communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-04749-7 OPEN

Machine learning reveals limited contribution of trans-only encoded variants to the HLA-DQ immunopeptidome

Jonas Birkelund Nilsson [1,6], Saghar Kaabinejadian [2,3,6], Hooman Yari [3], Bjoern Peters[4], Carolina Barra [1], Loren Gragert [5], William Hildebrand[3] & Morten Nielsen [1]✉

Human leukocyte antigen (HLA) class II antigen presentation is key for controlling and triggering T cell immune responses. HLA-DQ molecules, which are believed to play a major role in autoimmune diseases, are heterodimers that can be formed as both cis and trans variants depending on whether the α- and β-chains are encoded on the same (cis) or opposite (trans) chromosomes. So far, limited progress has been made for predicting HLA-DQ antigen presentation. In addition, the contribution of trans-only variants (i.e. variants not observed in the population 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 mass spectrometry immunopeptidomics data. The analysis demonstrates 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 reveals a limited to no contribution to the overall HLA-DQ immunopeptidome. In conclusion, this study furthers our understanding of HLA-DQ specificities and casts light on the relative role of cis versus trans-only HLA-DQ variants in the HLA class II antigen presentation space. The developed method, NetMHCIIpan-4.2, is available at https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.2.

[1] Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark. [2] Pure MHC, LLC, Oklahoma City, OK, USA. [3] Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [4] Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, La Jolla, CA 92037 California, USA. [5] Department of Pathology and Laboratory Medicine, Tulane University School of Medicine, New Orleans, LA 70112, USA. [6] These authors contributed equally: Jonas Birkelund Nilsson, Saghar Kaabinejadian. ✉email: morni@dtu.dk

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[1]. 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[2]. In many autoimmune diseases, HLA class II genes are major genetic susceptibility factors[1,3] 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[3]. These genetic and functional studies have indicated that both HLA-DR and DQ alleles are associated with the risk of T1D[3,4]. In addition, the associated DR-DQ haplotypes demonstrate a risk hierarchy, ranging from highly predisposing to highly protective[4]. 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[5].

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[6], thus necessitate a more thorough and systematic characterization of antigen presentation by HLA-DQ molecules to enable study of their function. Even though the field is moving forward rapidly[7], so far peptide binding motifs of only a limited number of HLA-DQ molecules have been exhaustively studied[8–10]. 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[11]. 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[11,12]. For example, DQA1*01 has only been detected to form stable heterodimers with DQB1*05 and 06 alleles. Likewise, the DQA1*02, 03, 04, 05, and 06 alleles form stable heterodimers only with the DQB1*02, 03, and 04[12–14].

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[3,11].

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 opposite (*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[11,15]. 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[16,17]. Hereafter, we refer to all stable DQα- and β-chain combinations mentioned above as cis, and the rest which includes any combination that has not been detected or reported as cis encoded will be referred to as "trans-only".

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[13]. Here, the assumption is that all observed DQ haplotypes, by natural selection, are able to form stable and functional cis and trans-encoded molecules. However, the role of trans-only encoded variants in antigen presentation and their contribution 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[18]). Current state-of-the-art prediction methods include NetMHCIIpan[19], 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[2], and development of different machine-learning prediction models to identify the rules of peptide–HLA binding[20,21]. However, experimental results indicate binding affinity (BA) to be a relatively weak correlate of antigen processing and presentation by HLA molecules[22]. 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)[2,20,23,24]. 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[25,26]. 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[27–29].

HLA class II binding predictions have been widely used to identify epitope candidates in infectious, cancer and autoimmune diseases[30]. 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 α- and β-chains in order to perform binding affinity experiments, and generation of large EL datasets have proven to be challenging. The latter mostly due to lack of application of HLA-DQ specific antibodies in large scale MS-immunopeptidomics experiments resulting in limited yield in the HLA-DQ purification process.

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 group 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.

Results

For the study, immunopeptidome data for 14 different HLA-DQ molecules was obtained from 16 homozygous B Lymphoblastoid Cell Lines (BLCLs) using LC-MS/MS. By using a DQ-specific antibody during the affinity purification, we were able to obtain a large dataset highly enriched in DQ peptide ligands. An overview of the cell lines' peptide counts, DQ HLA types and peptide length distributions is shown in Fig. 1. Overall, the data contains a total of 39,334 peptide ligands, with 14- and 15-mers being most prevalent. After enriching the novel data with random natural peptides assigned as negatives (see materials and methods), we combined it with the data used to train the NetMHCIIpan-4.1 prediction method, yielding a large dataset of eluted HLA class II ligands. From this, we set out to address three essential issues related to HLA-DQ, 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 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[31] 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 for direct comparison of the impact of the novel data one without (termed wo_Saghar_DQ). The models were then evaluated using cross-validation on a per-molecule basis within four different subsets of all the HLA class II molecules in the training 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 2 displays the result of this experiment and demonstrates that incorporation of the novel DQ data resulted in a significant performance gain for DQ as expected ($p = 0.011$ for all metrics, $n = 44$ molecules, one-tailed binomial test without ties). However, from these results it is apparent that the performance for DQ remains lower compared to that of non-DQ molecules. We assumed this to be a result of the DQ performance being calculated from a mix of both the novel data and the older

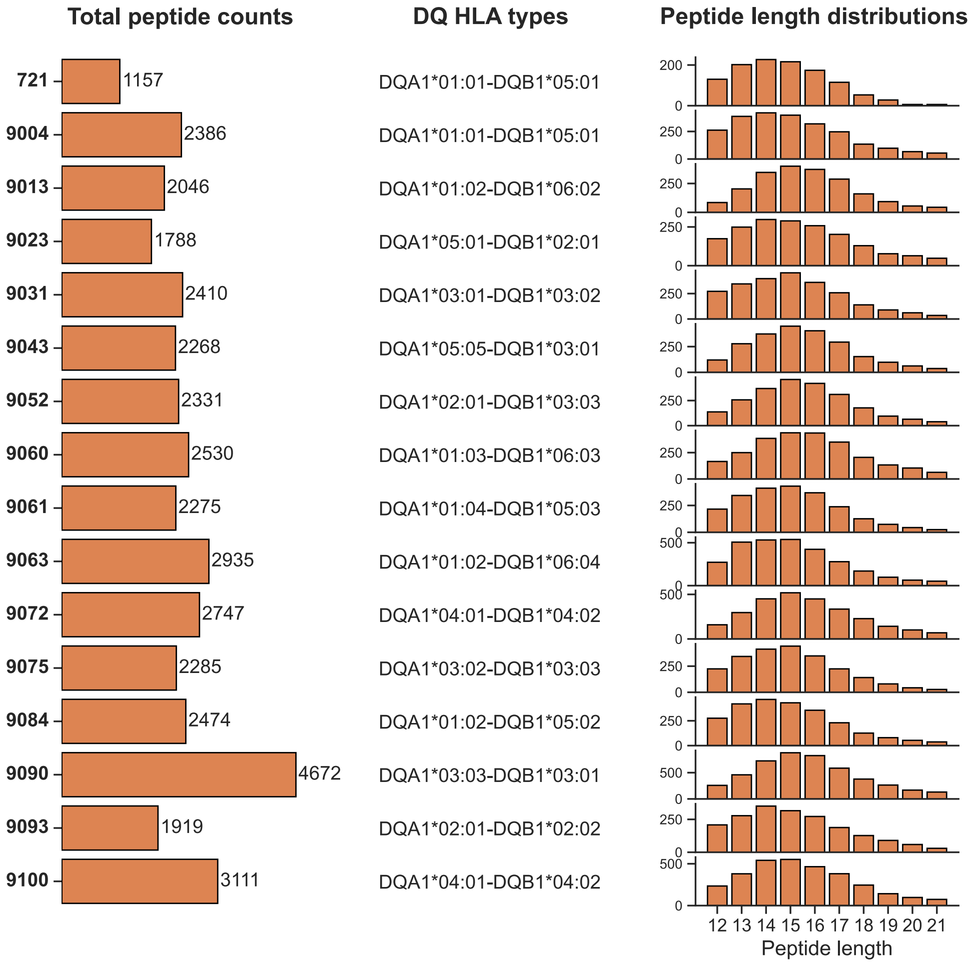

Fig. 1 Overview of the novel immunopeptidomics data. Each row corresponds to a dataset from a given DQ-homozygous cell line. Left panel: Bar plot of overall peptide counts. The numbers on the left correspond to the cell line IDs. Middle panel: DQ HLA types of the cell lines. Right panel: Peptide length distributions.

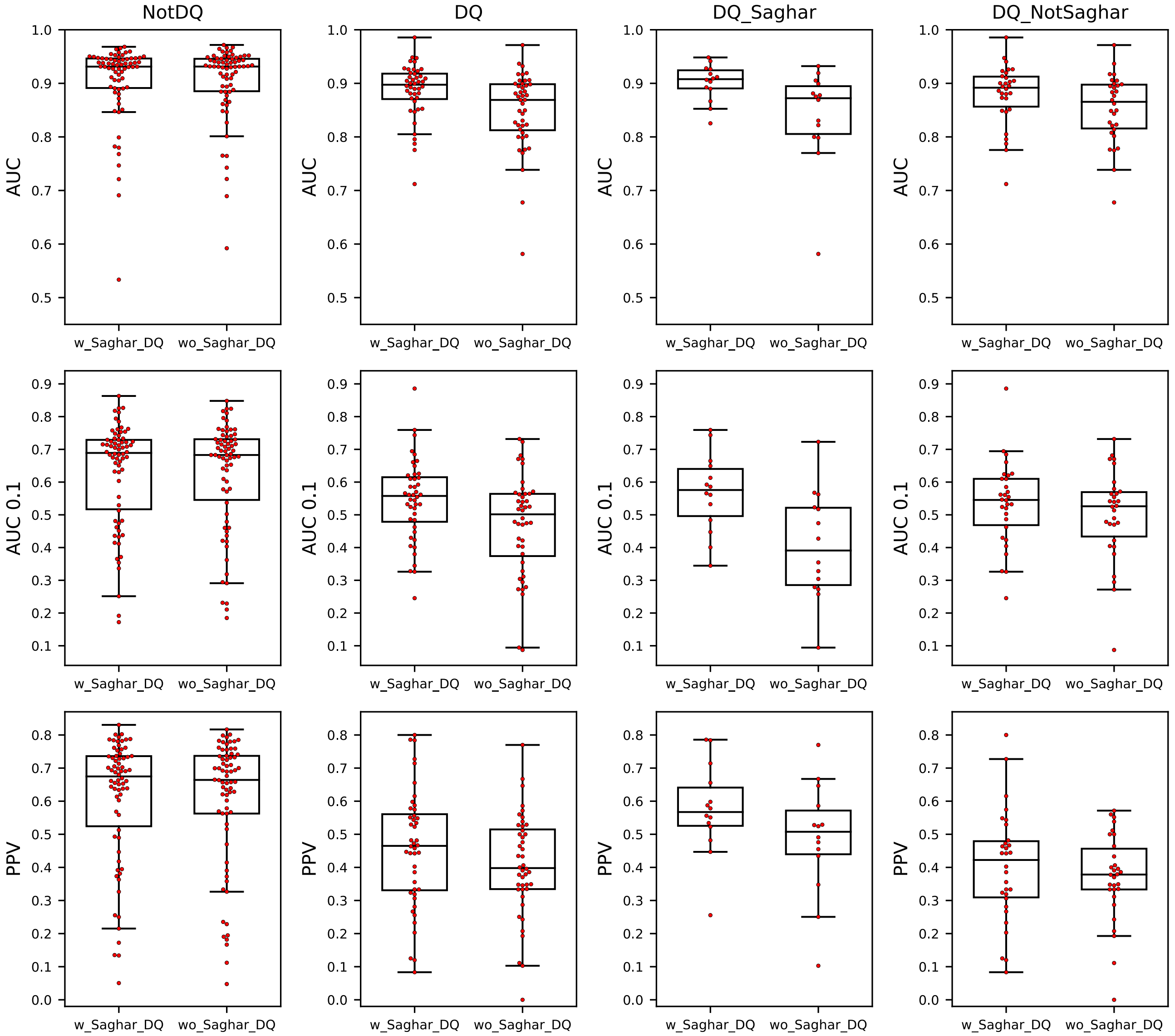

Fig. 2 AUC, AUC 0.1 and PPV predictive performance for the models trained with (w_Saghar_DQ) and without (wo_Saghar_DQ) the novel data. Each point is the performance metric for a unique HLA class II molecule. For details on the performance metrics refer to materials and methods. The columns correspond to four different subsets of HLA molecules, namely all non-HLA-DQ molecules (NotDQ, $n = 70$), all DQ molecules (DQ, $n = 44$), DQ molecules in the novel data set (DQ_Saghar, $n = 14$), and DQ molecules not present in the novel data (DQ_NotSaghar, $n = 30$). Each boxplot shows the median inside the interquartile range (IQR) between the upper and lower quartiles, with whiskers extending to at most 1.5 times the IQR.

NetMHCIIpan-4.1 training data. To demonstrate this, we evaluated the performance on the DQ_Saghar molecules limited to the novel data only. The result of this is shown in Fig. 3 and demonstrates that when focusing only on the novel data, the performance of DQ reaches a level comparable to that of non-DQ, with a significant gain in terms of PPV ($t = 1.19$, $p = 0.24$ for AUC, $t = 0.21$, $p = 0.83$ for AUC 0.1 and $t = 2.69$, $p = 0.009$ for PPV, $n = 14$ DQ molecules and $n = 70$ non-DQ molecules, two-sided t-tests). 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.

We next looked at the differences in peptides assigned to HLA-DQ molecules between the two methods across all samples. Here, we considered all peptides which were assigned to DQ with percentile rank <20 (i.e. as non-trash) in at least one of the methods[23]. Overall, the two methods share a high degree of overlap in the peptides assigned to DQ (60,959 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 1 for an overview of the peptide migrations). This suggests that at least part of the improved predictive performance of the novel model originates from an improved motif deconvolution.

To further quantify this, we show the mean consistency value per HLA molecule in the four molecule subsets in Supplementary Fig. 1. 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 (for details refer to materials and methods). Based on this analysis, an overall improved consistency is observed for the model trained with the novel DQ data ($p < 0.02$ in all cases except for the

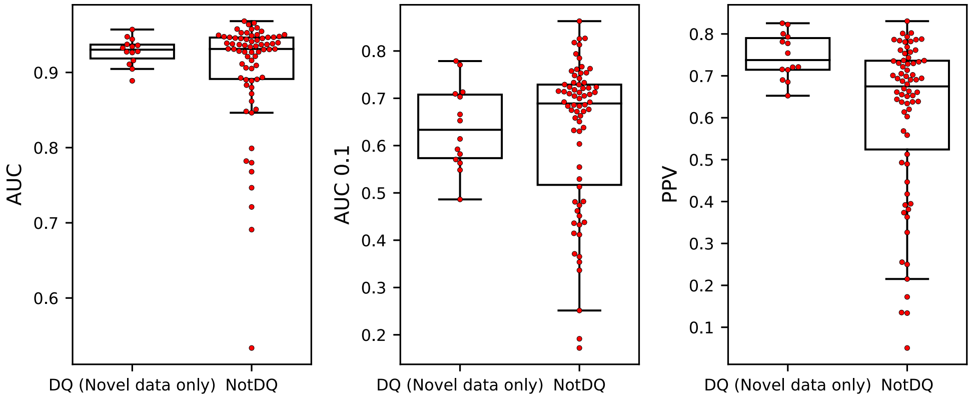

Fig. 3 Performance of the model trained including the novel data, evaluated on both the novel data alone restricted to DQ, as well as on non-DQ including the full dataset. Each point is the performance metric for an HLA class II molecule. Each boxplot shows the median inside the interquartile range (IQR) between the upper and lower quartiles, with whiskers extending to at most 1.5 times the IQR.

DQ_NotSaghar subset, one-tailed binomial test without ties). The consistency analysis for an example molecule contained in the novel data (DQA1*03:01-DQB1*03:02) is shown in Supplementary Fig. 2, illustrating that in most cases the improved motif consistency is caused by an increased peptide count across samples (see Supplementary Tables 2 and 3).

Furthermore, HLA-DQ binding motifs obtained by motif deconvolution of the novel MS data were visualized, along with sequence motifs based on predicted binders, in Supplementary Fig. 3. Here, the logos obtained by motif deconvolution are in most cases very similar when comparing the models trained with and without the novel data. 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 quantify these results, correlations between the deconvoluted and predicted logos for each method were calculated (Supplementary Fig. 4). This analysis showed significantly higher correlation for the method including the novel data ($p = 0.011$, $n = 16$ logo pairs, one-tailed binomial test without ties), indicating a highly consistent correspondence between the identified and predicted binding motifs.

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 trained including the novel data using peptide context 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. In line with what has been demonstrated earlier[2,27,31], the results of this comparison (Supplementary Fig. 5) demonstrated that the model trained including context significantly outperformed the model trained without context in all performance metrics and data subsets (the only exception being the DQ_NotSaghar subset). However, 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. In DQ-heterozygous cell lines, four possible α–β chain pairings can in principle be observed. For so-called cis-heterodimers, the α- and β-chain are expressed on the same chromosome and can thus be observed in haplotype sequencing. DQ molecules formed by pairing α- and β-chains between chromosomes are called trans-heterodimers. Some α–β pairings have not been observed as cis encoded (based on large HLA-haplotype sequencing population studies) and are thus here referred to as “trans-only” combinations. To assess the relative contribution of cis and trans-only DQ heterodimers in shaping the immunopeptidome, we investigated the distribution of peptides assigned to cis versus trans-only encoded DQ molecules across DQ-heterozygous datasets for the two models. Here, only datasets with at least 100 DQ-annotated peptides excluding trash in both methods were considered (for an overview of the datasets used in this analysis, refer to Supplementary Table 4). 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 across four categories were then investigated. These categories are all cis variants, cis-SA (cis variants part of the single-allelic DQ training data), cis-MA (cis variants part of the multi-allelic DQ training data), and trans-only variants.

The result of this analysis is shown in Fig. 4a for the two models and indicates that for the method including the novel data, trans-only molecules consistently cover a small proportion of the DQ annotations in each cell line. On the other hand, the cis molecules have generally high contribution, with the cis-SA molecules having the largest contribution. However, the cis-MA molecules were also found to have significantly larger contribution compared to the trans-only molecules in the model including the novel data ($t = 3.07$, $p = 0.005$, $n = 18$ cis-MA molecules and $n = 12$ trans-only molecules, two-sided t-test). Similar results were found when extending the cis-SA category to include cis-MA molecules with the same pseudo-sequence as a cis-SA molecule (Supplementary Fig. 6). Further, an overall higher contribution of trans-only molecules to the DQ peptide annotations was observed for the model trained without the novel data ($t = 2.1$, $p = 0.03$, $n = 12$ molecules, paired one-sided t-test). These results are striking, as they indicate that the motif deconvolution in the model including the novel data is not solely driven by the cis-SA molecules, but rather by an overall preference for cis-encoded variants compared to trans-only variants (see Supplementary Figs. 7 and 8).

To further investigate this, the DQ motif deconvolution of the two models for the Racle__TIL1 dataset is shown in Fig. 4b. Here, the model trained without the novel data assigns a large

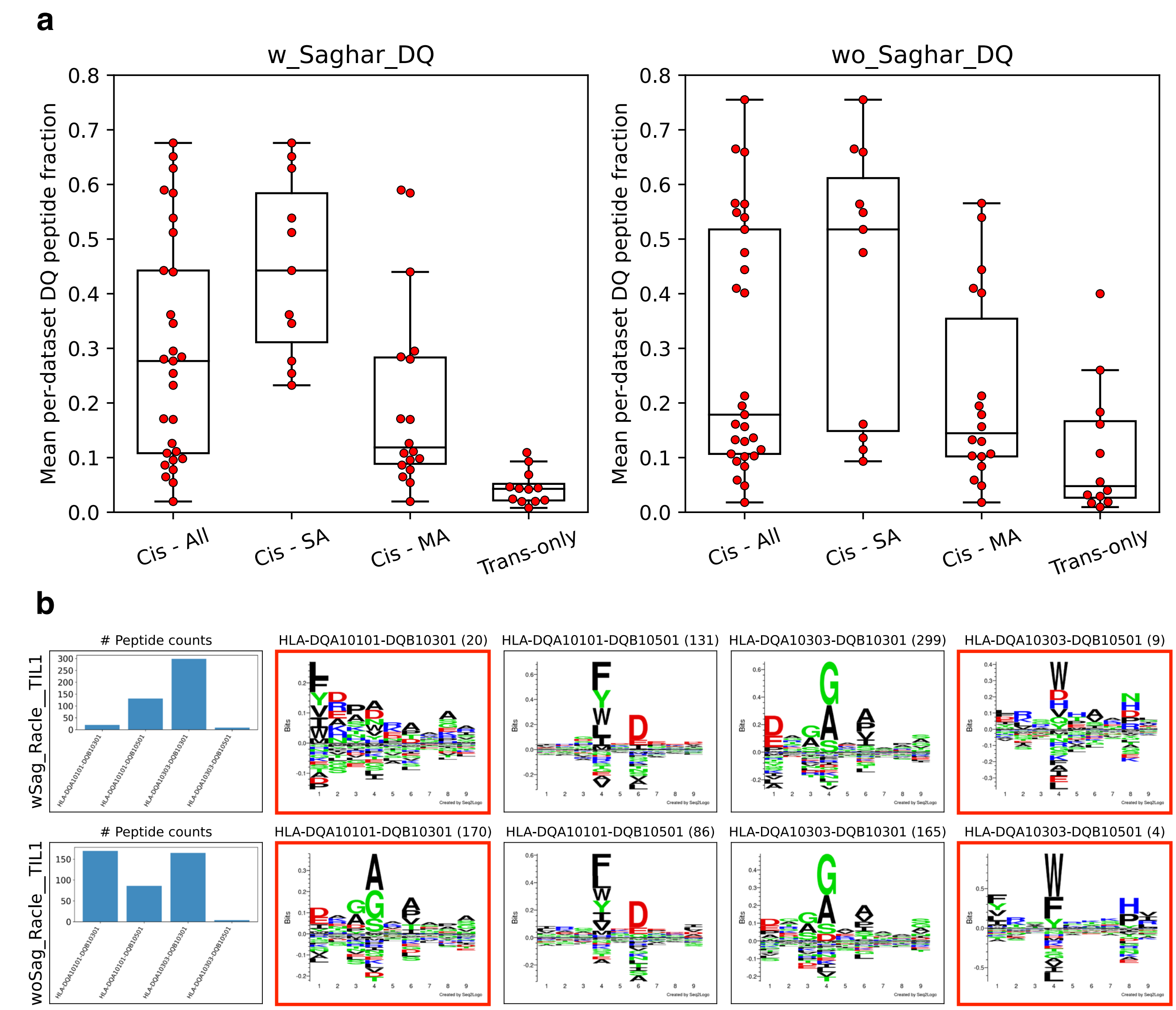

Fig. 4 Contribution of cis and trans-only DQ variants in DQ-heterozygous datasets. a Peptide-count contribution of cis and trans-only molecules in the methods with (w_Saghar_DQ) and without (wo_Saghar_DQ) the novel data. Each point shows the mean per-dataset peptide fraction for a given DQ molecule. For each method, trans-only molecules are shown in one boxplot (*n* = 12), while cis molecules are shown in three categories, namely all cis molecules (Cis–All, *n* = 29), cis molecules found in the DQ-SA training data (Cis–SA, *n* = 11), and cis molecules only found in the DQ-MA training data (Cis–MA, *n* = 18). Each boxplot shows the median inside the IQR between the upper and lower quartiles, with whiskers extending to at most 1.5 times the IQR. **b** DQ motif deconvolution for the Racle__TIL1 dataset. The rows correspond to the methods trained with (wSag) and without (woSag) the novel data, respectively. Peptide counts (excluding trash peptides) are displayed in parenthesis in the logo plot titles. Trans-only molecules are highlighted in red frames.

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[12,13]. 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[12,13].

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 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 have limited to no contribution to the total HLA-DQ immunopeptidome. However, it is important to underline that the predictions are highly influenced by the SA training data (illustrated by the dominant contribution of the cis-SA category). As such, we cannot rule out completely that the low number of annotations towards trans-only heterodimers may be impacted by the lack of SA training data for these molecules or a lower sequence similarity to the cis-SA molecules compared to that of the cis-MA molecules.

Difference in peptide length distributions of DR and DQ. When we compared the length distribution of DQ peptide ligands in the novel data with HLA-DR restricted peptides that were purified from the same set of BLCLs[23], it was revealed that the DQ ligands were in general shorter than the DR ligands (see Supplementary Fig. 9). By comparing the per-molecule median peptide lengths for the two loci, a significant difference was found ($t = 2.4$, $p < 0.03$, $n = 17$ DR molecules and $n = 14$ DQ molecules, two-sided t-test), with DR and DQ having average peptide length medians of 15.41 and 14.93, respectively. This analysis indicates that HLA-DQ molecules generally bind shorter peptides compared to HLA-DR. Moreover, in contrast to HLA-DQ alleles that are more consistent in their peptide length preferences, various HLA-DR molecules show subtle differences in their length preferences[23]. For example, HLA-DR*07:01, 09:01 and 14:01 show a preference for shorter peptides (14 mers) while the majority of DR alleles follow the common class II length preference (15 mer).

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.

Given the different sets of molecules covered by the two methods, we wanted to estimate each method's coverage when considering the entire DQ specificity space. As such, 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. The threshold of 0.025 was chosen based on the distance at which the model trained without the novel data could reach optimal performance on molecules not part of the method's DQ-SA training data (see Supplementary Fig. 10). Note, also, that 0.025 is a conservative distance threshold, and that we expect the model to maintain accuracy also for molecules falling beyond this value[32].

From this analysis, a significant gain in extended coverage was found ($\chi^2 = 4.73$, $p < 0.03$, $n = 154$ molecules, 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 5 and 6 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, $t = 2.69$, $p = 0.0083$, $n = 94$ covered molecules and $n = 60$ non-covered molecules, two-sided student t-test). These results suggest that the 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[33]. 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. Moreover, a few subclusters of non-covered molecules were observed (highlighted by motifs in red frames), which were found to correspond almost one-to-one with the non-covered clusters in a phylogenetic tree of the DQ pseudo-sequences (see Supplementary Fig. 11).

NetMHCIIpan-4.2. The model developed here including the novel DQ immunopeptidome data is made publicly available at https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.2. The method allows for prediction of HLA antigen presentation to all HLA-DQ molecules, and prediction can be made with or without context encoding.

Benchmark on independent DQ data. As a final showcase of our method's motif deconvolution power for DQ, we benchmarked our method against MixMHC2pred-2.0, another HLA class II predictor which was recently published[7]. The benchmark data was taken from Marcu et al.[34] and consists of eluted ligand data from 15 donor samples, which was enriched with random negative peptides (for more details on the benchmark data refer to materials and methods and see Supplementary Table 7 for an overview of the samples used).

We first evaluated the performance of the two methods without including peptide context information. Figure 6a shows this performance per sample on the entire data, indicating that our method significantly outperforms MixMHC2pred-2.0 on the independent dataset in all three metrics ($p < 0.02$ in all metrics, $n = 15$ samples, one-tailed binomial test without ties). Furthermore, Fig. 6b shows the performance per sample restricted to the union of peptides annotated towards DQ by either method, once again showing a significant performance gain in favor of NetMHCIIpan-4.2 ($p < 0.005$ in all metrics, $n = 15$ samples, one-tailed binomial test without ties). Repeating the benchmark including peptide context encoding also resulted in our method significantly outperforming MixMHC2pred-2.0 ($p < 0.005$ in all metrics for the entire data and $p = 3{\cdot}10^{-5}$ in all metrics for the union of DQ-annotated peptides, $n = 15$ samples, one-tailed binomial tests without ties (see Supplementary Fig. 12)). It should be noted that both methods identified a large proportion of trash peptides with percentile ranks >20 in the data (~21% and ~32% for NetMHCIIpan-4.2 and MixMHC2pred, respectively). This suggests a poor data quality in general, yielding substantially lower performance than observed in our cross-validation. The performance on this data is therefore not a true indicator of each method's predictive power. However, the overall performance

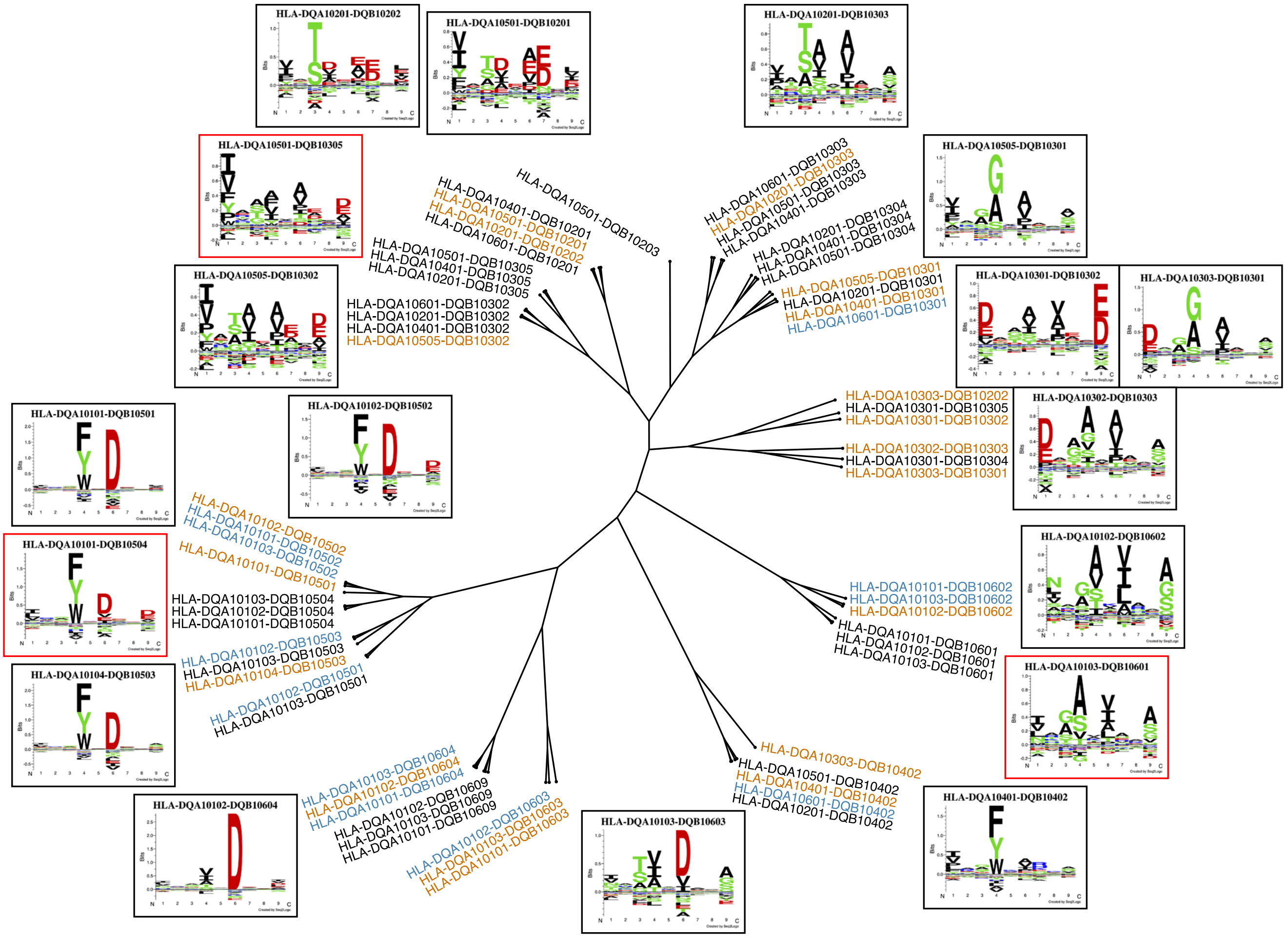

Fig. 5 HLA-DQ specificity tree. The tree is based on 61 DQ molecules including the 14 molecules described by the novel data. Orange molecules are covered by the method including the novel data with at least 100 peptides, and blue molecules are within a distance 0.025 of an orange molecule. Black molecules are non-covered (i.e. have peptide count <100 and have distance >0.025 to an orange molecule). Logos in black frames correspond to orange molecules. Logos in red frames correspond to molecules from branches with clusters of non-covered (black) molecules. 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.

gain of our method compared to MixMHC2pred-2.0 suggests that NetMHCIIpan-4.2 is more powerful in the motif deconvolution and identification of DQ ligands.

Investigating our method's motif deconvolution on the DQ-heterozygous samples, we observed that the trans-only molecules once again had limited to no contribution (see Supplementary Fig. 13a). In terms of observed cis variants found in the DQ-SA or DQ-MA training data (cis-SA and cis-MA, respectively), the cis-SA molecules had the largest contribution, with cis-MA having significantly larger contribution than the trans-only variants ($t = 4.64$, $p = 0.0002$, $n = 12$ cis-MA molecules and $n = 7$ trans-only molecules, two-sided t-test). Similar results were found when taking into account cis-MA molecules with the same pseudo-sequence as a cis-SA molecule (Supplementary Fig. 13b). This result contrasts with what was observed for MixMHC2pred, where close to an equal contribution was observed across the different molecule classes. Supplementary Figure 13c, d show the DQ motif deconvolution for the heterozygous samples from Marcu et al. 2021[34] by our method and MixMHC2pred, respectively. These motif deconvolutions overall reflect the results described above, with a very limited number of peptides assigned to trans-only variants by NetMHCIIpan-4.2, and a close to even contribution to all DQ molecules by MixMHC2pred-2.0.

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 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 2600 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

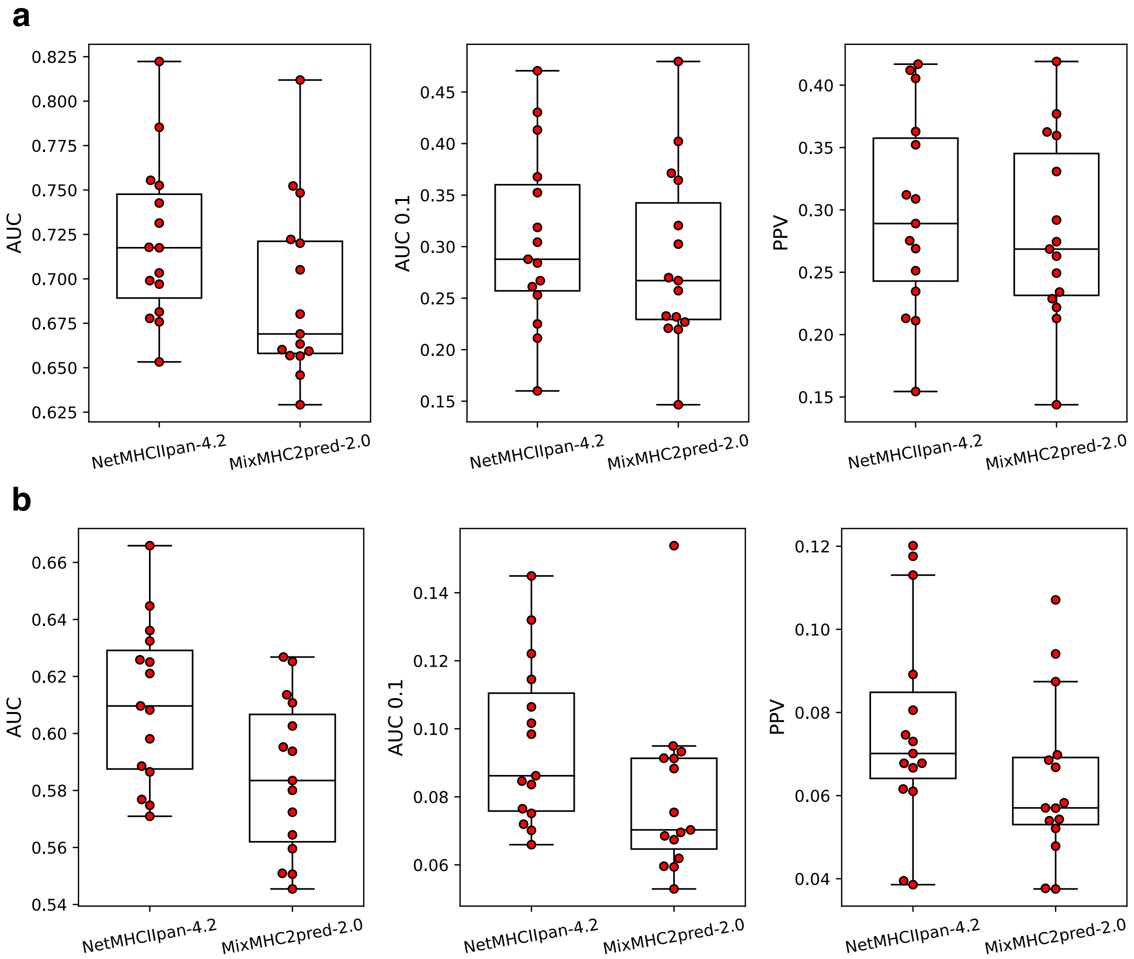

Fig. 6 Benchmark against MixMHC2pred-2.0 in terms of AUC, AUC 0.1 and PPV. Predictions were made without peptide context encoding in both methods. Each point is the performance metric for a given sample. Each boxplot ($n = 15$ samples in all cases) shows the median inside the IQR between the upper and lower quartiles, with whiskers extending to at most 1.5 times the IQR. **a** Performance per sample calculated on the entire data. **b** Performance per sample calculated on the union of DQ-annotated peptides between the two methods.

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 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 using known HLA-DQ cis-variants. One particular example here was the HLA-DQ molecule DQA1*01:01-DQB1*03:01, which 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 have minimal to no contribution to the total HLA-DQ immunopeptidome. This finding is striking since the definition of 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 inefficient assembly, lack of stability and surface expression and therefore loss of function[12,14]. 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.

As only cis-DQ variants are represented in the SA training data, we cannot rule out completely that the low number of annotations towards trans-only molecules is caused by a training data bias. This potential bias is also illustrated by the fact that for samples containing multiple cis-DQ molecules, our method consistently annotated fewer peptides to cis-variants not covered by the DQ-SA training data. Given this, it would be of great value to generate SA datasets for additional DQ molecules currently only covered by cis-MA data to uncover the true difference in peptide preferences and presentation hierarchies for these variants. Moreover, the independent MA dataset used to benchmark against MixMHC2pred was very noisy and thus did not give the best representation of each method's predictive power. As such, additional high-quality DQ-MA datasets are needed to further validate and compare the predictive power of the different methods, and to assess which method's approach to the handling of trans-only variants is better.

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 combinations can pair and form cis-variants, and all other combinations not observed as 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 heterogeneous groups and low frequency haplotypes[13].

Lastly, we demonstrated how the coverage of HLA-DQ molecules was largely increased by the models trained with the 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 (considering that *trans-only* HLA-DQ molecules have limited to no 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 what is observed for DR. We demonstrate that this to a very large degree can be attributed to the generally 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[29]). 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 total 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.

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 Data 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 4 C for 10 min. The cell pellets were immediately frozen in LN2 and stored at −80 until downstream processing[23]. 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). Immunoaffinity columns were generated by coupling 2 mg of the purified antibody to 1 mL of matrix (CNBr-activated Sepharose 4 Fast Flow, Amersham Pharmacia Biotech, Orsay, France)[23]. 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) and incubated at 4 C for 1 h on a rotary shaker. Lysates were centrifuged in an Optima XPN-80 ultracentrifuge (Beckman Coulter, IN, USA) at 4 C for 90 min (200,000 xg). Cleared supernatants were filtered using a 0.45 µm filter and were loaded on immunoaffinity columns overnight at 4 C. Columns were washed sequentially with 10 cv of wash buffers at pH:8.0[26] and were eluted with 0.2 N acetic acid. The HLA was denatured, and the peptides were isolated by adding glacial acetic acid (up to 10%) and heat (76 C for 10 min). The mixture of peptides and HLA-DQ was subjected to reverse phase high performance liquid chromatography (RP-HPLC).

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 over a 150-mm long Gemini C18 column, pore size 110 Å, particle size 5 µm (Phenomenex, Torrance, California, USA) 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. The sample was loaded on the column in a period of 18 min using a solvent system comprised of 2% solvent B at a flow rate of 120 µl/min. Then a two-segment gradient was run at 160 µl/min flow rate: 4 to 40% Solvent B for 40 min, followed by 40 to 80% Solvent B for 8 min[23]. 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 min, followed by 40 to 80% B for 7 min.

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 2400 V. Information-dependent analysis (IDA) of peptide ions was acquired based on a survey scan in the TOF-MS positive-ion mode over a range of 300 to 1250 m/z for 0.25 s. Following each survey scan, up to 22 ions with a charge state of 2–5 and intensity of at least 200 counts per second were subjected to collision-induced dissociation (CID) for tandem MS analysis (MS/MS) over a maximum period of 3.3 s. Selection of a particular ion m/z was excluded for 30 s after three initial MS/MS experiments. Dynamic collision energy was utilized to automatically adjust the collision voltage based upon ion size and charge[23]. 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) at a precursor mass error tolerance of 30 ppm and a fragment mass error tolerance of 0.02 Da. 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, deamidation, 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. This data was filtered to exclude potential HLA class I binders and other co-immunoprecipitated contaminants, resulting in a list of peptides of length 12-21[23].

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.

Our final novel data set consists of 39,334 positive and 369,313 negative peptides covering 14 unique HLA-DQ molecules. The positive peptides of this dataset are available in Supplementary Data 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 points from 177 samples/cell lines, and the BA data consist of 129,110 data points.

The data was partitioned into five subsets for cross-validated method training and evaluation using the common-motif approach[35] 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.

Model training. Models were trained using the NNAlign_MA machine learning framework[31] in a manner similar to that for NetMHCIIpan-4.0[2]. 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 backpropagation 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 the 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 was defined in both the peptide's N- and C-terminal as three residues from the source protein flanking the peptide, along with three starting residues from the peptide, all concatenated into a 12-mer amino acid sequence. For further details refer to Barra et al. 2018[27].

Performance evaluation and MHC restriction deconvolution. For MA datasets, the HLA annotation for each peptide is based on which of the HLA molecules expressed in the given cell line received the highest prediction score. To balance the differences between HLAs' prediction score distributions, percentile normalized prediction scores were generated for each molecule by ranking the prediction scores against a distribution of prediction scores of random natural peptides. As an example, if a peptide ligand receives a percentile rank score of 1, it means that 1% of the random peptides had a higher prediction score than the peptide ligand for the given HLA[19,36].

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). Each metric was calculated in a per-HLA manner from the "raw" prediction scores after HLA annotation. Further, the 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. 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.

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[2]. 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 >20 in either of the two models) was 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 from the pseudo-distance of the two molecules, i.e.

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

where s(X, Y) is the summed BLOSUM 50 similarity between the pseudo-sequences of molecule X and Y[37]. Here, each pseudo-sequence was defined from a 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[32].

Estimation of prevalent stable HLA-DQ molecules. A list of HLA-DQ α- and β-chains forming prevalent stable HLA-DQ heterodimers was constructed by first obtaining lists of DQA1 and DQB1 alleles with annotated worldwide allele frequencies. This was done by querying the allelefrequencies.net database[38] 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 1, only including molecules with a combined allele frequency >0.00005. This resulted in a list of 154 HLA-DQ molecules.

Table 1 List of DQA1-DQB1 haplotypes extracted from Creary et al. and Petersdorf et al.[13,14].

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

Due to the relatively small sample size for some populations included in the study by Creary et al.[13], the reported alleles and haplotypes may not reflect all haplotypes observed in the entire populations. Therefore, in this table only low-resolution (2 digit) DQ haplotypes were included to define the observed DQA1-DQB1 haplotypes.

Estimation of worldwide haplotype frequencies. Worldwide HLA-DQ haplotype frequencies were estimated by querying the allelefrequencies.net database[38] for high resolution DQ haplotypes in populations of size 100 and above, average across population as described above for HLA-DQ frequencies.

HLA-DQ specificity trees. 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[33] and visualized using the Iroki phylogenetic tree viewer[39].

A similar tree was constructed based on clustering of the DQ pseudo-sequences. This tree was calculated with ClustalW-2.1[40] using its phylogenetic tree function, and again visualized using the Iroki tree viewer[39].

Independent benchmark. For our benchmark against MixMHC2pred-2.0[7], an independent dataset was taken from Marcu et al.[34], which consists of eluted ligand data from 15 donor samples (listed in Supplementary Table 7). This data was processed in the same way as the training data, i.e. peptides were mapped to the human proteome to define context, and were subsequently enriched with random negative peptides. To reduce bias, peptides which were present in the EL training data of our method were not included in the benchmark. This yielded a total of 163,933 positive and 2,900,818 negative peptides covering 66 unique HLA class II molecules.

Predictions on the benchmark data were made both with and without peptide context encoding. For peptides located near the beginning or end of the source protein, missing context residues were represented by "-" and "A" in MixMHC2pred-2.0 and our method, respectively. Further, in both our method and MixMHC2pred, the HLA annotation for each peptide was based on the lowest percentile rank score reported by the given method for the HLA molecules in the given sample.

Performance was evaluated on a per-sample basis in terms of AUC, AUC 0.1, and PPV. For our method, we calculated the performance values in the same way as in the cross-validation using the 'raw' prediction scores, while for MixMHC2pred-2.0 the performance was calculated using its reported percentile rank scores.

Data visualization. Data visualizations in the manuscript figures were made in Python 3.8 using the Matplotlib library (version 3.5.1) and the seaborn library (version 0.12.0). Sequence logos were constructed using Seq2Logo-2.0[41].

Statistics and reproducibility. Statistical analyses were done in Python 3.8 using the scipy library (version 1.9.1). For each statistical test, the sample size was based on the number of samples or HLA molecules present in the data. Further, a standard significance level of 0.05 was used in each test. For the performance evaluations, the statistical tests were mainly performed using one-tailed binomial tests excluding ties. The alternative hypothesis in these tests is thus that the method trained with the novel data is more likely to perform better on a given sample or HLA molecule than the other method.

Reproducibility of our experimental and computational results was ensured by highly detailed descriptions of the experimental designs and making all relevant datasets available (see 'Data availability'). For the experimental data generation, we used two sets of different homozygous BLCLs sharing the same HLA-DQ allele to confirm reproducibility of the motifs obtained for those alleles (721.221 and IHW09004 shared the DQA1*01:01-DQB1*05:01 allele and IHW09072 and IHW9100 shared the DQA1*04:01-DQB1*04:02 allele).

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[42] partner repository with the dataset identifier PXD040860 and 10.6019/PXD040860. HLA typing for the 16 BLCLs used in the study are included in Supplementary Data 1. The novel immunopeptidomics data generated for this study is available in Supplementary Data 2. Numerical source data used to generate the main figures are included in Supplementary Data 3. The training dataset used in the study can be downloaded from the NetMHCIIpan-4.2 web server on the 'Training data sets' page, available at https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.2/.

Received: 1 October 2022; Accepted: 23 March 2023;
Published online: 21 April 2023

References

1. Rocha, N. & Neefjes, J. MHC class II molecules on the move for successful antigen presentation. *EMBO J.* **27**, 1–5 (2008).
2. Reynisson, B. et al. 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 (2020).
3. Arango, M. T. et al. HLA-DRB1 the notorious gene in the mosaic of autoimmunity. *Immunol. Res.* **65**, 82–98 (2017).
4. Erlich, H. et al. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk analysis of the type 1 diabetes genetics consortium families. *Diabetes* **57**, 1084–1092 (2008).
5. Hu, X. et al. 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 (2015).
6. Stepniak, D. et al. Large-scale characterization of natural ligands explains the unique gluten-binding properties of HLA-DQ2. *J. Immunol.* **180**, 3268–3278 (2008).
7. Racle, J. et al. Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *bioRxiv* https://doi.org/10.1101/2022.06.26.497561 (2022).
8. 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 (2014).
9. Sidney, J. et al. Divergent motifs but overlapping binding repertoires of six HLA-DQ molecules frequently expressed in the worldwide human population. *J. Immunol.* **185**, 4189–4198 (2010).
10. Vartdal, F. et al. The peptide binding motif of the disease associated HLA-DQ (α 1* 0501, β 1* 0201) molecule. *Eur. J. Immunol.* **26**, 2764–2772 (1996).
11. Tollefsen, S. et al. Structural and functional studies of trans-encoded HLA-DQ2.3 (DQA1*03:01/DQB1*02:01) protein molecule. *J. Biol. Chem.* **287**, 13611–13619 (2012).
12. Kwok, W. W., Kovats, S., Thurtle, P. & Nepom, G. T. HLA-DQ allelic polymorphisms constrain patterns of class II heterodimer formation. *J. Immunol.* **150**, 2263–2272 (1993).
13. Creary, L. E. et al. 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 (2021).
14. Petersdorf, E. W. et al. HLA-DQ heterodimers in hematopoietic cell transplantation. *Blood* **139**, 3009–3017 (2022).
15. Lundin, K. E. et al. T lymphocyte recognition of a celiac disease-associated cis- or trans-encoded HLA-DQ alpha/beta-heterodimer. *J. Immunol.* **145**, 136–139 (1990).
16. Kwok, W. W. & Nepom, G. T. Structural and functional constraints on HLA class II dimers implicated in susceptibility to insulin dependent diabetes mellitus. *Baillieres Clin. Endocrinol. Metab.* **5**, 375–393 (1991).
17. McFarland, B. J. & Beeson, C. Binding interactions between peptides and proteins of the class II Major Histocompatibility Complex. *Med Res. Rev.* **22**, 168–203 (2002).
18. Nielsen, M., Andreatta, M., Peters, B. & Buus, S. Immunoinformatics: predicting peptide–MHC binding. *Annu Rev. Biomed. Data Sci.* **3**, 191–215 (2020).
19. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. 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 (2020).
20. Gfeller, D. & Bassani-Sternberg, M. Predicting antigen presentation-What could we learn from a million peptides? *Front Immunol.* **9**, 1716 (2018).
21. Nielsen, M., Lund, O., Buus, S. & Lundegaard, C. MHC Class II epitope predictive algorithms. *Immunology* **130**, 319–328 (2010).
22. 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).
23. Kaabinejadian, S. et al. 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 (2022).
24. Alvarez, B., Barra, C., Nielsen, M. & Andreatta, M. Computational tools for the identification and interpretation of sequence Motifs in immunopeptidomes. *Proteomics* **18**, 1700252 (2018).
25. Caron, E. et al. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol. Cell. Proteom.* **14**, 3105–3117 (2015).
26. Purcell, A. W., Ramarathinam, S. H. & Ternette, N. Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nat. Protoc.* **14**, 1687–1707 (2019).
27. Barra, C. et al. Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med* **10**, 84 (2018).
28. Paul, S. et al. Determination of a predictive cleavage motif for eluted major histocompatibility complex class II ligands. *Front. Immunol.* **9**, 1795 (2018).
29. Racle, J. et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).
30. Wang, P. et al. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinforma.* **11**, 568 (2010).
31. Alvarez, B. et al. NNAlign_MA; MHC peptidome deconvolution for accurate MHC Binding Motif characterization and improved T-cell epitope predictions. *Mol. Cell. Proteom.* **18**, 2459–2477 (2019).
32. Karosiene, E. et al. 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 (2013).
33. Thomsen, M. C. F., Lundegaard, C., Buus, S., Lund, O. & Nielsen, M. MHCcluster, a method for functional clustering of MHC molecules. *Immunogenetics* **65**, 655–665 (2013).
34. Marcu, A. et al. HLA ligand atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J. Immunother. Cancer* **9**, e002071 (2021).
35. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinforma.* **8**, 238 (2007).
36. 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**, 33 (2016).
37. Hoof, I. et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13 (2009).
38. Gonzalez-Galarza, F. F., Christmas, S., Middleton, D. & Jones, A. R. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res.* **39**, D913–D919 (2011).
39. Moore, R. M., Harrison, A. O., McAllister, S. M. & Polson, S. W. & Eric Wommack, K. Iroki: Automatic customization and visualization of phylogenetic trees. *PeerJ* **8**, e8584 (2020).
40. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
41. 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**, W281–W287 (2012).
42. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

Acknowledgements

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (NIAID), under award number 75N93019C00001. We would like to sincerely thank Dr. Rico Buchli (Pure Protein, LLC) for providing the SPVL3 affinity columns for this study. We also thank Steven Cate (University of Oklahoma Health Sciences Center) and Sean Osborn (Pure MHC, LLC) for HLA typing of the BLCLs and very helpful discussions.

Author contributions

S.K. and M.N. designed the study. The experimental data used in the study was generated by S.K., with contribution from H.Y. and W.H. J.B.N. and M.N. generated the computational results and figures. B.P., C.B. and L.G. contributed towards methodology regarding the cis and trans-only DQ analysis and provided scientific feedback. The manuscript was written by J.B.N., S.K. and M.N., with contributions from all authors. All authors have read and approved the final version of the paper.

Competing interests

S.K. is an employee at Pure MHC, LLC. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-023-04749-7.

Correspondence and requests for materials should be addressed to Morten Nielsen.

Peer review information *Communications Biology* thanks Shanfeng Zhu, David Gfeller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Zhijuan Qiu.