HHS Public Access

Author manuscript

J Immunol Methods. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

J Immunol Methods. 2015 July ; 422: 28–34. doi:10.1016/j.jim.2015.03.022.

Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes

Sinu Paul[1], **Cecilia S. Lindestam Arlehamn**[1], **Thomas J. Scriba**[2], **Myles B.C. Dillon**[1], **Carla Oseroff**[1], **Denise Hinz**[1], **Denise M. McKinney**[1], **Sebastian Carrasco Pro**[3], **John Sidney**[1], **Bjoern Peters**[1], and **Alessandro Sette**[1]

[1]La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA, 92037, USA

[2]South African Tuberculosis Vaccine Initiative (SATVI) and School of Child and Adolescent Health, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[3]Laboratorio de Bioinformática y Biología Molecular, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru

Abstract

Computational prediction of HLA class II restricted T cell epitopes has great significance in many immunological studies including vaccine discovery. In recent years, prediction of HLA class II binding has improved significantly but a strategy to globally predict the most dominant epitopes has not been rigorously defined. Using human immunogenicity data associated with sets of 15-mer peptides overlapping by 10 residues spanning over 30 different allergens and bacterial antigens, and HLA class II binding prediction tools from the Immune Epitope Database and Analysis Resource (IEDB), we optimized a strategy to predict the top epitopes recognized by human populations. The most effective strategy was to select peptides based on predicted median binding percentiles for a set of seven DRB1 and DRB3/4/5 alleles. These results were validated with predictions on a blind set of 15 new allergens and bacterial antigens. We found that the top 21% predicted peptides (based on the predicted binding to seven DRB1 and DRB3/4/5 alleles) were required to capture 50% of the immune response. This corresponded to an IEDB consensus percentile rank of 20.0, which could be used as a universal prediction threshold. Utilizing actual binding data (as opposed to predicted binding data) did not appreciably change the efficacy of global predictions, suggesting that the imperfect predictive capacity is not due to poor algorithm performance, but intrinsic limitations of HLA class II epitope prediction schema based on HLA binding in genetically diverse human populations.

Corresponding author: Sinu Paul, Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037 USA, Phone: (858) 752-6925, Fax: (858) 752-6987, spaul@liai.org.

Keywords

Epitope prediction; HLA class II restricted T cell epitopes

1. Introduction

The prediction and identification of HLA class II restricted T cell epitopes is a task of significance for several different applications. These include, to name a few, efforts to elucidate the epitopes responsible for the induction of allergen specific T cells, the study of immune response against complex pathogens with large genomes, such as *Mycobacterium tuberculosis* (MTB), or the identification and removal of unwanted epitopes in protein-based drugs.

Class II molecules are alpha/beta heterodimers encoded by four different loci in humans, DRA/DRB1, DRA/DRB3/4/5, DPA/DPB and DQA/DQB. With the exception of DRA, all other chains are highly polymorphic (Robinson et al., 2003). The extensive polymorphism of HLA class II molecules in the general population does represent a formidable obstacle to epitope identification approaches. However, it has been recognized that the majority of molecules expressed in the general population can be reconciled to a manageable number by focusing on those most frequently expressed (McKinney et al., 2013). At the same time, extensive similarities exist within the peptides bound by different allelic variants, and even across different loci (Greenbaum et al., 2011). Finally and perhaps most significantly, it has been shown that peptides capable of binding multiple HLA class II molecules (i.e. promiscuous peptides) often account for a large fraction, if not the majority, of antigen specific T cell responses (Oseroff et al., 2010; Paul et al., 2013a).

Bioinformatic predictions of MHC binding capacity have proven to be a key component of various epitope identification approaches. While historically less impressive than the case for HLA class I, the performance of various methods for the prediction of HLA class II binding peptides has been subject to significant improvement over the last few years as more novel and sophisticated computational approaches have been implemented, as reviewed and evaluated in several studies (Paul et al., 2013a; Nielsen et al., 2010; Wang et al., 2010). However, to date, definition of an optimal strategy to employ these algorithms to allow efficient prediction of promiscuous class II restricted T cell epitopes, or dominant epitopes frequently recognized in an outbred cohort, has been difficult.

During the last few years we have generated T cell recognition data in humans for several panels of overlapping peptides completely spanning entire antigens of immunological interest. These antigens included four house dust mite allergens (referred as HDM data set) (Hinz, D., in preparation), ten allergens linked to pollen grass allergies (TG) (Oseroff et al., 2010), four MTB antigens recognized by healthy donors with latent MTB infection (LTBI) from the San Diego region (TB-SD) (Arlehamn et al., 2012), and eleven different MTB antigens recognized by healthy donors with LTBI from the Cape Town (South Africa) region (TB-CT) (Mc Kinney, D., in preparation). Each set of peptides was tested with similar methodology, in 20-40 different HLA typed individuals of diverse ethnicity. Overall, a total of 1151 peptides were tested, in studies involving more than 95 donors.

In the present study we have utilized these data sets to perform an evaluation of different strategies to implement HLA binding predictions for the purpose of selecting epitopes with the capacity to elicit HLA class II restricted T cell immune responses. To validate the approach defined herein, independent blind analyses were subsequently performed using overlapping peptide sets spanning six different cockroach allergens (Oseroff et al., 2012; Dillon, M., in preparation) and the five antigens included in the vaccine against whooping cough (*Bordetella pertussis*) (Dillon, M., in preparation).

2. Materials and methods

2.1 Immunogenicity studies

Sets of overlapping 15 or 16mer peptides spanning various allergen and bacterial antigens were screened for immune reactivity as previously described (Oseroff et al., 2010; Arlehamn et al., 2012; Oseroff et al., 2012). Antigen specific cytokine production in donor peripheral-blood mononuclear cells (PBMC) was measured in dual or single ELISPOT assays. Responses to timothy grass, cockroach and house dust mite peptides were measured following *in vitro* stimulation with respective allergen extracts, and responses to *Bordetella pertussis* peptides following stimulation with corresponding vaccine antigens. Responses to mycobacterial antigens were analyzed *ex vivo*. Peptide specific responses were expressed as spot-forming cells (SFCs)/10^6 PBMC. Donors were HLA typed at each class II locus to four-digit resolution by SSO/SSP HLA typing (One Lambda reagents, Canoga Park, CA, USA) or deep sequencing methods (2012McKinney et al., submitted).

2.2 MHC purification and binding Assays

The binding affinity of peptides in TG and TB-SD data sets to the 26 most frequent alleles was experimentally determined. Quantitative measurement of peptide binding capacity for HLA class II molecules was performed in competition assays based on the inhibition of binding of a high affinity radiolabeled peptide to purified MHC molecules. Purification of class II MHC molecules by affinity chromatography, and the performance of binding assays were done essentially as detailed elsewhere (Sidney et al., 2013). Briefly, EBV transformed homozygous cell lines were used as sources of MHC molecules. A high affinity radiolabeled peptide (0.1-1 nM) was co-incubated at room temperature or 37°C with purified MHC in the presence of a cocktail of protease inhibitors. Following a two-day incubation, MHC bound radioactivity was determined by capturing MHC/peptide complexes on Ab coated Lumitrac 600 plates (Greiner Bio-one, Frickenhausen, Germany), and measuring bound cpm using the TopCount (Packard Instrument Co., Meriden, CT) microscintillation counter. The concentration of peptide yielding 50% inhibition of the binding of the radiolabeled peptide was calculated. Under the conditions utilized, where [label]<[MHC] and IC50 ≥ [MHC], the measured IC50 values are reasonable approximations of the true Kd values. Each competitor peptide was tested at six different concentrations covering a 100,000-fold range, and in three or more independent experiments. As a positive control, the unlabeled version of the radiolabeled probe was also tested in each experiment.

2.3 Prediction of binding affinity

Peptide binding affinity for HLA class II alleles was predicted using the MHC II binding prediction tool available at the IEDB (www.iedb.org) (Vita et al., 2014, Zhang et al., 2008, Kim et al., 2012). Allele-specific consensus percentile ranks of all algorithms queried by the IEDB tool were utilized (Wang et al., 2010). A percentile rank is generated by comparing the selected peptide's predicted binding affinity against that of a large set of similarly sized peptides randomly selected from the SWISS-PROT database (Kim et al., 2012). Percentile rank provides a uniform scale allowing comparisons across different predictors. A lower percentile rank value indicates higher affinity. In the case of consensus method, median of the percentile ranks of the three methods involved is considered as the IEDB consensus percentile rank.

2.4 Correction for epitope redundancy

Responses against two consecutive peptides are often due to the same minimal epitope. To avoid counting the same epitope twice, two consecutive responses with magnitudes within 2.5-fold of each other were merged into a single antigenic region, and the higher SFC value utilized. The region was considered successfully predicted if either of the two peptides was predicted, and "credit" for prediction was given only once.

3. Results and discussion

3.1 Evaluation of optimal prediction strategies for HLA class II epitopes

Predicted binding affinity of peptides in the data sets (Table 1) for a previously described set of 26 HLA class II alleles that are most frequent in the general worldwide population (Greenbaum et al., 2011) (Table 2), was determined as described above. To evaluate the efficacy of various approaches to employing these predictions to identify the most dominant epitope responses, the percentage (fraction) of peptides in each data set needed to capture 50% of the total response (50% of the total SFC values in the data set - expressed as SFCs/10^6 PBMCs) was utilized as a performance metric.

As a first approach, we considered the "promiscuous binding capacity" of each peptide, where promiscuity is defined by the number of alleles bound (i.e., peptides binding more alleles being more promiscuous binders). For this purpose, a peptide was considered binder for a specific allele if its IEDB predicted consensus percentile rank was ≤ 20. This approach was originally devised by us based on a single dataset (TG, Oseroff et al., 2010). Using this approach, as shown in Figures 1a and b, an average of 30.91% (range 25.35%-40.08%) peptides were needed to capture 50% of the total response in the data set.

As a second approach, we considered the "median consensus percentile rank" of each peptide, defined as the median of the IEDB consensus percentile ranks predicted for the set of 26 selected alleles. This approach was the most effective, with the top 26.26% (range 16.90%-38.38%) of the peptides capturing 50% of the total response (Figure 1a, c).

3.2 Comparison of the median percentile rank approach with best percentile and allele specific binding thresholds

In addition to the promiscuous binding capacity (promiscuity) and median consensus percentile rank, other strategies were also evaluated. Considering that the dominance of a particular peptide might be a reflection of very high binding affinity for a given allele rather than promiscuity, one approach we tested was based on “best percentile rank”. In this approach, the peptides were sorted based on the best percentile rank (the lowest percentile rank value among the 26 most frequent alleles) and the percentage of peptides required to capture 50% of total SFC was identified. This method required on average 27.05% of the top peptides to capture 50% of the response.

In yet another approach, we followed a strategy based on allele-specific binding affinity thresholds that had improved the efficacy of class I predictions (Paul et al., 2013b). For this analysis, all previously identified 15-mer epitopes with defined HLA class II restriction were retrieved from the IEDB. The allele-specific thresholds were estimated for each of the 26 alleles in terms of binding affinity predicted by SMM_align method (IC_{50}) (Nielsen et al., 2007), taking into account the number of predicted binders in the set of epitopes retrieved from the IEDB (based on a general threshold of IC_{50} 1000nM) and the SMM_align IC_{50} value demarcating the top 75% peptides of the same epitope set. The total promiscuity for each peptide was then recalculated based on the allele-specific thresholds and the fraction of peptides required to capture 50% response was identified. This approach required 33.58% of peptides to capture 50% response.

Both these approaches were found to be less efficient than the median consensus percentile rank method, requiring on average 27.05% and 33.58% peptides respectively (vs. 26.26%), to capture 50% response based on the 26 most frequent alleles (data not shown).

3.3 Exclusion of DP locus improves predictive efficacy

As different HLA class II loci appear to contribute differentially to human responses (Oseroff et al., 2010), we hypothesized that examining the performance as a function of the class II locus may improve predictions. The average % of peptides required to capture 50% SFC for different combinations of DRB1, DRB3/4/5, DQ, and DP alleles are shown in Figure 2. The best results (23.82%) were obtained when DP alleles were left out. The lower performance of methods incorporating DP molecules might be due to the fact that less binding data is available for these molecules leading to inferior prediction algorithms or it could be that DP molecules are less often restricting elements for dominant T cell responses.

3.4 Optimal results obtained with a set of seven DRB1 and DRB3/4/5 alleles

We next examined the effect of varying the specific alleles included in the prediction panel. Frequency thresholds for inclusion were varied independently for each locus (i.e., DQ, DRB1 and DRB3/4/5). The best results (21.41% of peptides needed to capture 50% SFC) were observed when the three DRB1 alleles with frequency ≥ 12% (DRB1*03:01, DRB1*07:01, DRB1*15:01) were used along with the four DRB3/4/5 alleles (DRB3*01:01, DRB3*02:02, DRB4*01:01, DRB5*01:01) (data not shown). This empirical optimization is probably reflective of the fact that DR alleles are the most dominant locus restricting HLA

class II responses in humans. It is noteworthy that the seven allelic variants cover the main HLA class II supertypes (Greenbaum et al., 2011).

3.5 Predictions based on alleles frequent in specific donor cohorts

HLA frequencies vary in cohorts of individuals with different ethnicity. Accordingly, we examined the performance of the "median consensus percentile rank" predictions utilizing alleles custom selected for the cohorts in which the peptides were tested. Specifically, we included all alleles with frequencies of ≥ 10% in a given donor population. As above, the best results (19.69%) were obtained using only DRB1 and DRB3/4/5 alleles (Figure 3). At the same time, the improvement seen when using cohort specific allele sets is minor suggesting that tailoring the prediction to a specific population has limited value.

3.6 Defining a universal prediction threshold

The percent of total peptides required to capture 50% of the response, as calculated here on a protein-by-protein basis, is not available when considering individual peptides. To derive a standard prediction threshold, we calculated the median IEDB consensus percentile rank, using predictions for the seven DRB1 and DRB3/4/5 alleles highlighted above, associated with the selected set of peptides yielding 50% of the response. This value was found to be 20.0 (median consensus percentile rank from the seven selected alleles).

3.7 Validation of the results with blind prediction using new data sets

The analyses above suggested that the optimal approach for efficient selection of epitope candidates would be based on determining the median consensus percentile rank across a selected panel of seven DR alleles (3 DRB1 alleles with frequency ≥12% in conjunction with 4 DRB3/4/5 alleles). To validate these results we examined overlapping peptides for two additional sets of proteins of immunological interest: 1) cockroach allergens and 2) acellular pertussis vaccine antigens.

When the range of approaches tried above was implemented against the two blind sets, the best performance was again achieved with the "median consensus percentile rank" approach. When the universal median IEDB consensus percentile threshold defined above (20.0) with the panel of seven DR alleles was utilized, the average % of SFC captured using peptides with median IEDB consensus percentile rank ≤ 20.0 was found to be 48.55%, confirming the validity of this prediction threshold.

3.8 Comparison with experimentally measured binding data

The analysis showed that on average, that approximately to top ~20% scoring peptides are needed to capture 50% of the immune response. In order to examine whether this rather high number of peptides is due to lower efficacy of HLA class II binding prediction algorithms, we compared the performance based on predicted binding affinity with that of experimentally measured binding affinity. For this analysis predicted and measured binding affinity for the seven selected DR alleles was assessed in the context of two cohorts (TG and TB-SD).

It was found that the peptide selection strategy based on the predicted binding affinity required actually fewer peptides to capture 50% of the immune response, compared to the measured binding affinity (15.63% vs. 32.86% peptides) (Figure 4). No significant difference was observed when binding data from two other allele categories were used (the most frequent panel alleles and the panel alleles excluding DP locus). This shows that the overall lower efficacy in prediction of HLA class II immunogenicity is an inherent issue of class II alleles rather than underperformance of HLA class II binding prediction algorithms.

4. Conclusions

We scrutinized the use of HLA class II binding predictions to identify sets of epitopes with high immunological activity. The results validate previous observations that promiscuous binders account for a large fraction of the total response. However, in comparison to HLA class I predictions, the results are sobering, as the overall performance is remarkably less effective. This is in line with other recent studies (Chaves et al., 2012). We considered the possibility that these results may be due to a generally lower performance of class II binding prediction algorithms. However when actual binding data, as opposed to predicted binding data, was used, no significant improvement was noted. These results suggest that the imperfect predictive capacity is due to intrinsic limitations of HLA class II epitope prediction schema based on HLA binding in genetically diverse human populations.

At the same time, our results provide guidance for practical implementation of predictions, and identify specific subsets of HLA molecules that are most effectively considered by prediction schemes. The synthesis of approximately 20% of the peptides in a set of 15-mer peptides overlapping by 10 residues allows covering a 200-residue protein (otherwise covered by 38 overlapping peptides) with 8 peptides, which still affords significant cost savings, and enables the screening of large genomes with experimental designs based on predicted epitope peptide pools.

Acknowledgements

This project has been funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under Contract Numbers HHSN272201200010C (IEDB), HHSN272200900044C (TB), HHSN272200700048C (Allergy), HHSN272200900052C (ICAC2), HHSN272201000052I (Rho) and UM1AI114271 (ICAC3) and Grant number NIH U19 AI100275 (Allergy) in addition to Global Health Grant OPP1066265 from Bill & Melinda Gates Foundation.

Abbreviations

IEDB	Immune Epitope Database and Analysis Resource
MTB	Mycobacterium tuberculosis
HDM	House dust mite
TG	Timothy grass
TB-SD	Tuberculosis – San Diego cohort
TB-CT	Tuberculosis – Cape Town cohort

LTBI Latent Tuberculosis Infection

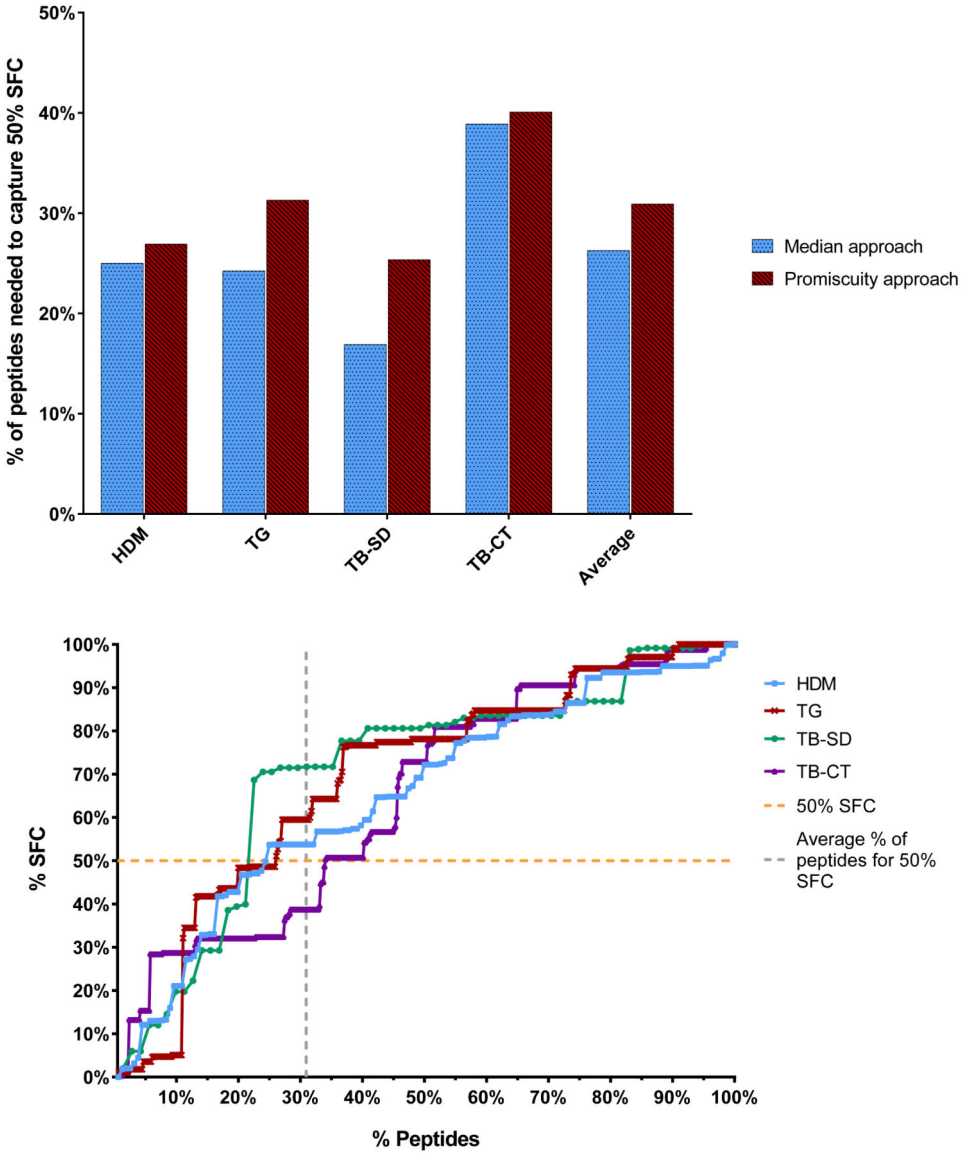
6. References

1. Arlehamn CS, Sidney J, Henderson R, Greenbaum JA, James EA, Moutaftsi M, Coler R, McKinney DM, Park D, Taplitz R, Kwok WW, Grey H, Peters B, Sette A. Dissecting mechanisms of immunodominance to the common tuberculosis antigens ESAT-6, CFP10, Rv2031c (hspX), Rv2654c (TB7.7), and Rv1038c (EsxJ). J. Immunol. 2012; 188:5020–5031. [PubMed: 22504645]
2. Chaves FA, Lee AH, Nayak JL, Richards KA, Sant AJ. The utility and limitations of current Web-available algorithms to predict peptides recognized by CD4 T cells in response to pathogen infection. J. Immunol. 2012; 188:4235–4248. [PubMed: 22467652]
3. Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics. 2011; 63.6:325–335. [PubMed: 21305276]
4. Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, Greenbaum J, Lundegaard C, Sette A, Lund O, Bourne PE. Immune epitope database analysis resource. Nucleic Acids Res. 2012; 40:W525–W530. [PubMed: 22610854]
5. McKinney DM, Southwood S, Hinz D, Oseroff C, Arlehamn CSL, Schulten V, Taplitz R, Broide D, Hanekom WA, Scriba TJ. A strategy to determine HLA class II restriction broadly covering the DR, DP, and DQ allelic variants most commonly expressed in the general population. Immunogenetics. 2013; 65.5:1–14. [PubMed: 23053058]
6. Nielsen M, Lund O, Buus S, Lundegaard C. MHC Class II epitope predictive algorithms. Immunology. 2010; 130:319–328. [PubMed: 20408898]
7. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007; 8:238. [PubMed: 17608956]
8. Oseroff C, Sidney J, Kotturi MF, Kolla R, Alam R, Broide DH, Wasserman SI, Weiskopf D, McKinney DM, Chung JL. Molecular determinants of T cell epitope recognition to the common Timothy grass allergen. The Journal of Immunology. 2010; 185:943–955. [PubMed: 20554959]
9. Oseroff C, Sidney J, Tripple V, Grey H, Wood R, Broide DH, Greenbaum J, Kolla R, Peters B, Pomés A. Analysis of T cell responses to the major allergens from German cockroach: Epitope specificity and relationship to IgE production. The Journal of Immunology. 2012; 189:679–688. [PubMed: 22706084]
10. Paul S, Kolla RV, Sidney J, Weiskopf D, Fleri W, Kim Y, Peters B, Sette A. Evaluating the Immunogenicity of Protein Drugs by Applying In Vitro MHC Binding Data and the Immune Epitope Database and Analysis Resource. Clinical and Developmental Immunology. 2013a; 2013
11. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. J. Immunol. 2013b; 191:5831–5839. [PubMed: 24190657]
12. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SG. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 2003; 31:311–314. [PubMed: 12520010]
13. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, Grey HM, Sette A. Measurement of MHC/Peptide Interactions by Gel Filtration or Monoclonal Antibody Capture. Current protocols in immunology. 2013:18.3.1–18.3.36.
14. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, Peters B. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2014; 2014
15. Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics. 2010; 11.1:568. [PubMed: 21092157]

16. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J, Lund O, Lundegaard C, Nielsen M, Ponomarenko J, Sette A, Zhu Z, Peters B. Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res. 2008; 36:W513–8. [PubMed: 18515843]

Highlights

- A new scheme for prediction of HLA class II restricted T cell epitopes.
- 21% of top peptides capture 50% of immune response.
- Definition of a universal threshold (Median IEDB consensus percentile rank = 20.0).
- The scheme is validated using 2 blind data sets.



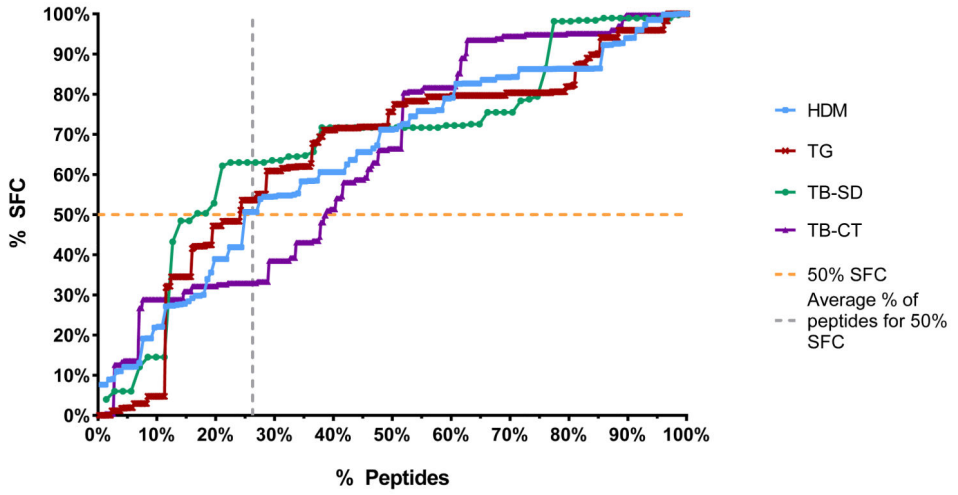


Figure 1.

a: Performance of two approaches for implementing HLA class II binding predictions to identify T cell epitopes. The % of peptides needed for each method to identify a panel of epitopes accounting for 50% of the total antigen specific response (SFC) is shown for 4 different systems, as described in the text: HDM (House dust mite), TG (Timothy grass), TB-SD (MTB), and TB-CT (MTB). Blue bars show performance based on ranking peptides according to the median consensus percentile rank against a panel of the 26 most common HLA class II alleles. Red bars show performance based on ranking peptides according to the number of alleles predicted to bind (promiscuity). A lower % of peptides indicates better performance.

b: The % of response predicted as a function of the % of the total peptides predicted for the four data sets with the “promiscuous binding capacity” approach using the 26 most frequent class II alleles.

c: The % of response predicted as a function of the % of the total peptides predicted for the four data sets with the “Median percentile” approach using the 26 most frequent class II alleles.

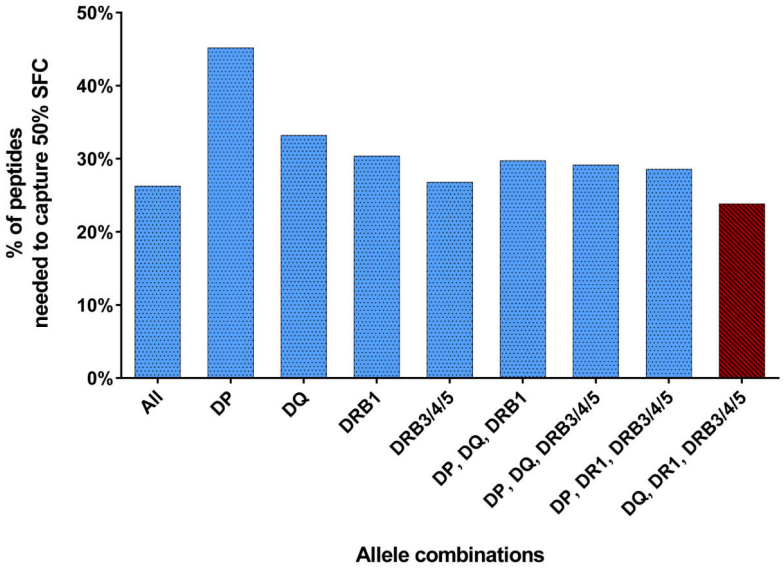


Figure 2. Performance of the "median consensus percentile rank" approach as a function of variable inclusion of the DP, DQ, DRB1 and DRB3/4/5 loci

The performance was best when DP locus was excluded (red bar).

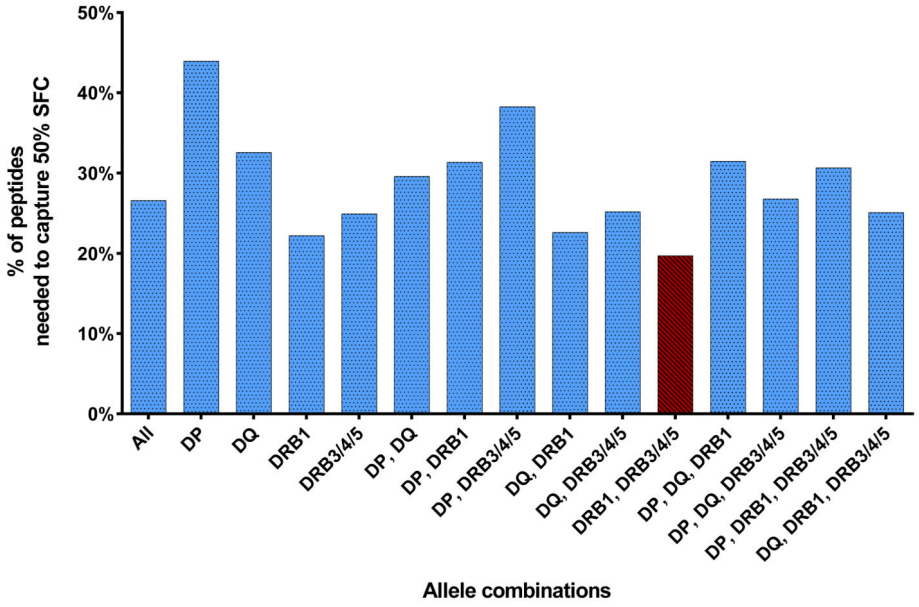


Figure 3. Average % of peptides required to capture 50% SFC for different allele combinations using the alleles with frequency >10% in each specific corresponding donor cohort

The prediction was best when alleles from DRB1 and DRB3/4/5 alleles were used (red bar)

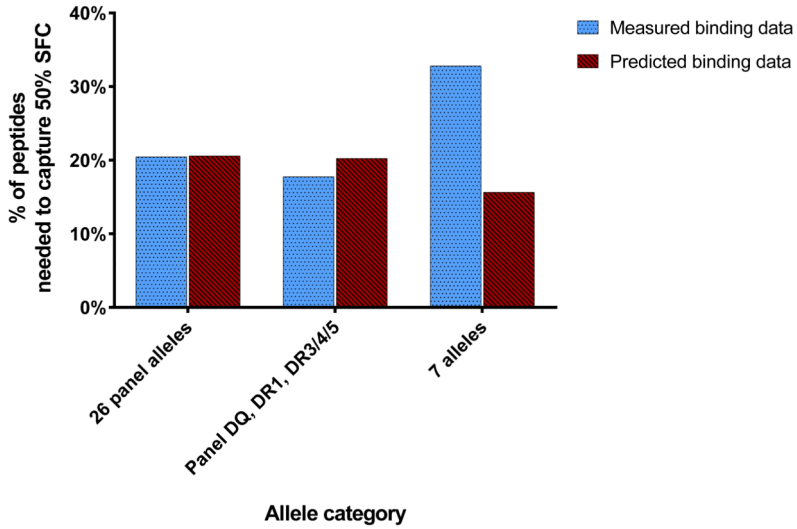


Figure 4. Comparison of the prediction performance based on experimentally measured binding data with that of predicted binding data for TG and TB-SD data sets

Blue bars show the % of peptides needed to capture 50% SFC using experimentally measured HLA class II binding data while red bars show the same using predicted binding data. The prediction strategy using 7 selected alleles performed better with predicted binding compared to measured binding.

Table 1

Data sets used in the analyses

Data set	No. of antigens	Antigens	No. of peptides per antigen	No. of peptides per data set	No. of donors	Ethnicity	Reference
HDM	4	proDer p 1.0105	59	156*	20	White	Hinz, D., in preparation
		proDer f 1.0101	59				
		Der p 2.0101	24				
		Der f 2.0103	24				
TG	10	Phl p 1	51	425	25	Mixed (predominantly white)	Oseroff et al., 2010
		Phl p 2	23				
		Phl p 3	18				
		Phl p 4	103				
		Phl p 5.0103	61				
		Phl p 6	26				
		Phl p 7	14				
		Phl p 11	27				
		Phl p 12	25				
		Phl p 13	77				
TB-SD	4	Rv1038c	9	71	18	Mixed	Arlehamn et al., 2012
		Rv2031c	27				
		Rv3874	18				
		Rv3875	17				
TB-CT	11	Rv0125	69	499	32		McKinney, D. M., in preparation
		Rv0288	18				
		Rv1196	77				
		Rv1813c	27				
		Rv1886c	63				
		Rv2608	114				
		Rv2660c	13				
		Rv3619	17				
		Rv3620c	18				
		Rv3804c	66				
		Rv3875	17				
Cockroach	6	Bla g 1	189	463	19		Oseroff et al., 2012
		Bla g 2	69				
		Bla g 4	35				

Data set	No. of antigens	Antigens	No. of peptides per antigen	No. of peptides per data set	No. of donors	Ethnicity	Reference
		Bla g 5	39				
		Bla g 6	76				
		Bla g 7	55				
Pertussis	9	fhaB	468	785	23		Dillon, M.B.C., in preparation
		fim2	26				
		fim3	25				
		prn	131				
		ptxA	40				
		ptxB	30				
		ptxC	28				
		ptxD	21				
		ptxE	16				

* 10 peptides are shared by different antigens in the Der p/f data set. Thus, the total no. of unique peptides is 10 less than the sum of peptides in individual antigens.

Table 2

26 HLA class II alleles that are most frequent in the general worldwide population, and thus were included in the analyses

Locus	Allele	Phenotype frequency	Gene frequency
DRB1	DRB1*0101	5.44	2.76
	DRB1*0301	13.72	7.11
	DRB1*0401	4.58	2.32
	DRB1*0405	6.15	3.13
	DRB1*0701	13.52	7.01
	DRB1*0802	4.87	2.46
	DRB1*0901	6.17	3.13
	DRB1*1101	11.84	6.11
	DRB1*1201	3.94	1.99
	DRB1*1302	7.71	3.93
	DRB1*1501	12.18	6.29
	Total	71.09	46.23
DRB3/4/5	DRB3*0101	26.12	14.04
	DRB3*0202	34.25	18.92
	DRB4*0101	41.75	23.68
	DRB5*0101	15.98	8.34
	Total	87.73	77.30
DQA1/DQB1	DQA1*0501/DQB1*0201	11.29	5.81
	DQA1*0501/DQB1*0301	35.14	19.47
	DQA1*0301/DQB1*0302	19.05	10.03
	DQA1*0401/DQB1*0402	12.78	6.61
	DQA1*0101/DQB1*0501	14.65	7.61
	DQA1*0102/DQB1*0602	14.59	7.58
	Total	81.61	57.11
DPB1	DPA1*02:01/DPB1*01:01	16.01	8.35
	DPA1*01:03/DPB1*02:01	17.47	9.15
	DPA1*01/DPB1*04:01	36.20	20.13
	DPA1*03:01/DPB1*04:02	41.63	23.60
	DPA1*02:01/DPB1*05:01	21.68	11.50
	Total	94.49	76.53