 is estimated by combining peptide binding data across the entire test data set where the HLA restriction is predicted to match the given molecule. This means that the reported performance is calculated from a mix of multiple data sets, including the NetMHCIIpan-4.1 training data. To quantify the effect of this, we evaluated the performance of the DQ_Saghar molecules limited to the novel data only. The result of this is shown in figure 2, and demonstrates that when focusing only on the novel data, the performance of DQ reaches a level comparable to that of non-DQ (median AUC values for DQ_Saghar and NotDQ are 0.9303 and 0.9312, respectively). This result is important as it suggests that the low performance earlier reported for DQ is at least in part imposed by a low quality and quantity of the earlier DQ data.

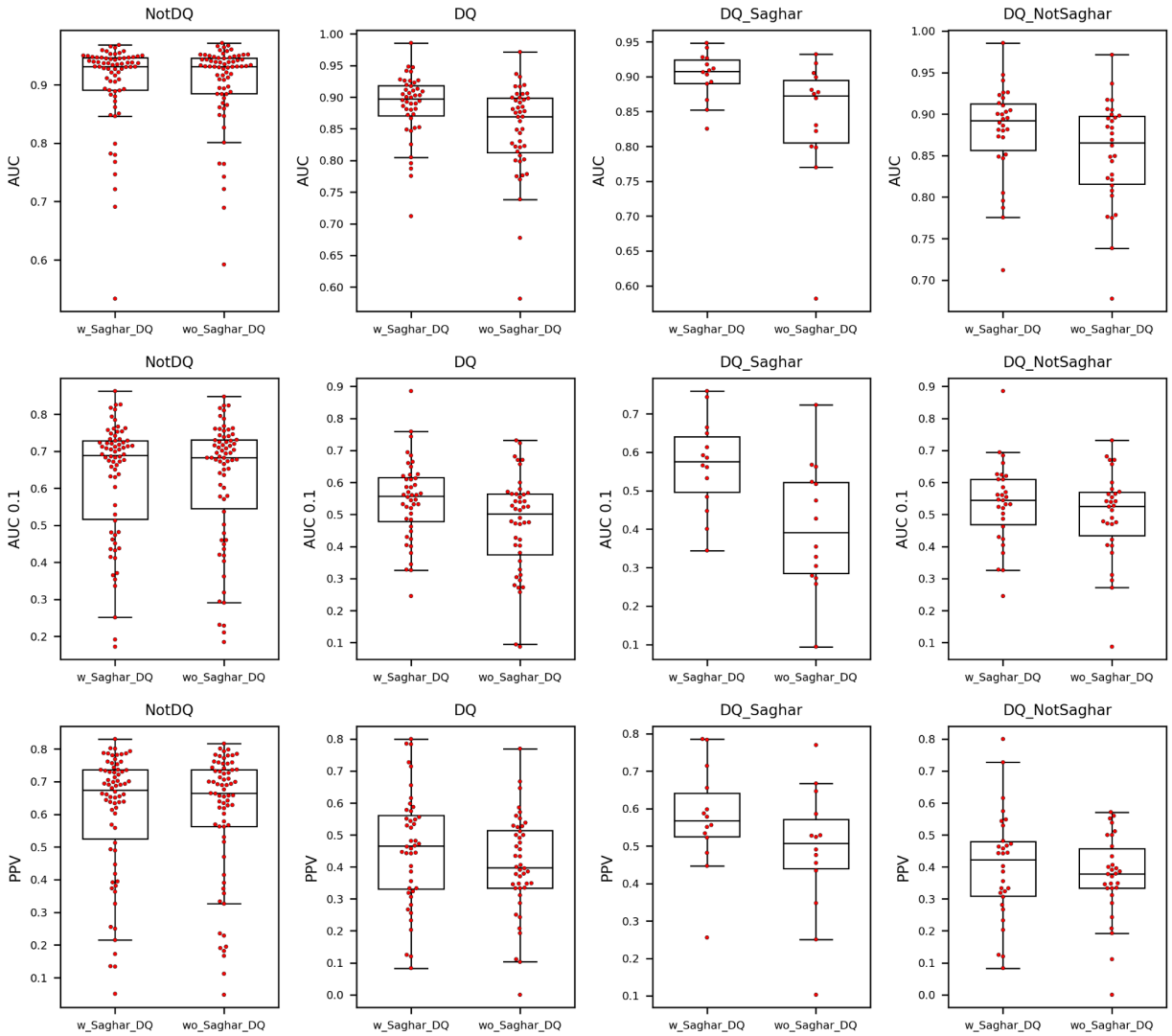


Figure 1: AUC, AUC 0.1 and PPV predictive performance for the two models on different data subsets. Each point is the performance metric for a unique HLA class II molecule. For details on the performance metrics refer to materials and methods. NotDQ includes all non-HLA-DQ molecules; DQ all DQ molecules, DQ_Saghar includes all DQ molecules in the novel data set, and DQ_NotSaghar DQ molecules not present in the novel data.

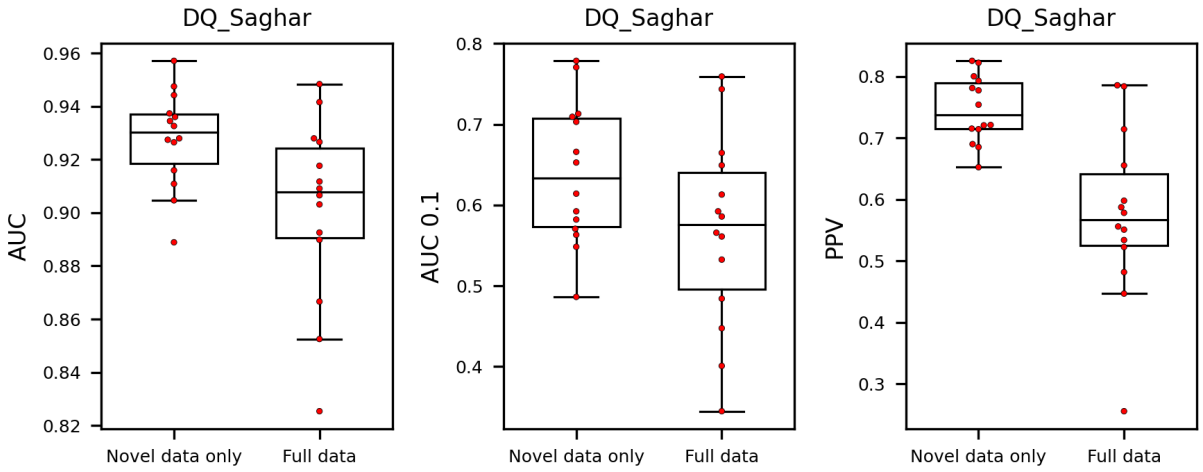


Figure 2: Performance of model trained including the novel data, evaluated on both the novel data alone, as well as the full dataset. Each point is the performance metric for an HLA-DQ molecule present in the novel data.

We next looked at the differences in peptides assigned to HLA-DQ molecules between the two methods. Here, trash peptides with percentile rank greater than 20 were excluded (Kaabinejadian et al., 2022). Overall the two methods share a high degree of overlap in the peptides assigned to DQ (60959 annotations were shared by both models, 9309 annotations were unique for the method trained including the novel data and 4316 unique for the method trained without). This increased DQ coverage for the model trained including the novel data predominantly comes from peptides assigned to DR (and to some degree trash and DP) by the model trained without the novel data (see supplementary table 3).

To further quantify the improved performance of the model trained including the novel data, we show the mean consistency value per HLA molecule in the four data subsets in supplementary figure 1. Here the consistency analysis was conducted as described earlier (Reynisson et al., 2020a) and detailed in materials and methods. In short, position-specific scoring matrices were constructed for each molecule in a given cell line from the predicted binding cores in the individual positive peptides, and the consistency was quantified by the correlation of such matrices for the same molecule between different cell line data sets. Based on this analysis, an overall improved consistency is observed for the model trained with the novel DQ data. The difference is significant in all four subsets of HLA molecules except for the DQ_NotSaghar subset ($p<0.02$). Together, these observations demonstrate that incorporating the novel HLA-DQ data has allowed for an enriched identification of HLA-DQ peptide ligands, rescuing peptides otherwise assigned to alternative DR/DP molecules, resulting in improved motif deconvolution consistency and improved predictive power.

The above results were complemented by a comparison to a model training including the novel data using context peptide encoding. In short, context encoding refers to a scenario where information from the regions flanking the peptide is extracted from the source protein sequence, and included as additional input to the machine learning model. The inclusion of such context

information has in several earlier studies been demonstrated to boost the performance for HLA ligand identification (Alvarez et al., 2019; Barra et al., 2018; Reynisson et al., 2020a). The results of this comparison are shown in supplementary figure 2, and confirm these earlier findings. Also here, the model trained including context was found to significantly outperform the model trained without context in all performance metrics and data subsets (the only exception being the DQ_NotSaghar subset). Only when evaluating the motif consistency, did the context model fail to demonstrate a significantly increased performance (see supplementary figure 3). Given that the main focus of the remaining part of the manuscript is to investigate motif-deconvolution and the role of cis versus trans-only DQ α and β chain pairing in this context, we focus on the simpler model trained without context information from here on.

Distribution of annotations to cis vs trans-only DQ molecules

We next moved to the main objective of this work, and investigated the distribution of peptides assigned to cis versus trans-only encoded DQ molecules across DQ-heterozygous datasets for the two models. Only datasets with at least 100 peptides assigned to DQ after excluding trash were included in this analysis. The proportion of DQ-annotated peptides assigned to each molecule was then calculated for each dataset containing that molecule. Finally, the mean per-dataset peptide fraction was reported for each DQ molecule, and the distribution of these means for molecules belonging to the cis and trans-only categories was investigated. The result of this is shown in figure 3 for the two models, and indicates that for the method including the novel data, trans-only molecules occupy a consistently small proportion of the DQ annotations in each cell line. On the other hand, the cis molecules have generally high contribution, apart for a limited set of outliers with low (<10%) mean fractions. When investigating these outliers, they were all found to originate from cell lines in which none of the 4 DQ molecules are trans-only. As such, these molecules get low peptide counts in these cell lines due to other molecules being favored (Ettinger et al., 1998). For the model including the novel data, the cis and trans-only contributions were found to be significantly different as expected ($p=0.0002$, two-sample t-test). Further, an overall higher contribution of trans-only molecules to the DQ peptide annotations was observed for the model trained without the novel data ($p=0.03$, paired one-sided t-test). This is especially clear when looking at difference in contribution of trans-only molecules in individual cell lines between the two models (supplementary figure 4).

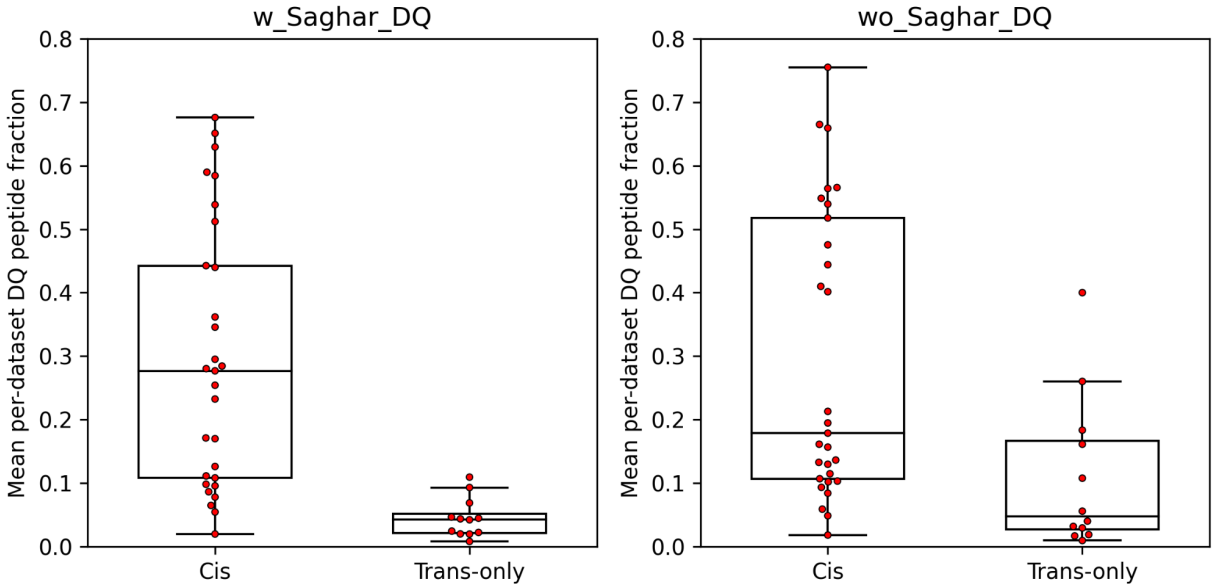


Figure 3: Peptide-count contribution of cis and trans-only molecules in the methods with (w_Saghar_DQ) and without (wo_Saghar_DQ) the novel data. Each point shows the mean per-dataset peptide fraction for a given DQ molecule.

To further investigate this, the DQ motif deconvolution of the two models for the Racle__TIL1 dataset is shown in figure 4. Here, the model trained without the novel data assigns a large proportion of peptides (170 out of 425) to HLA-DQA1*01:01-DQB1*03:01, which is a trans-only molecule known to not form a stable heterodimer (Creary et al., 2021; Kwok et al., 1993). On the other hand, in the model trained with the novel data, almost no peptides are assigned to this molecule (20 out of 459). Instead the peptides are assigned to the cis molecule HLA-DQA1*03:03-DQB1*03:01. Note also, that for both models a very minor proportion of peptides are assigned to HLA-DQA1*03:03-DQB1*05:01, another trans-only heterodimer known to be unstable (Creary et al., 2021; Kwok et al., 1993). Overall, these results demonstrate that the model including the novel DQ data allows for proper motif deconvolution with limited assignment of peptides to trans-only HLA-DQ molecules. Further, the overall very low proportion of peptides assigned to trans-only molecules, combined with the overall increased HLA-DQ peptide volume and motif consistency of the model trained including the novel data, strongly suggests that trans-only HLA-DQ molecules are unstable, non-functional and have limited to no contribution to the total HLA-DQ immunopeptidome.

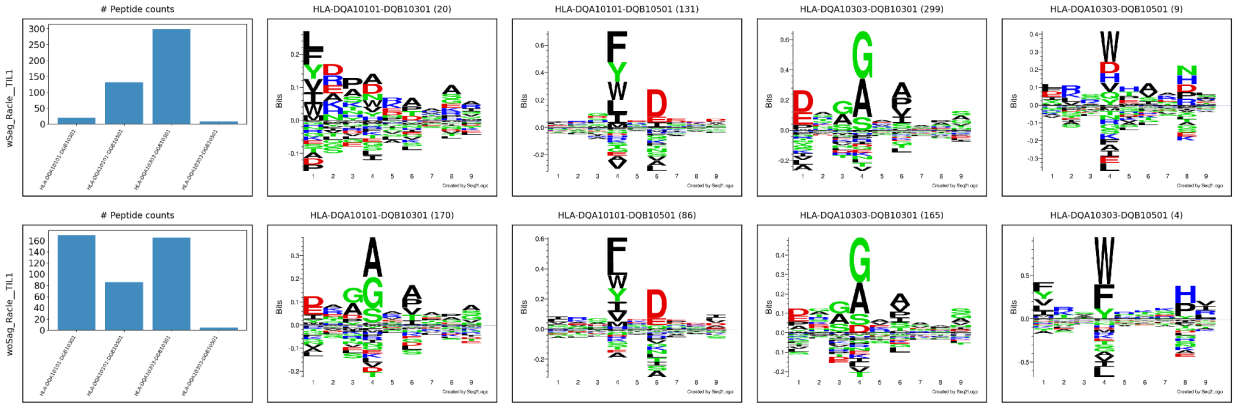


Figure 4: DQ motif deconvolution for the Racle__TIL1 dataset. wSag and woSag refers to the methods trained with and without the novel data, respectively. Trash peptides with percentile rank greater than 20 are not included in the logo plots.

Improved motif consistency

To further illustrate how the novel DQ data has improved the DQ motif deconvolution, consistency correlation and logo plots for a representative molecule are shown in supplementary figure 5 for the models trained with and without the novel data. Here, the analysis is shown for HLA-DQA1*03:01-DQB1*03:02, which is one of the DQ molecules in the novel data. The correlation plots in the first panel of the figure illustrate the overall improved consistency for the model trained with the novel data for this molecule. When looking at the sequence logo plots in the two methods, we see highly consistent motifs across all cell lines in the method including the novel data. In contrast, for the method without the novel data we see consistently lower peptide counts across cell lines, resulting in noisier motifs with less consistency across the different cell lines. The consistent increase in peptide counts for the method including the novel data is observed for the vast majority of molecules with improved consistency (see supplementary tables 4 and 5).

Furthermore, binding motifs for the DQ molecules contained in the novel data obtained by motif deconvolution of the MS data were visualized, along with sequence motifs based on predicted binders, in supplementary figure 6. Here, the logos obtained from the motif deconvolution are in most cases very similar when comparing the models trained with and without the novel data. This illustrates the strong pan-specificity of NNAlign_MA with respect to assigning peptides to molecules that the method has not learned fully during training. However, the predicted sequence logos based on top scoring random natural peptides indicate that the model trained without the novel DQ data has failed to fully learn the correct binding motifs of all the novel DQ molecules, especially with respect to the P1 amino acid preferences. To further quantify these results, correlations between the two methods' deconvoluted logos were calculated, along with correlations between the deconvoluted and predicted logos for each method. Firstly, PCC values were calculated between the PSSMs generated from the motif convolutions for each DQ contained in the novel data for the models trained with and without this novel data. The result of

this is shown in supplementary figure 7A displaying correlations between the motif deconvolution logos for the methods trained with and without the novel data, showing very high correlations between the two models for all DQ molecules (median PCC 0.995). Secondly, for each novel DQ, the PCC value between the motif deconvolution and predicted PSSM was calculated for each of the methods with and without the novel data. Supplementary figure 7B shows the result of this analysis. Here, a higher correlation is observed for the method with the novel data in 15 out of 16 PSSM pairs, indicating a highly consistent correspondence between the identified and predicted sequence logos (median PCC values are 0.95 and 0.89 for the method with and without the novel data, respectively).

Difference in peptide length distributions of DR and DQ

When analyzing the peptides assigned to DQ in the model with the novel data, it was found that their lengths were in general shorter than the peptides annotated towards other loci. To illustrate this, we visualized in supplementary figure 8 the length distributions of peptides for both DR and DQ molecules. Here, only DQ molecules with at least 1000 peptides and DR molecules with at least 2000 peptides (both excluding trash) were included. This figure indicates that HLA-DQ molecules, with a preference for mainly 14-15 mers, generally bind shorter peptides compared to HLA-DR. On the other hand, the DR molecules show peptide length distributions most often peaking between 15 and 17, apart from one outlier with a mode of 14. Overall, the average mode of peptide lengths is 14.7 for the DQ molecules and 15.5 for the DR molecules. One outlier distribution for DQ was observed, namely for HLA-DQA1*01:02-DQB1*06:02 (a molecule present in the novel data), whose peptide length mode was 16. When investigating the peptides for this molecule, it was found that the longer peptides were generally part of the original NetMHCIIpan-4.1 data, which thus shifts the overall distribution towards longer lengths. The cause of this discrepancy in peptide lengths between the novel and earlier data for this molecule remains unknown.

Coverage of DQ

Next, we wanted to assess the number of DQ molecules present in the cross-validation predictions by each model which were properly covered (i.e. had a large number of peptides assigned during training), and hence where the models are expected to achieve accurate predictive power. The peptide count for a given DQ molecule was estimated as the accumulated sum of peptides from each cell line containing that molecule (excluding trash peptides). Here, only peptides annotated to DQ molecules in a given cell line corresponding to at least 5% of the total number of DQ peptides were included in its count (this was done to avoid including accumulation of low count noise). A given DQ molecule was then said to be covered if the summed peptide count over all cell lines was at least 100. This analysis resulted in 24 DQ molecules being covered by the model trained including the novel data, and 23 being covered when excluding these data. None of the 24 DQ molecules covered by the model including the novel data were found to be trans-only, whereas the model without the novel data covered two trans-only DQ molecules, namely HLA-DQA1*01:01-DQB1*03:01 (as described earlier) and HLA-DQA1*01:03-DQB1*03:02. Of the remaining 21 molecules, 20 were included in the molecules covered by the model trained with the novel data.

Now, for each of the two methods, we investigated the proportion of 154 prevalent DQ molecules that had a distance of at most 0.025 to a molecule covered by the model (this set of molecules is here referred to as ‘extended coverage’). For details on how this distance was determined and how the list of prevalent DQ molecules was defined refer to materials and methods. Note, that 0.025 is a very conservative distance threshold, and that we expect the model to maintain accuracy also for molecules falling beyond this value (Karosiene et al., 2013). Here, a significant gain in extended coverage was found ($p=0.027$, chi-squared test), with the model including the novel data covering 94 out of 154 molecules, while the model without the novel data only covered 75 out of 154 molecules (see supplementary tables 6 and 7 for a list of covered and non-covered DQ molecules for the model trained including the novel data). When comparing the covered and non-covered molecules for the method including the novel data, the non-covered group had significantly lower worldwide haplotype frequency data as obtained from Allelefrequencies.net (for detail on how these frequencies were obtained refer to material and methods) compared to the covered group (average frequencies for the two groups were 0.0134 and 0.0025, $p<0.03$, student t-test). These results suggest that the list of non-covered DQ molecules are of limited importance seen from a population coverage perspective.

For visualizing the coverage of the DQ space, a specificity tree was constructed. Here, we used the list of 154 prevalent HLA-DQ molecules as the starting point. This list was first reduced to a set of 61 molecules with unique specificities (for details see methods) which were included in the subsequent analysis. Next, a specificity tree was constructed covering the 61 DQ molecules applying the MHCCluster method (Thomsen et al., 2013). In short, the MHCCluster method estimates the similarity between two MHC molecules using the correlation between predicted binding values for a large set of random natural peptides. Figure 5 shows the resulting specificity tree along with predicted binding motifs for the 14 novel DQ molecules. The tree displays wide coverage of the DQ space, as all the novel molecules are spread more or less uniformly across the different branches of the tree, and all branches are covered by one or more DQ molecules in close distance to the DQ molecules covered by the training data.

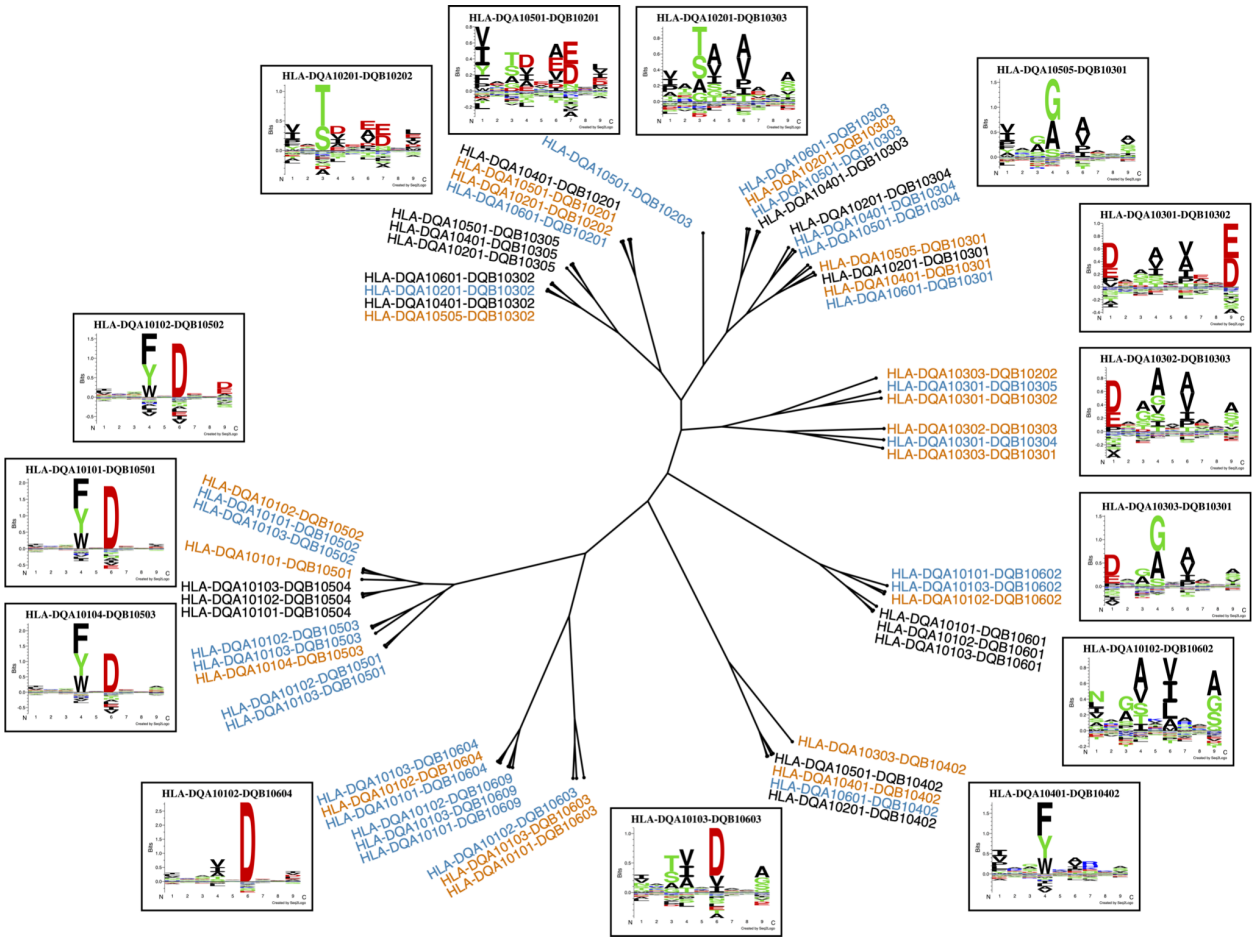


Figure 5: HLA-DQ specificity tree. The tree is based on 61 DQ molecules including the 14 molecules described by the novel data (highlighted in the figure by boxed sequence logos). Orange molecules are covered by the method including the novel data with at least 100 peptides, and blue molecules are within a distance 0.05 of an orange molecule. Black molecules are non-covered (i.e. have peptide count less than 100 and have distance greater than 0.05 to an orange molecule). The specificity tree was calculated from the pairwise similarities between the predictions scores for the DQ molecules for a set of 100,000 random natural 13-17mer peptides. Logos were constructed for the top 1% highest scoring binding cores for these 100,000 peptides.

NetMHCIIpanDQ

The model developed here including the novel DQ immunopeptidome data is made publicly available at https://services.healthtech.dtu.dk/service.php?NetMHCIIpanDQ-1.0. The method allows for prediction of HLA antigen presentation to all HLA-DQ molecules, and prediction can be made with or without context encoding.

Discussion

In this work, we have demonstrated how rational data generation combined with refined immunoinformatics data mining can boost the performance of HLA class II antigen presentation predictions and move towards closing the performance gap between HLA-DR and HLA-DQ.

We generated high quality MS-immunopeptidomics data from a series of 16 HLA-DQ homozygous cell lines covering a total of 14 different frequent HLA-DQ molecules in different populations worldwide. Using an in-house HLA-DQ specific antibody enabled identification of MS-immunopeptidomics datasets of an, in a DQ context, unprecedented volume with an average of 2,600 unique peptides identified in each cell line. Integrating this large volume of data with earlier data from the development of NetMHCIIpan-4.1 allowed us to boost the HLA-DQ antigen presentation predictive performance to a level comparable to that of HLA-DR. Investigating the accuracy of the motif deconvolution of the two methods trained with and without the novel data demonstrated an overall improved motif consistency across all HLA molecules. This observation demonstrates how integration of the novel HLA-DQ data results in an overall improved HLA-restriction assignment of the individual MS-HLA-peptides leading to more accurate motif characterizations across all three HLA class II loci. The main source of this improvement was demonstrated to be an increased volume of peptide assignment to HLA-DQ molecules during the motif deconvolution. This resulted in improved motif accuracy for both HLA-DQ imposed by the larger volume of peptides, and non HLA-DQ molecules imposed by the removal of peptides mis-assigned as DQ restricted by the model not including the novel DQ data.

Next, moving into the issue of cis versus trans-only HLA-DQ α and β chain combinations, we demonstrated that in contrast to the method without the novel data, the model trained including the novel data performed the DQ motif deconvolution almost solely by use of known HLA-DQ cis-variants. One particular example here was the HLA-DQ molecule HLA-DQA1*01:01-DQB1*03:01, which in earlier studies has been demonstrated to be non-functional (Kwok et al., 1993). This molecule was assigned a large number of peptides in the model trained without the novel data. However, when including the novel data, the peptide assignment to this molecule was almost completely depleted. This result combined with the overall increased HLA-DQ peptide volume and motif consistency of the model trained including the novel data, strongly suggests that trans-only HLA-DQ α and β combinations are likely non-functional and have minimal or no contribution to the total HLA-DQ immunopeptidome. This finding is striking since the definition for cis and trans-only dimerization defined here precisely follows the rules proposed earlier for forming stable/unstable HLA-DQ heterodimers. Specifically, the rules indicate that structural constraints do not favor dimerization of DQA1*01 with DQB1*02, 03, and 04 alleles, resulting in their lack of function (Kwok et al., 1993; Petersdorf et al., 2022). These results thus demonstrate how such rules can be learned directly from MS-immunopeptidome data using tailored data mining methods and rationally defined data sets, suggesting that similar types of analysis should be extended to HLA-DP to further our understanding of cis versus trans α- and β chain pairing.

Note, that the definition of cis and trans-only HLA-DQ α and β chain combinations applied in this work is contingent on the current haplotype data available and the assumption that all observed haplotype α and β pair can pair and form cis-variants, and likewise that all unobserved such cis-variants are trans-only. The current data defining these categories are limited in volume, and larger sample sizes are required for more accurate analyses particularly for the more heterogenetic groups and low frequency haplotypes (Creary et al., 2021).

Lastly, we demonstrated how the coverage of HLA-DQ molecules was largely increased by the models trained with novel data, and illustrated this by constructing an HLA-DQ tree showing coverage of all branches. This suggests that the current model covers all HLA-DQ binding specificities (at least within the assumption that *trans-only* HLA-DQ molecules have limited contribution to the overall HLA-DQ immunopeptidome).

Overall, this work has demonstrated how careful data generation using a DQ-specific antibody and affinity purification combined with refined data mining and motif deconvolution can be applied towards closing the performance gap in peptide binding prediction between HLA-DR and HLA-DQ. Despite the large performance gain demonstrated here, the accuracy for HLA-DQ remains below compared to what is observed for DR. We demonstrate that this to a very large degree can be attributed to the general lower quantity and quality of ligands obtained in earlier DQ immunoprecipitation studies where most often DQ (and DP) data have been obtained using a pan-HLA class II antibody (after first depleting for HLA-DR (Racle et al., 2019)). Focusing solely on the novel data generated in this study, we find that both the quantity and quality of the obtained DQ ligands are on par with what is found for HLA-DR, resulting in predictive performance for the associated dataset being equal between the two. This result has large impacts and suggests that modeling DQ is a task of equal complexity to that of HLA-DR, and that the current lower performance of DQ compared to DR is driven by low quantity and quality of data; a situation that can be resolved by generation of high quality and volume data as outlined in this study.

In conclusion, other than demonstrating an overall improved predictive performance and coverage of HLA-DQ molecules, a key result of our work is an improved understanding of the relative contribution of cis versus trans-only paired molecules to the overall HLA-DQ immunopeptidome demonstrating a very limited role of the latter in complementing the specificity space. We believe these findings will provide a foundation for further research defining the molecular role of HLA-DQ in the onset of cellular immunity within autoimmune and infectious diseases.

References

- Abelin, J.G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S.P., Creech, A.L., Serrano, L.R., Nasir, G., Nasrullah, Y., et al. (2019). Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. Immunity *51*, 766-779.e17. https://doi.org/10.1016/j.immuni.2019.08.012.
- Alvarez, B., Barra, C., Nielsen, M., and Andreatta, M. (2018). Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes. Proteomics *18*, 1700252. https://doi.org/https://doi.org/10.1002/pmic.201700252.
- Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M., and Nielsen, M. (2019). NNAlign_MA; MHC peptidome Deconvolution for Accurate MHC Binding Motif Characterization and Improved T-cell Epitope Predictions. Molecular and Cellular Proteomics *18*, 2459–2477. https://doi.org/10.1074/mcp.TIR119.001658.
- Arango, M.T., Perricone, C., Kivity, S., Cipriano, E., Ceccarelli, F., Valesini, G., and Shoenfeld, Y. (2017). HLA-DRB1 the notorious gene in the mosaic of autoimmunity. Immunol Res *65*, 82–98. https://doi.org/10.1007/s12026-016-8817-7.
- Barra, C., Alvarez, B., Paul, S., Sette, A., Peters, B., Andreatta, M., Buus, S., and Nielsen, M. (2018). Footprints of antigen processing boost MHC class II natural ligand predictions. Genome Med *10*. https://doi.org/10.1186/s13073-018-0594-6.
- Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., et al. (2016). Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun *7*, 13404. https://doi.org/10.1038/ncomms13404.
- Bergseng, E., Dørum, S., Arntzen, M.Ø., Nielsen, M., Nygård, S., Buus, S., de Souza, G.A., and Sollid, L.M. (2014). 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. https://doi.org/10.1007/s00251-014-0819-9.
- Caron, E., Kowalewski, D.J., Koh, C.C., Sturm, T., Schuster, H., and Aebersold, R. (2015). Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. Molecular and Cellular Proteomics *14*, 3105–3117. https://doi.org/10.1074/mcp.O115.052431.
- Creary, L.E., Sacchi, N., Mazzocco, M., Morris, G.P., Montero-Martin, G., Chong, W., Brown, C.J., Dinou, A., Stavropoulos-Giokas, C., Gorodezky, C., et al. (2021). High-resolution HLA allele and haplotype frequencies in several unrelated populations determined by next generation sequencing: 17th International HLA and Immunogenetics Workshop joint report. Hum Immunol *82*, 505–522. https://doi.org/https://doi.org/10.1016/j.humimm.2021.04.007.
- Erlich, H., Valdes, A.M., Noble, J., Carlson, J.A., Varney, M., Concannon, P., Mychaleckyj, J.C., Todd, J.A., Bonella, P., Fear, A.L., et al. (2008). HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk analysis of the type 1 diabetes genetics consortium families. Diabetes *57*, 1084–1092. https://doi.org/10.2337/db07-1331.
- Ettinger, R.A., Liu, A.W., Nepom, G.T., and Kwok, W.W. (1998). Exceptional Stability of the HLA-DQA1*0102/DQB1*0602 αβ Protein Dimer, the Class II MHC Molecule

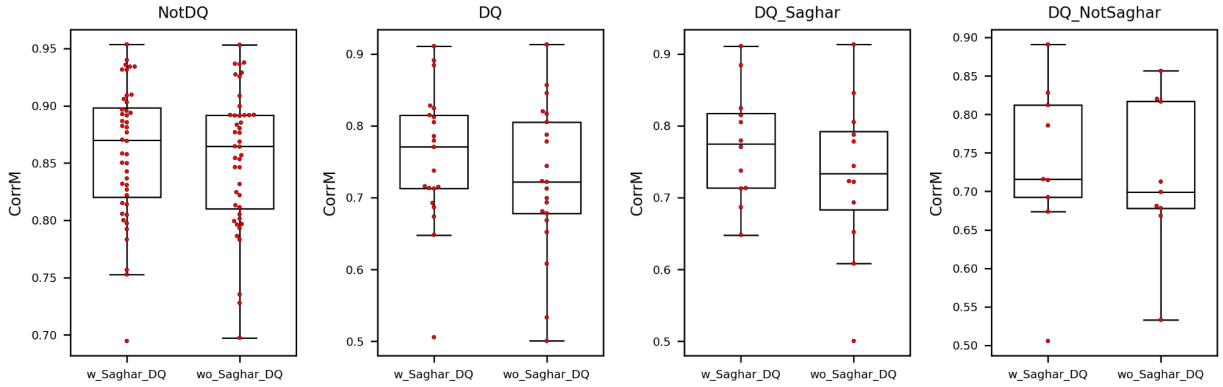
Associated with Protection from Insulin-Dependent Diabetes Mellitus. The Journal of Immunology *161*, 6439–6445. .

- Gfeller, D., and Bassani-Sternberg, M. (2018). Predicting antigen presentation-What could we learn from a million peptides? Front Immunol *9*, 1716. https://doi.org/10.3389/fimmu.2018.01716.
- Gonzalez-Galarza, F.F., Christmas, S., Middleton, D., and Jones, A.R. (2011). Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Res *39*. https://doi.org/10.1093/nar/gkq1128.
- Hoof, I., Peters, B., Sidney, J., Pedersen, L.E., Sette, A., Lund, O., Buus, S., and Nielsen, M. (2009). NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics *61*, 1–13. https://doi.org/10.1007/s00251-008-0341-z.
- Hu, X., Deutsch, A.J., Lenz, T.L., Onengut-Gumuscu, S., Han, B., Chen, W.M., Howson, J.M.M., Todd, J.A., de Bakker, P.I.W., Rich, S.S., et al. (2015). Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. Nat Genet *47*, 898–905. https://doi.org/10.1038/ng.3353.
- Kaabinejadian, S., Barra, C., Alvarez, B., Yari, H., Hildebrand, W.H., and Nielsen, M. (2022). Accurate MHC Motif Deconvolution of Immunopeptidomics Data Reveals a Significant Contribution of DRB3, 4 and 5 to the Total DR Immunopeptidome. Front Immunol *13*, 835454. https://doi.org/10.3389/fimmu.2022.835454.
- Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S., and 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. https://doi.org/10.1007/s00251-013-0720-y.
- Kwok, W.W., and Nepom, G.T. (1991). Structural and functional constraints on HLA class II dimers implicated in susceptibility to insulin dependent diabetes mellitus. Baillieres Clin Endocrinol Metab *5*, 375–393. https://doi.org/https://doi.org/10.1016/S0950-351X(05)80137-5.
- Kwok, W.W., Kovats, S., Thurtle, P., and Nepom, G.T. (1993). HLA-DQ allelic polymorphisms constrain patterns of class II heterodimer formation. The Journal of Immunology *150*, 2263–2272. .
- Lundin, K.E., Sollid, L.M., Qvigstad, E., Markussen, G., Gjertsen, H.A., Ek, J., and Thorsby, E. (1990). T lymphocyte recognition of a celiac disease-associated cis- or trans-encoded HLA-DQ alpha/beta-heterodimer. The Journal of Immunology *145*, 136–139. .
- McFarland, B.J., and Beeson, C. (2002). Binding interactions between peptides and proteins of the class II Major Histocompatibility Complex. Med Res Rev *22*, 168–203. https://doi.org/https://doi.org/10.1002/med.10006.
- Moore, R.M., Harrison, A.O., McAllister, S.M., Polson, S.W., and Eric Wommack, K. (2020). Iroki: Automatic customization and visualization of phylogenetic trees. PeerJ *8*, e8584. https://doi.org/10.7717/PEERJ.8584.
- Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med *8*. https://doi.org/10.1186/s13073-016-0288-x.

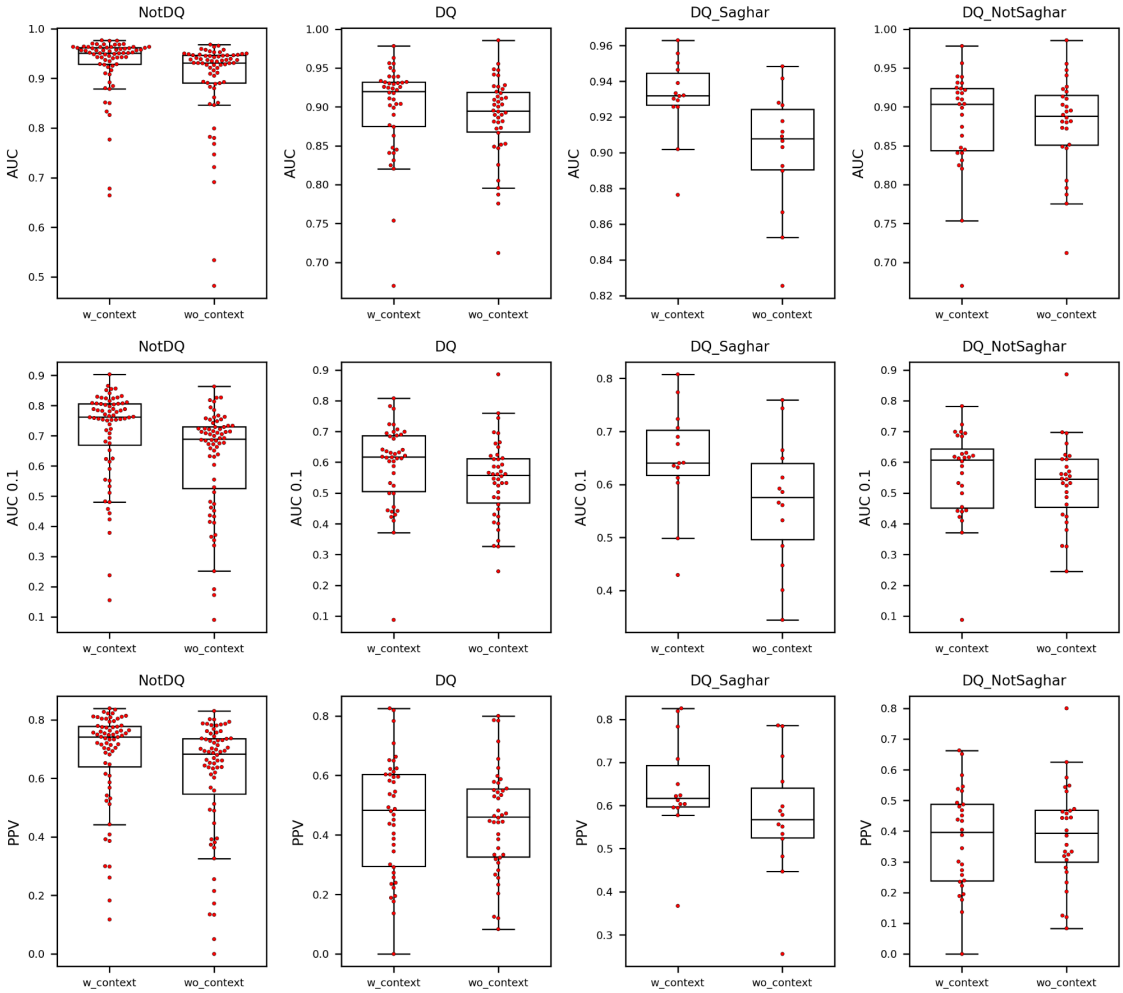
- Nielsen, M., Lundegaard, C., and Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics *8*, 238. https://doi.org/10.1186/1471-2105-8-238.
- Nielsen, M., Lund, O., Buus, S., and Lundegaard, C. (2010). MHC Class II epitope predictive algorithms. Immunology *130*, 319–328. https://doi.org/https://doi.org/10.1111/j.1365-2567.2010.03268.x.
- Nielsen, M., Andreatta, M., Peters, B., and Buus, S. (2020). Immunoinformatics: Predicting Peptide–MHC Binding. Annu Rev Biomed Data Sci *3*, 191–215. https://doi.org/10.1146/annurev-biodatasci-021920-100259.
- Paul, S., Karosiene, E., Dhanda, S.K., Jurtz, V., Edwards, L., Nielsen, M., Sette, A., and Peters, B. (2018). Determination of a predictive cleavage motif for eluted major histocompatibility complex class II ligands. Front Immunol *9*, 1795. https://doi.org/10.3389/fimmu.2018.01795.
- Petersdorf, E.W., Bengtsson, M., Horowitz, M., McKallor, C., Spellman, S.R., Spierings, E., Gooley, T.A., Stevenson, P., and Transplantation, on behalf of the I.H.W.G. in H.C. (2022). HLA-DQ heterodimers in hematopoietic cell transplantation. Blood *139*, 3009–3017. https://doi.org/10.1182/blood.2022015860.
- Purcell, A.W., Ramarathinam, S.H., and Ternette, N. (2019). Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. Nat Protoc *14*, 1687–1707. https://doi.org/10.1038/s41596-019-0133-y.
- Racle, J., Michaux, J., Rockinger, G.A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., et al. (2019). Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat Biotechnol *37*, 1283–1286. https://doi.org/10.1038/s41587-019-0289-6.
- Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W.H., Peters, B., and Nielsen, M. (2020a). Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. J Proteome Res *19*, 2304–2315. https://doi.org/10.1021/acs.jproteome.9b00874.
- Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020b). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res *48*, W449–W454. https://doi.org/10.1093/nar/gkaa379.
- Rocha, N., and Neefjes, J. (2008). MHC class II molecules on the move for successful antigen presentation. Embo Journal *27*, 1–5. https://doi.org/10.1038/sj.emboj.7601945.
- Sidney, J., Steen, A., Moore, C., Ngo, S., Chung, J., Peters, B., and Sette, A. (2010). Divergent motifs but overlapping binding repertoires of six HLA-DQ molecules frequently expressed in the worldwide human population. Journal of Immunology *185*, 4189–4198. https://doi.org/10.4049/jimmunol.1001006.
- Stepniak, D., Wiesner, M., de Ru, A.H., Moustakas, A.K., Drijfhout, J.W., Papadopoulos, G.K., van Veelen, P.A., and Koning, F. (2008). Large-scale characterization of natural ligands explains the unique gluten-binding properties of HLA-DQ2. Journal of Immunology *180*, 3268–3278. https://doi.org/10.4049/jimmunol.180.5.3268.

- Thomsen, M.C.F., Lundegaard, C., Buus, S., Lund, O., and Nielsen, M. (2013). MHCcluster, a method for functional clustering of MHC molecules. Immunogenetics *65*, 655–665. https://doi.org/10.1007/s00251-013-0714-9.
- Tollefsen, S., Hotta, K., Chen, X., Simonsen, B., Swaminathan, K., Mathews, I.I., Sollid, L.M., and Kim, C.Y. (2012). Structural and functional studies of trans-encoded HLA-DQ2.3 (DQA1*03:01/DQB1*02:01) protein molecule. Journal of Biological Chemistry *287*, 13611–13619. https://doi.org/10.1074/jbc.M111.320374.
- Vartdal, F., Johansen, B.H., Friede, T., Thorpe, C.J., Stevanović, S., Eriksen, J.E., Sletten, K., Thorsby, E., Rammensee, H.-G., and Sollid, L.M. (1996). The peptide binding motif of the disease associated HLA-DQ (α 1* 0501, β 1* 0201) molecule. Eur J Immunol *26*, 2764–2772. https://doi.org/https://doi.org/10.1002/eji.1830261132.
- Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., and Peters, B. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. B M C Bioinformatics *568*, 568. https://doi.org/10.1186/1471-2105-11-568.

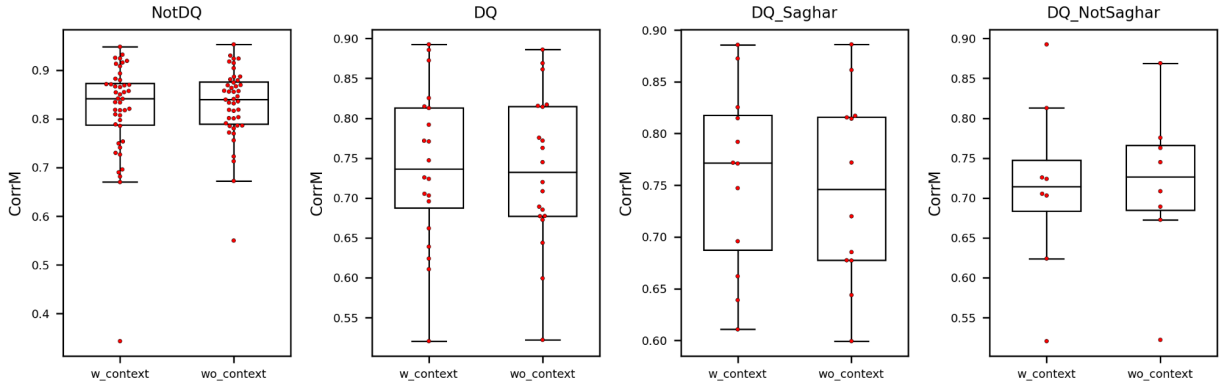
Supplementary figures



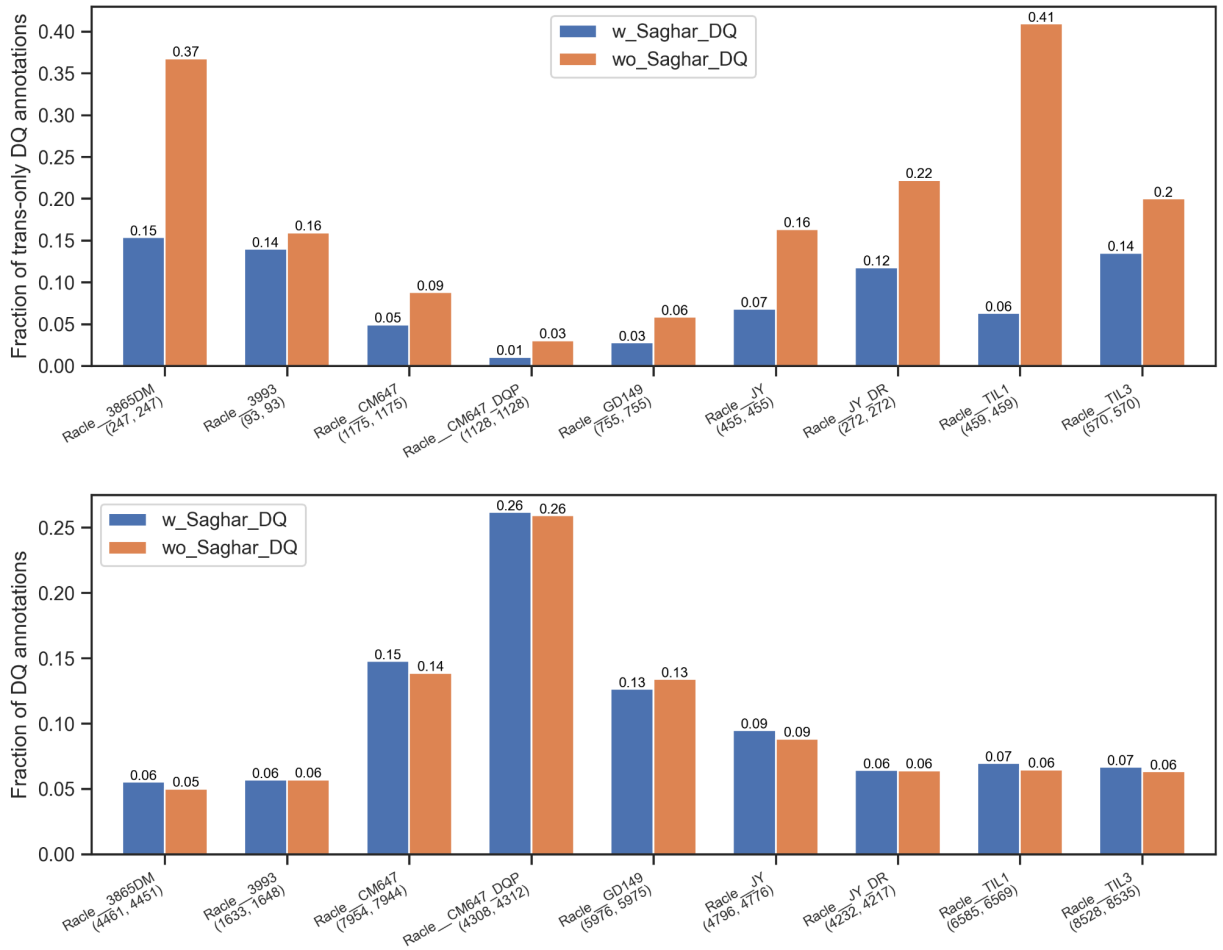
Supplementary figure 1: Mean consistency values for HLA molecule deconvolutions shared between multiple data sets for the two models. Each point reflects an HLA class II molecule, and values shown are the mean over the different data set comparisons. The different subgroups of HLA molecules are defined in the text and the Figure 1 caption.



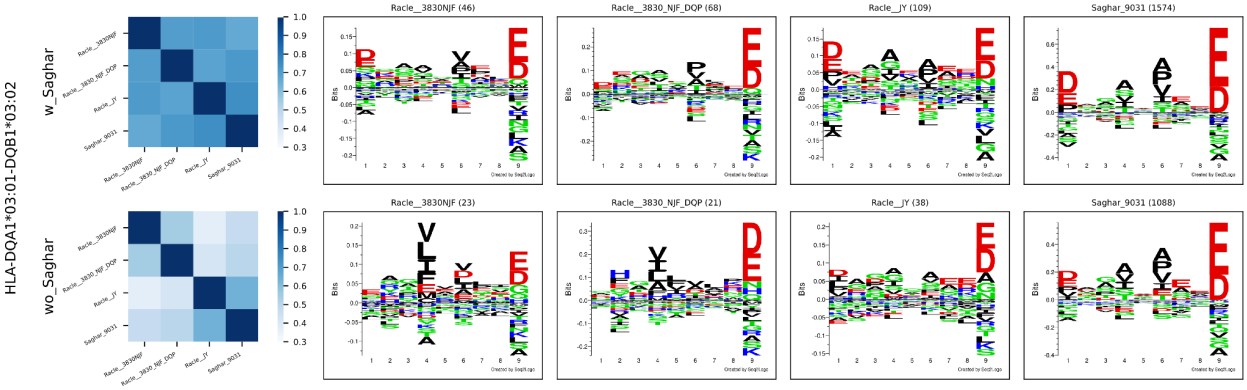
Supplementary figure 2: Performance comparison of models trained with the novel DQ data, (w_context) and without (wo_context) context encoding.



Supplementary figure 3: Consistency correlation analysis of models trained with the novel data, with (w_context) and without (wo_context) context encoding. The analysis was performed as in the main analysis in the manuscript, i.e. the union of identified trash peptides was removed, and a minimum threshold of 20 peptides was required in order for a given PSSM in the analysis to be generated.

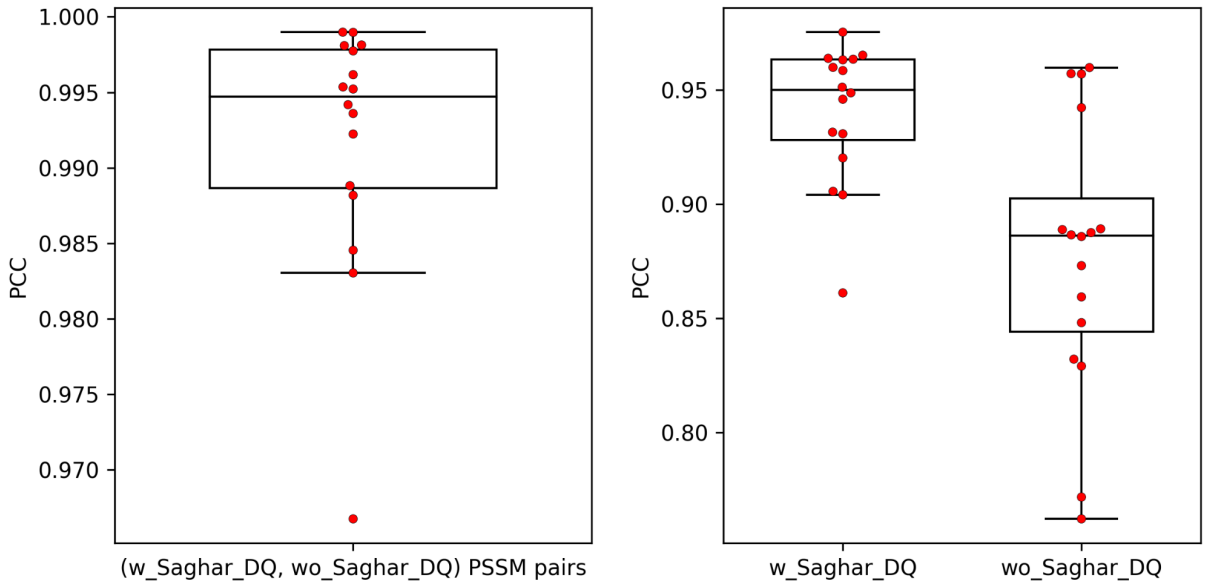


Supplementary figure 4: Distribution of DQ-annotated peptides for the methods trained with (w_Saghar) and without (wo_Saghar) the novel data. Only datasets in which peptides were assigned to both cis and trans-only DQ molecules in both methods were included in this figure. Furthermore, only cell lines in which at least 5% of the total non-trash peptide count was assigned to DQ molecules were included. **A:** Distribution of peptides assigned to trans-only molecules per sample. The numbers in parentheses show the total number of DQ-assigned peptides in each sample for each of the two methods (w_Saghar, wo_Saghar). **B:** Distribution of peptides assigned to DQ molecules per sample. The numbers in parentheses show the total number of peptides in each sample for each of the two methods (w_Saghar, wo_Saghar). Note, that peptide HLA annotation was performed including the complete list of possible HLA molecules for each cell line. This also in the case where HLA-specific antibodies (such as L243 for HLA-DR) were used in the immuno-precipitation.

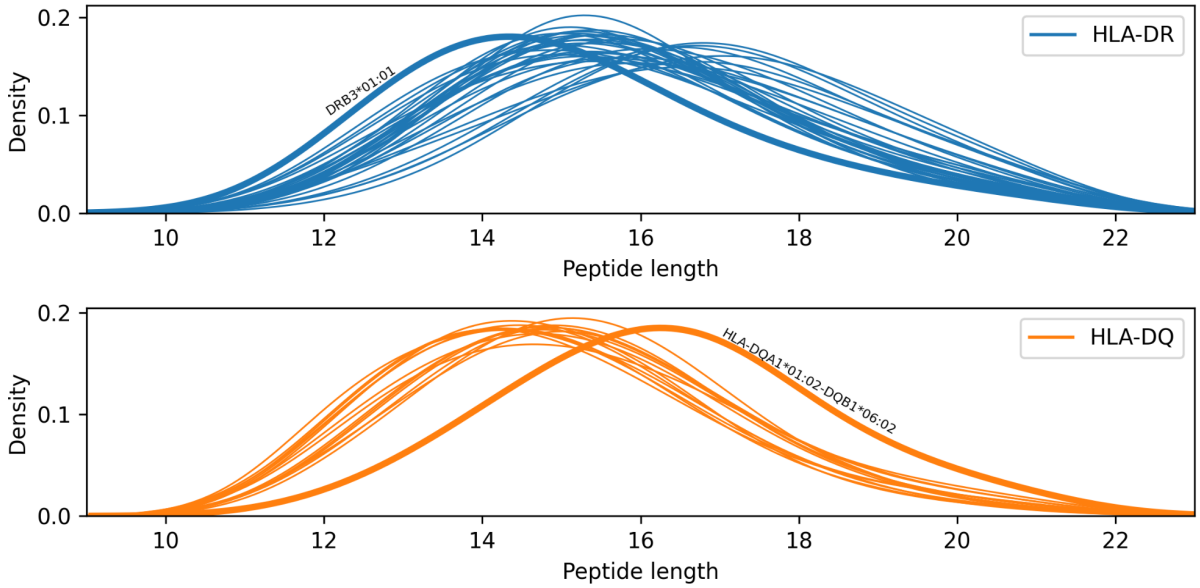


Supplementary figure 5. Consistency analysis for HLA-DQA1*03:01-DQB1*03:02. Here, only a subset of the cell lines used in the consistency analysis are included. Left panel: Correlation matrices showing the consistency values per cell line pair. Right panel: Sequence logos constructed from the peptide sets used in the consistency correlation analysis, with a minimum threshold of 20 peptides required for a given logo. The rows correspond to the methods including (w_Saghar) and not including (wo_Saghar) the novel data.

Supplementary figure 6: Identified binding motifs for the 14 DQ molecules from the novel data in the two methods. Left panel: Motif deconvolution of the novel datasets in the models trained with (wSag) and without (woSag) the novel data (the two rows with 4 logos correspond to the two HLA-DQ molecules shared between two cell lines). Right panel: Motifs predicted for the HLA-DQ molecules by the two models. Predicted motifs are generated from the top 1% of 100,000 random natural 13-17 mer peptides.



Suplementary figure 7: Correlations between deconvoluted and predicted sequence logos in the methods with (w_Saghar) and without (wo_Saghar) the novel data. Each point corresponds to a DQ molecule from a given cell line in the novel data. Left: correlations between the motif deconvolution PSSMs in the methods trained with and without the novel data. Right: correlations between motif deconvolution logos and predicted logos for each of the two methods.



Supplementary figure 8: Kernel density plots of length distributions of peptides annotated towards HLA-DR and HLA-DQ. Each line corresponds to the peptide length distribution for an HLA-DR (top) or HLA-DQ (bottom) molecule. A minimum peptide count of 1000 for DQ and 2000 for DR (excluding trash peptides) was required for a given molecule to be included. The distributions perceived as outliers are marked with a thicker line and annotated with the molecule name.

Supplementary tables

Supplementary table 1: Overview of HLA types per EL dataset. Below are shown the 16 cell lines in the novel data. The HLA typing for all EL datasets used in this publication can be found in the external datafile *allelelist.xlsx*. Note that the HLA typing for the two datasets Abelin__HLA_DR_SILAC_Donor2_DC_UV_16hr and Abelin__HLA_DR_SILAC_Donor2_DC_UV_24hr were modified to include all four possible DQ combinations.

Cell line	**HLA typing**
Saghar_721	DRB1*01:01,HLA-DQA1*01:01-DQB1*05:01,HLA-DPA1*01:03-DPB1*02:01
Saghar_9004	DRB1*01:01,HLA-DQA1*01:01-DQB1*05:01,HLA-DPA1*01:03-DPB1*04:01
Saghar_9013	DRB1*15:01,DRB5*01:01,HLA-DQA1*01:02-DQB1*06:02,HLA-DPA1*01:03-DPB1*04:02
Saghar_9023	DRB1*03:01,DRB3*01:01,HLA-DQA1*05:01-DQB1*02:01,HLA-DPA1*02:01-DPB1*01:01
Saghar_9031	DRB1*04:01,DRB4*01:03,HLA-DQA1*03:01-DQB1*03:02,HLA-DPA1*01:03-DPB1*04:01
Saghar_9043	DRB1*11:01,DRB3*02:02,HLA-DQA1*05:05-DQB1*03:01,HLA-DPA1*02:01-DPB1*10:01
Saghar_9052	DRB1*07:01,DRB4*01:03N,HLA-DQA1*02:01-DQB1*03:03,HLA-DPA1*01:03-DPB1*04:01
Saghar_9060	DRB1*13:01,DRB3*02:02,HLA-DQA1*01:03-DQB1*06:03,HLA-DPA1*02:02-DPB1*19:01
Saghar_9061	DRB1*14:01,DRB3*02:02,HLA-DQA1*01:04-DQB1*05:03,HLA-DPA1*01:03-DPB1*04:01
Saghar_9063	DRB1*13:02,DRB3*03:01,HLA-DQA1*01:02-DQB1*06:04,HLA-DPA1*01:03-DPB1*16:01
Saghar_9072	DRB1*08:02,HLA-DQA1*04:01-DQB1*04:02,HLA-DPA1*01:03-DPB1*04:02
Saghar_9075	DRB1*09:01,DRB4*01:03,HLA-DQA1*03:02-DQB1*03:03,HLA-DPA1*01:03-DPB1*04:01
Saghar_9084	DRB1*16:01,DRB5*02:02,HLA-DQA1*01:02-DQB1*05:02,HLA-DPA1*01:03-DPB1*04:01
Saghar_9090	DRB1*04:01,DRB4*01:03,HLA-DQA1*03:03-DQB1*03:01,HLA-DPA1*01:03-DPB1*04:01,HLA-DPA1*01:03-DPB1*14:01,HLA-DPA1*02:01-DPB1*04:01,HLA-DPA1*02:01-DPB1*14:01
Saghar_9093	DRB1*07:01,DRB4*01:03,HLA-DQA1*02:01-DQB1*02:02,HLA-DPA1*01:03-DPB1*02:01,HLA-DPA1*01:03-DPB1*04:01
Saghar_9100	DRB1*08:02,HLA-DQA1*04:01-DQB1*04:02,HLA-DPA1*01:03-DPB1*03:01,HLA-DPA1*01:03-DPB1*04:02

Supplementary table 2: Novel peptide ligand dataset, in the format PEPTIDE 1 CELL_LINE_ID CONTEXT. The data can be found in the external datafile *saghar_peptide_data.xlsx*.

Supplementary table 3: HLA annotations of peptides assigned to DQ by the model with (w_Saghar) the novel data in the method without (wo_Saghar) the novel data.

	Distribution of peptide annotation in wo_Saghar method		
Peptides uniquely assigned to HLA-DQ in w_Saghar method	HLA-DP	HLA-DR	Trash
9309	1816	4688	2805

Supplementary table 4: Overview of peptides assigned to DQ molecules in the methods with (w_Saghar) and without (wo_Saghar) the novel data. Trash peptides with percentile rank greater than 20 are not included in the metrics. A bold value indicates either a higher average peptide count or a lower mean/median percentile rank in a given method.

	Average number of peptides per dataset		Average of per-dataset mean percentile rank		Median of per-dataset median percentile rank	
Molecule	**w_Saghar**	**wo_Saghar**	**w_Saghar**	**wo_Saghar**	**w_Saghar**	**wo_Saghar**
HLA-DQA1*01:01-DQB1*02:01	6.00	**7.50**	**3.2818**	3.7402	**0.6713**	0.8226
HLA-DQA1*01:01-DQB1*03:01	20.00	**170.00**	**0.3084**	1.8631	**0.0086**	0.4087
HLA-DQA1*01:01-DQB1*03:03	27.00	**58.00**	**0.4281**	0.5240	**0.0078**	0.0086
HLA-DQA1*01:01-DQB1*05:01	**209.53**	155.06	2.6547	**1.1562**	1.1783	**0.1096**
HLA-DQA1*01:01-DQB1*06:02	**30.50**	27.67	1.1216	**0.9356**	**0.2011**	0.2064
HLA-DQA1*01:01-DQB1*06:03	26.40	**59.20**	**1.2573**	2.6273	**0.1503**	2.1822
HLA-DQA1*01:02-DQB1*02:02	13.00	**28.75**	**0.1723**	1.3248	**0.0073**	0.0094
HLA-DQA1*01:02-DQB1*03:01	53.00	**99.00**	**0.3111**	1.5251	**0.0083**	0.0552
HLA-DQA1*01:02-DQB1*03:02	6.50	**8.50**	**0.3849**	0.8472	**0.0073**	0.0078
HLA-DQA1*01:02-DQB1*05:01	27.33	**43.33**	2.2276	**0.8075**	0.1011	**0.0092**
HLA-DQA1*01:02-DQB1*05:02	**512.20**	368.00	2.7410	**1.1782**	0.6769	**0.0513**
HLA-DQA1*01:02-DQB1*06:02	550.88	**551.88**	4.1209	**3.8584**	1.9511	**1.5487**
HLA-DQA1*01:02-DQB1*06:04	**1548.50**	1520.00	**1.8068**	1.9438	**0.3963**	0.4895
HLA-DQA1*01:03-DQB1*03:02	23.50	**54.50**	**0.3737**	0.9327	**0.0071**	0.0091
HLA-DQA1*01:03-DQB1*05:01	6.60	**12.80**	0.6552	**0.3573**	**0.0140**	0.0165
HLA-DQA1*01:03-DQB1*05:03	**2.00**	1.50	4.2220	**0.0102**	4.2220	**0.0102**
HLA-DQA1*01:03-DQB1*06:03	**353.10**	325.60	**2.6588**	3.0594	**0.5745**	0.5782
HLA-DQA1*01:04-DQB1*05:03	**549.00**	406.33	2.8193	**0.7154**	**1.7952**	0.2739
HLA-DQA1*01:04-DQB1*06:03	3.00	**3.50**	**3.2792**	3.5277	4.1573	**2.3288**
HLA-DQA1*01:05-DQB1*05:01	**89.00**	27.00	2.5292	**1.4787**	1.2150	**0.0631**
HLA-DQA1*01:05-DQB1*06:03	27.00	**51.00**	**0.6273**	0.9981	**0.0100**	0.0225
HLA-DQA1*01:10-DQB1*02:02	6.00	**32.00**	**0.0063**	0.0663	**0.0067**	0.0073

HLA-DQA1*01:10-DQB1*05:01	17.00	**22.00**	**0.0983**	0.9877	**0.0077**	0.0085
HLA-DQA1*01:10-DQB1*06:03	**205.00**	189.00	**1.9616**	1.9742	0.5409	**0.4139**
HLA-DQA1*02:01-DQB1*02:01	**75.67**	73.33	**2.2314**	2.3259	0.5515	**0.3520**
HLA-DQA1*02:01-DQB1*02:02	1188.91	**1207.82**	3.4357	**2.9958**	1.1275	**1.1813**
HLA-DQA1*02:01-DQB1*03:01	8.00	**10.50**	6.1621	**4.9487**	**2.2676**	2.6416
HLA-DQA1*02:01-DQB1*03:03	**868.00**	711.00	3.8598	**2.8138**	2.0423	**0.9332**
HLA-DQA1*02:01-DQB1*05:01	11.00	**24.00**	**0.0094**	1.5804	**0.0060**	0.0080
HLA-DQA1*02:01-DQB1*05:02	**18.33**	15.67	**0.0088**	2.0617	**0.0059**	0.0083
HLA-DQA1*02:01-DQB1*06:02	6.00	**7.00**	1.0585	**0.0980**	0.0355	**0.0091**
HLA-DQA1*02:01-DQB1*06:03	15.00	15.00	**0.0088**	0.2958	0.0085	**0.0072**
HLA-DQA1*03:01-DQB1*02:01	**14.67**	14.33	3.0440	**2.0955**	**0.0953**	0.9899
HLA-DQA1*03:01-DQB1*03:01	**54.67**	16.83	3.5246	**0.7732**	1.1628	**0.0092**
HLA-DQA1*03:01-DQB1*03:02	**152.86**	92.64	4.2735	**1.5354**	2.0231	**0.1016**
HLA-DQA1*03:01-DQB1*06:02	**4.00**	1.00	2.3313	**0.4498**	**0.2132**	0.4498
HLA-DQA1*03:01-DQB1*06:03	8.00	**10.00**	0.6198	**0.0115**	0.0270	**0.0088**
HLA-DQA1*03:02-DQB1*03:03	**1565.00**	1003.00	3.3282	**2.6386**	1.4368	**0.6301**
HLA-DQA1*03:03-DQB1*02:02	**48.33**	15.00	0.8371	**0.1817**	**0.0096**	0.0109
HLA-DQA1*03:03-DQB1*03:01	**593.25**	389.50	3.0677	**1.0865**	0.8015	**0.0512**
HLA-DQA1*03:03-DQB1*04:02	**82.00**	17.00	**0.8113**	1.2414	**0.0109**	0.0204
HLA-DQA1*03:03-DQB1*05:01	**9.00**	4.00	**0.0143**	0.0253	**0.0085**	0.0258
HLA-DQA1*04:01-DQB1*03:01	96.67	**371.00**	**1.4585**	3.5739	**0.0179**	1.5131
HLA-DQA1*04:01-DQB1*04:02	**967.60**	801.80	3.0511	**2.3695**	1.4733	**0.5541**
HLA-DQA1*05:01-DQB1*02:01	344.00	**347.79**	**2.2603**	2.6852	**0.2518**	0.3707
HLA-DQA1*05:01-DQB1*03:01	1008.50	1008.50	4.6531	**4.3624**	1.7700	**1.6834**
HLA-DQA1*05:01-DQB1*03:02	19.67	**22.00**	3.5080	**2.4423**	**0.2787**	0.5770
HLA-DQA1*05:01-DQB1*05:01	2.00	**4.00**	**0.0071**	0.0238	**0.0070**	0.0082

HLA-DQA1*05:05-DQB1*02:02	**64.00**	62.40	2.4021	**2.1639**	**0.1911**	0.3171
HLA-DQA1*05:05-DQB1*03:01	577.55	**594.90**	**3.3196**	3.8926	**1.4065**	1.8265
HLA-DQA1*05:05-DQB1*03:02	18.67	**25.33**	**1.2711**	1.5989	0.0884	**0.0495**
HLA-DQA1*05:05-DQB1*05:02	**24.00**	9.00	0.0093	**0.0068**	**0.0064**	0.0071

Supplementary table 5: Overview of consistency analysis in the methods with (w_Saghar) and without (wo_Saghar) the novel data. The molecules are sorted in descending order by the difference in mean consistency. Further, the metrics are calculated on the peptide sets used in the consistency analysis, with the union of identified trash peptides removed. As such, the metrics regarding differences in percentile ranks may not correspond one-to-one with the values in supplementary table 4.

Molecules	Difference in mean consistency (w_Saghar - wo_Saghar)	Number of cell lines with at least 20 peptides in both methods	Number of cell lines with larger peptide count in w_Saghar	Number of cell lines with lower mean %rank in w_Saghar	Fraction of cell lines with either larger peptide count or lower mean %rank in w_Saghar
HLA-DQA1*03:01-DQB1*03:02	**0.148**	**6**	**6**	**1**	**1.000**
HLA-DQA1*03:01-DQB1*03:01	**0.147**	**2**	**2**	**0**	**1.000**
HLA-DQA1*03:03-DQB1*03:01	**0.119**	**5**	**5**	**1**	**1.000**
HLA-DQA1*01:01-DQB1*05:01	**0.079**	**10**	**10**	**3**	**1.000**
HLA-DQA1*04:01-DQB1*04:02	**0.039**	**5**	**3**	**2**	**0.600**
HLA-DQA1*01:02-DQB1*05:02	**0.037**	**5**	**5**	**1**	**1.000**
HLA-DQA1*02:01-DQB1*03:03	**0.035**	**2**	**2**	**1**	**1.000**
HLA-DQA1*01:10-DQB1*06:03	**0.034**	**2**	**2**	**1**	**1.000**
HLA-DQA1*05:05-DQB1*03:01	**0.027**	**16**	**2**	**8**	**0.625**
HLA-DQA1*01:02-DQB1*06:02	**0.020**	**7**	**3**	**1**	**0.429**
HLA-DQA1*02:01-DQB1*02:01	**0.017**	**3**	**2**	**1**	**0.667**
HLA-DQA1*05:01-DQB1*02:01	**0.016**	**12**	**4**	**9**	**0.917**
HLA-DQA1*05:05-DQB1*02:02	**0.014**	**3**	**0**	**2**	**0.667**
HLA-DQA1*02:01-DQB1*02:02	**0.010**	**9**	**1**	**8**	**1.000**
HLA-DQA1*05:05-DQB1*03:02	**0.005**	**2**	**1**	**1**	**1.000**
HLA-DQA1*05:01-DQB1*03:01	**0.002**	**4**	**3**	**0**	**0.750**
HLA-DQA1*01:02-DQB1*06:04	-0.002	2	1	2	1.000
HLA-DQA1*01:03-DQB1*03:02	-0.008	2	0	2	1.000
HLA-DQA1*01:03-DQB1*06:03	-0.009	7	5	4	0.857
HLA-DQA1*01:01-DQB1*06:03	-0.027	2	0	2	1.000
HLA-DQA1*04:01-DQB1*03:01	-0.031	3	0	3	1.000

Supplementary table 6: DQ molecules covered by the method including the novel data. The highlighted molecules are covered by a peptide count of at least 100, while the remaining molecules have a distance of at most 0.025 to one of the highlighted molecules.

HLA-DQA1*01:01-DQB1*05:01	HLA-DQA1*01:01-DQB1*05:02	HLA-DQA1*03:01-DQB1*02:02	HLA-DQA1*05:01-DQB1*03:19
HLA-DQA1*01:01-DQB1*06:03	HLA-DQA1*01:01-DQB1*05:03	HLA-DQA1*03:01-DQB1*03:03	HLA-DQA1*05:01-DQB1*03:22
HLA-DQA1*01:02-DQB1*05:02	HLA-DQA1*01:01-DQB1*06:02	HLA-DQA1*03:01-DQB1*03:19	HLA-DQA1*05:01-DQB1*03:62
HLA-DQA1*01:02-DQB1*06:02	HLA-DQA1*01:01-DQB1*06:04	HLA-DQA1*03:01-DQB1*03:22	HLA-DQA1*05:02-DQB1*02:01
HLA-DQA1*01:02-DQB1*06:04	HLA-DQA1*01:02-DQB1*05:01	HLA-DQA1*03:01-DQB1*03:62	HLA-DQA1*05:02-DQB1*02:02
HLA-DQA1*01:03-DQB1*06:03	HLA-DQA1*01:02-DQB1*05:03	HLA-DQA1*03:01-DQB1*03:87	HLA-DQA1*05:02-DQB1*03:01
HLA-DQA1*01:04-DQB1*05:03	HLA-DQA1*01:02-DQB1*05:46	HLA-DQA1*03:01-DQB1*04:01	HLA-DQA1*05:02-DQB1*03:02
HLA-DQA1*01:10-DQB1*06:03	HLA-DQA1*01:02-DQB1*06:03	HLA-DQA1*03:01-DQB1*04:02	HLA-DQA1*05:03-DQB1*02:01
HLA-DQA1*02:01-DQB1*02:01	HLA-DQA1*01:03-DQB1*05:02	HLA-DQA1*03:02-DQB1*02:01	HLA-DQA1*05:03-DQB1*02:02
HLA-DQA1*02:01-DQB1*02:02	HLA-DQA1*01:03-DQB1*06:02	HLA-DQA1*03:02-DQB1*02:02	HLA-DQA1*05:03-DQB1*03:01
HLA-DQA1*02:01-DQB1*03:03	HLA-DQA1*01:03-DQB1*06:04	HLA-DQA1*03:02-DQB1*03:01	HLA-DQA1*05:03-DQB1*03:02
HLA-DQA1*03:01-DQB1*03:01	HLA-DQA1*01:04-DQB1*05:01	HLA-DQA1*03:02-DQB1*03:02	HLA-DQA1*05:03-DQB1*03:22
HLA-DQA1*03:01-DQB1*03:02	HLA-DQA1*01:04-DQB1*05:02	HLA-DQA1*03:02-DQB1*03:22	HLA-DQA1*05:05-DQB1*02:01
HLA-DQA1*03:02-DQB1*03:03	HLA-DQA1*01:04-DQB1*06:02	HLA-DQA1*03:02-DQB1*04:01	HLA-DQA1*05:05-DQB1*03:22
HLA-DQA1*03:03-DQB1*02:02	HLA-DQA1*01:04-DQB1*06:03	HLA-DQA1*03:02-DQB1*04:02	HLA-DQA1*05:06-DQB1*03:01
HLA-DQA1*03:03-DQB1*03:01	HLA-DQA1*01:04-DQB1*06:04	HLA-DQA1*03:03-DQB1*02:01	HLA-DQA1*05:06-DQB1*03:02
HLA-DQA1*03:03-DQB1*04:02	HLA-DQA1*01:05-DQB1*05:01	HLA-DQA1*03:03-DQB1*03:02	HLA-DQA1*05:08-DQB1*02:01
HLA-DQA1*04:01-DQB1*03:01	HLA-DQA1*01:05-DQB1*05:02	HLA-DQA1*03:03-DQB1*03:03	HLA-DQA1*05:08-DQB1*03:01
HLA-DQA1*04:01-DQB1*04:02	HLA-DQA1*01:05-DQB1*05:03	HLA-DQA1*03:03-DQB1*03:22	HLA-DQA1*05:08-DQB1*03:02
HLA-DQA1*05:01-DQB1*02:01	HLA-DQA1*01:05-DQB1*06:02	HLA-DQA1*03:03-DQB1*04:01	HLA-DQA1*06:01-DQB1*03:01
HLA-DQA1*05:01-DQB1*03:01	HLA-DQA1*01:05-DQB1*06:03	HLA-DQA1*04:01-DQB1*03:22	HLA-DQA1*06:01-DQB1*03:22
HLA-DQA1*05:05-DQB1*02:02	HLA-DQA1*01:05-DQB1*06:04	HLA-DQA1*04:01-DQB1*04:01	HLA-DQA1*06:01-DQB1*04:01
HLA-DQA1*05:05-DQB1*03:01	HLA-DQA1*02:01-DQB1*03:87	HLA-DQA1*05:01-DQB1*02:02	HLA-DQA1*06:01-DQB1*04:02
HLA-DQA1*05:05-DQB1*03:02	HLA-DQA1*03:01-DQB1*02:01	HLA-DQA1*05:01-DQB1*03:02	

Supplementary table 7: Prevalent DQ molecules with distance greater than 0.025 to molecules covered by the method including the novel data.

HLA-DQA1*01:01-DQB1*05:04	HLA-DQA1*01:04-DQB1*06:09	HLA-DQA1*04:01-DQB1*02:01	HLA-DQA1*05:03-DQB1*03:03
HLA-DQA1*01:01-DQB1*06:01	HLA-DQA1*01:05-DQB1*06:01	HLA-DQA1*04:01-DQB1*02:02	HLA-DQA1*05:03-DQB1*03:04
HLA-DQA1*01:01-DQB1*06:05	HLA-DQA1*02:01-DQB1*03:01	HLA-DQA1*04:01-DQB1*03:02	HLA-DQA1*05:03-DQB1*03:05
HLA-DQA1*01:01-DQB1*06:09	HLA-DQA1*02:01-DQB1*03:02	HLA-DQA1*04:01-DQB1*03:03	HLA-DQA1*05:03-DQB1*04:01
HLA-DQA1*01:02-DQB1*05:04	HLA-DQA1*02:01-DQB1*03:04	HLA-DQA1*04:01-DQB1*03:04	HLA-DQA1*05:03-DQB1*04:02
HLA-DQA1*01:02-DQB1*06:01	HLA-DQA1*02:01-DQB1*03:05	HLA-DQA1*04:01-DQB1*03:05	HLA-DQA1*05:05-DQB1*03:03
HLA-DQA1*01:02-DQB1*06:05	HLA-DQA1*02:01-DQB1*03:19	HLA-DQA1*05:01-DQB1*02:03	HLA-DQA1*05:05-DQB1*03:04
HLA-DQA1*01:02-DQB1*06:09	HLA-DQA1*02:01-DQB1*03:22	HLA-DQA1*05:01-DQB1*03:03	HLA-DQA1*05:05-DQB1*03:05
HLA-DQA1*01:03-DQB1*05:01	HLA-DQA1*02:01-DQB1*03:62	HLA-DQA1*05:01-DQB1*03:04	HLA-DQA1*05:05-DQB1*04:01
HLA-DQA1*01:03-DQB1*05:03	HLA-DQA1*02:01-DQB1*04:01	HLA-DQA1*05:01-DQB1*03:05	HLA-DQA1*05:05-DQB1*04:02
HLA-DQA1*01:03-DQB1*05:04	HLA-DQA1*02:01-DQB1*04:02	HLA-DQA1*05:01-DQB1*03:87	HLA-DQA1*05:08-DQB1*03:03
HLA-DQA1*01:03-DQB1*06:01	HLA-DQA1*03:01-DQB1*03:04	HLA-DQA1*05:01-DQB1*04:01	HLA-DQA1*06:01-DQB1*02:01
HLA-DQA1*01:03-DQB1*06:05	HLA-DQA1*03:01-DQB1*03:05	HLA-DQA1*05:01-DQB1*04:02	HLA-DQA1*06:01-DQB1*02:02
HLA-DQA1*01:03-DQB1*06:09	HLA-DQA1*03:02-DQB1*03:04	HLA-DQA1*05:02-DQB1*03:03	HLA-DQA1*06:01-DQB1*03:02
HLA-DQA1*01:04-DQB1*06:01	HLA-DQA1*03:02-DQB1*03:05	HLA-DQA1*05:02-DQB1*04:02	HLA-DQA1*06:01-DQB1*03:03