Received Date : 18-Dec-2015
Revised Date : 09-Jan-2016
Accepted Date : 14-Jan-2016
Article type : Original Article

T cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome

Anne Bresciani[1,2], Sinu Paul[1], Nina Schommer[1], Myles B. Dillon[1], Tara Bancroft[1], Jason Greenbaum[1], Alessandro Sette[1], Morten Nielsen[2,3] and Bjoern Peters[1*]

[1]*La Jolla Institute for Allergy and Immunology, Department of Vaccine Discovery, 9420 Athena Circle, La Jolla, CA 92037, USA*

[2]*Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark*

[3]*Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, San Martín, B 1650 HMP, Buenos Aires, Argentina*

** communicating author*

Corresponding author mail id: ***bpeters@lji.org***

Keywords

Bioinformatics

T cell recognition

Epitopes

Abbreviations

HMP – Human Microbiome Project

MTB – Mycobacterium tuberculosis

BP – Bordetella pertussis

CR – cockroaches

SFC - spot forming cell

PBMC - peripheral blood mononuclear cell

Abstract

Several mechanisms exist to avoid or suppress inflammatory T cell immune responses that could prove harmful to the host due to targeting self-antigens or commensal microbes. We hypothesized that these mechanisms could become evident when comparing the immunogenicity of a peptide from a pathogen or allergen with the conservation of its sequence in the human proteome or the healthy human microbiome. Indeed, performing such comparisons on large sets of validated T cell epitopes, we found that epitopes that are similar with self-antigens above a certain threshold showed lower immunogenicity, presumably as a result of negative selection of T cells capable of recognizing such peptides. Moreover, we also found a reduced level of immune recognition for epitopes conserved in the commensal microbiome, presumably as a result of peripheral tolerance. These findings indicate that the existence (and potentially the polarization) of T cell responses to a given epitope is influenced and to some extend predictable based on its similarity to self-antigens and commensal-antigens.

Introduction

Several methods are available that can accurately predict binding of peptides to MHCI and MHCII molecules (1–5). Binding to an MHC molecule is an essential though not sufficient criterion for a peptide to be recognized by T cells as an immune target. Other factors play a role, such as the ability of a MHCI binding peptide to be processed from its source protein (6–11), and the amino acid composition of the peptide, which has been linked to immunogenicity presumably by some residues being more visible to the T cell receptor (12, 13). However, there are significant factors influencing T cell immunogenicity beyond the factors mentioned above that remain unknown. This is particularly evident for MHC class II restricted epitopes, where binding predictions correlate well with measured binding affinities (5), but when attempting to predict immunogenic peptides the performance is far from perfect (14). We have recently demonstrated that a combination of HLA class II binding predictions selecting a set of top 20% candidate peptides will cover 50% of the immune response (15). While this is practically very useful, it also demonstrates that mechanisms beyond MHC binding affinity shape immune recognition patterns (14).

One effect that is expected to influence immunogenicity of a given peptide is the suppression of immune responses that could be harmful to the host. T cells reacting to peptides conserved in the human proteome are expected to be deleted by negative selection during T cell maturation. In addition, it has been postulated that inflammatory T cells reactive to peptides found in commensal microorganisms will be suppressed by regulatory T cells - a mechanism called peripheral tolerance. However, the extent to which these mechanisms imprint on the T cell immune

repertoire and thereby impact the immune recognition of peptides from pathogens or allergens has not been systematically analyzed or quantified.

In this study, we examined if there is evidence for tolerization in immune recognition patterns by correlating immune responses of peptides from bacterial pathogens and allergens with the sequence conservation of these peptides in the human proteome and in proteins identified in the Human Microbiome Project (HMP). We find that there is evidence for both.

Methods

Peptide immunogenicity dataset assembly

Peptides from three independent studies were utilized. The first dataset consisted of 15-mer peptides from *Mycobacterium tuberculosis* (MTB) antigens tested for recognition in IFNg ELISPOT assays using peripheral blood mononuclear cells (PBMC) from individuals latently infected with MTB (16, 17). The second dataset consisted of 16-mer peptides overlapping antigens included in the *B. pertussis* vaccine, tested for recognition in IFNg ELISPOT assays by PBMCs from previously vaccinated individuals (manuscript in preparation). The third dataset consisted of 15-mer peptides contained in antigens encoded in cockroach proteins and tested for recognition in IL-5 ELISPOTs by PBMCs from allergic individuals (18, 19). All assays were performed in triplicates and included media stimulated cells as a background control. Individual experiment were considered positive if the number of spot forming cells (SFC) above background was more than 20 per million input cells and the number of IFNg producing cells after peptide stimulation was significantly above background based on a Student's t-test p-value < 0.05 and a stimulation index > 2.0.

This combination of positivity criteria is commonly used in our laboratory, including in the studies cited above. To classify a peptide overall as positive in the donor cohort, the SFC from individual experiments that met the positivity criteria were added up, and the total SFC as well as the number of positive experiments were utilized as positivity cutoffs as specified in the results section.

Human proteome and microbiome sequence dataset assembly

Protein sequences from the human proteome were downloaded from UniProt (www.uniprot.org) (20) using the query: *keyword:"Complete proteome" AND organism:"Homo sapiens (Human) [9606]"*. The sequences were downloaded using the "Download" option, choosing "Download all" and Format: FASTA (canonical & isoform). Protein sequences from the human gut microbiome were retrieved from the Human Microbiome Project (HMP) (21, 22). The data was downloaded from the HMP Data Analysis and Coordination Center (www.hmpdacc.org/HMRGD), where annotated reference genomes could be found. The complete set of annotated reference genomes was downloaded as protein sequences in fasta format, by choosing "Download all" in protein multifasta (PEP) format for with the body site specified as 'gastrointestinal tract'.

Quantifying peptide similarity to protein sequences

For a given peptide of length N, we define the similarity score of that peptide with sequence *a* as the highest score for an equal length amino acid stretch *b* in the set of target proteins, where the match score is given by the formula

$$\frac{bl(a,b)}{\sqrt{bl(a,a) * bl(b,b)}}$$

in which bl(a,b) is the sum of the BLOSUM62 matrix (23) values for substituting residues in peptide *a* for the residues in amino acid stretch *b*.

Results

Assembly of a peptide immunogenicity dataset

Immunogenicity can be assessed by different assays, and differs substantially between antigenic systems due to variances in the route of exposure to the antigens. We thus only compared immunogenicity of peptides from the same antigenic source and tested in the same assay systems. We utilized datasets from three studies representing different modes of exposure: infection with *Mycobacterium tuberculosis* (MTB), vaccination with *Bordetella pertussis* (BP) and inhalation of allergens from cockroaches (CR). In each of these datasets, a large numbers of peptides (>500) had been tested in a consistent fashion by ELISPOT assays in a large (>30) number of donors. For the cockroach dataset, peptides were considered positive if they had significant responses in at least two donors (t-test p-value < 0.05, SI >2, SFC>20), and if the total number of spot forming cells per million (SFC) summed over all donors in the cohort was greater than one hundred. For the *M. tuberculosis* and *B. pertussis* datasets, reactivities were higher in general, and the cutoff for positivity was set to 3 reacting donors and a total SFC>200. Peptides were considered negative if they did not give a significant response in any single donor. Peptides with intermediate reactivities were discarded. These selection criteria take into account inherent differences in assays and immunization procedures to ensure that the positive set capture 70% or more of the total reactivity. **Table 1** lists the number of positive and negative peptides as well as the number of donors for each of the

datasets. **Supplementary Table 1** lists the peptide sequences and their immunogenicity classification.

Quantifying peptide similarity

In order to compare the bacterial and allergen derived peptides tested for immunogenicity with the human proteome and human microbiome, we needed to define a quantitative score of similarity. A simple approach is to count the number of different amino acids in two sequences, but this neglects that some exchanges of amino acids alter the properties of a peptide significantly more than others. To account for this, we quantified amino acid similarity using a BLOSUM similarity score described in the methods section, which quantifies amino acid similarity based on large scale protein alignments. This score will give peptides with an identical match a similarity score of 1.0, regardless of the amino acid composition and length of the peptides. The bigger differences between two sequences are according to the BLOSUM matrix, the lower the score. **Figure 1** illustrates the ranges of similarity scores for peptides with varying numbers of amino acids exchanges. While a single amino acid exchange results in scores in the range from 0.901 to 0.987, (90% confidence interval displayed in Figure 1), for multiple exchanges the score range is much broader (from 0.536 to 0.771). The score ranges for a given number of substitutions are provided here as a reference of how the BLOSUM scores should be interpreted.

Correlating peptide immunogenicity with similarity to the human proteome

To identify if negative selection of self-reactive T cells reduces the immunogenicity of peptides that are similar to the human proteome, we calculated the peptide similarity of immunogenic peptides (epitopes) and negative peptides from our three datasets

to the human proteome. **Figure 2** shows the cumulative distribution of similarity scores for three datasets. For all three datasets, epitopes had a slightly but significantly lower median similarity to self-peptides than non-epitopes, as shown in **Table 1** (pertussis: 0.485 vs. 0.493, p = 0.049; tuberculosis: 0.507 vs. 0.515, p = 0.037; cockroach: 0.513 vs. 0.534, p = 0.008). This confirms that immunogenic MHC class II restricted peptides have a tendency to be less similar to self-peptides.

Correlating peptide immunogenicity with similarity to the human gut microbiome

Next, we assessed in an analog fashion if there was a detectable reduction in immune reactivity for peptides that had similar matches in the human gut microbiome. **Figure 3** shows the cumulative distribution of similarity scores for epitopes and non-epitopes from the three datasets. For both the pertussis and cockroach datasets, epitopes had a significantly lower median similarity to the gut microbiome compared to non-epitopes as assessed by a one-tailed Mann-Whitney test (pertussis: 0.558 vs. 0.571, p=0.012; cockroach: 0.580 vs. 0.599, p = 0.0002). The MTB dataset showed the same trend, but did not reach the significance of $p < 0.05$ in this test (0.584 vs. 0.590; p = 0.11). All median similarity scores can be seen in **Table 3**. Thus MHC class II restricted epitopes also have a tendency to be less similar to peptides found in the healthy gut microbiome.

Combining human and microbiome conservation scores

To determine if the similarity scores of a peptide to the human proteome and the human gut microbiome can be combined to better predict immunogenicity, we performed a linear regression of the two scores calculating the total score = offset -

$\beta_{microbiome}$ * $score_{microbiome}$ - β_{human} * $score_{human}$, where the three model parameters are 1) a constant offset, and 2) $\beta_{microbiome}$ and 3) β_{human} are weights of the gut microbiome and the human proteome similarity scores, respectively. The model parameters were fitted by calculating the total score of a set of peptides, and minimizing the squared difference to their immunogenicity, with immunogenic peptides set to 1.0 and non-immunogenic peptides set to 0.0. Fitted model parameters determined in 20-fold cross-validation are listed in **Table 4**. The fitted model parameters indicated that the microbiome score gets assigned nearly double the weight of the human proteome score (0.66 vs. 0.39), suggesting that it has higher predictive power in this model. The average cross-validated distance for the combined model is 0.1458. This distance is statistically significantly lower when comparing it with a model including only the human proteome score (distance = 0.1471; p = 0.025 (one-sided, paired t-test)), and shows the same trend but does not reach statistical significance when compared to a model including only the microbiome score (distance = 0.1459, p = 0.13). These data suggest that both scores provide independent information on the immunogenicity of a peptide, and that the microbiome score has higher predictive value in this simple model.

Discussion

Previous studies for MHC class I restricted epitopes had shown that there is evidence for negative selection against peptides that are similar to the human proteome (24). In this study, we have expanded those findings for MHC class II restricted epitopes, and in addition demonstrated for the first time that there is a correlation between a peptide's T cell immune reactivity and its conservation in the microbiome. This suggests that there is an imprint on the availability of T cells

recognizing certain peptide targets that shape the development of immune response against subsequent exposures.

Our study provides proof of principle that it should be possible to include similarity to self and the microbiome as selection factors into prediction pipelines for MHC class II restricted epitopes and adaptive T cell immunotherapy (25). Given the discrepancy between our ability to predict MHC class II binding and MHC class II T cell immunogenicity (14), any such additional factors are highly desirable. However, at the current stage, the magnitude of the detected differences in similarity to either the human proteome or the microbiome are very small. As a result, incorporating the current similarity scores into epitope prediction pipelines would be expected to only give marginal improvements. Additional improvements will be necessary, such as better understanding which species in the microbiome have selective impact on the epitope repertoire, or by developing better similarity matrices that quantify which amino acid substitutions are considered conservative in the context of T cell immune recognition. Similarly, the present study was limited to three datasets, representing different types of antigen exposure (infection, vaccination and allergen exposure). Future studies will explore in much more detail if and how this observation holds in different antigenic systems, and how the scores can best be combined with other factors to derive overall immunogenicity predictions.

Acknowledgements

This work has been funded by National Institutes of Health contract HHSN272201200010C.

Disclosures

The authors declare that they have no conflict of interest.

References

1. Karosiene, E., C. Lundegaard, O. Lund, and M. Nielsen. 2012. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 64: 177–86.

2. Kim, Y., J. Ponomarenko, Z. Zhu, D. Tamang, P. Wang, J. Greenbaum, C. Lundegaard, A. Sette, O. Lund, P. E. Bourne, M. Nielsen, and B. Peters. 2012. Immune epitope database analysis resource. *Nucleic Acids Res.* 40: W525–30.

3. Nielsen, M., C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, and O. Lund. 2008. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput. Biol.* 4: e1000107.

4. Peters, B., H.-H. Bui, S. Frankild, M. Nielsen, C. Lundegaard, E. Kostem, D. Basch, K. Lamberth, M. Harndahl, W. Fleri, S. S. Wilson, J. Sidney, O. Lund, S. Buus, and A. Sette. 2006. A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules. *PLoS Comput. Biol.* 2: e65.

5. Wang, P., J. Sidney, C. Dow, B. Mothé, A. Sette, and B. Peters. 2008. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol.* 4: e1000048.

6. Lankat-Buttgereit, B., and R. Tampe. 2002. The Transporter Associated With Antigen Processing: Function and Implications in Human Diseases. *Physiol Rev* 82: 187–204.

7. Paz, P., N. Brouwenstijn, R. Perry, and N. Shastri. 1999. Discrete Proteolytic Intermediates in the MHC Class I Antigen Processing Pathway and MHC I–

Dependent Peptide Trimming in the ER. *Immunity* 11: 241–251.

8. Craiu, A., T. Akopian, A. Goldberg, and K. L. Rock. 1997. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc. Natl. Acad. Sci.* 94: 10850–10855.

9. Kloetzel, P. M. 2001. Antigen processing by the proteasome. *Nat. Rev. Mol. Cell Biol.* 2: 179–87.

10. Larsen, M. V, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, and M. Nielsen. 2007. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* 8: 424.

11. Stranzl, T., M. V. Larsen, C. Lundegaard, and M. Nielsen. 2010. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62: 357–68.

12. Alexander, J., J. Sidney, S. Southwood, J. Ruppert, C. Oseroff, A. Maewal, K. Snoke, H. M. Serra, R. T. Kubo, and A. Sette. 1994. Development of high potency universal DR-restricted helper epitopes by modification of high affinity DR-blocking peptides. *Immunity* 1: 751–61.

13. Calis, J. J. A., M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Keşmir, and B. Peters. 2013. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* 9: e1003266.

14. Chaves, F. A., A. H. Lee, J. L. Nayak, K. A. Richards, and A. J. Sant. 2012. The utility and limitations of current Web-available algorithms to predict peptides recognized by CD4 T cells in response to pathogen infection. *J. Immunol.* 188: 4235–48.

15. Paul, S., J. Sidney, B. Peters, and A. Sette. 2014. Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology,*

and Health Informatics - BCB '14 ACM Press, New York, New York, USA. 733–738.

16. Carpenter, C., J. Sidney, R. Kolla, K. Nayak, H. Tomiyama, C. Tomiyama, O. A. Padilla, V. Rozot, S. F. Ahamed, C. Ponte, V. Rolla, P. R. Antas, A. Chandele, J. Kenneth, S. Laxmi, E. Makgotlho, V. Vanini, G. Ippolito, A. S. Kazanova, A. V Panteleev, W. Hanekom, H. Mayanja-Kizza, D. Lewinsohn, M. Saito, M. J. McElrath, W. H. Boom, D. Goletti, R. Gilman, I. V Lyadova, T. J. Scriba, E. G. Kallas, K. Murali-Krishna, A. Sette, and C. S. Lindestam Arlehamn. 2015. A side-by-side comparison of T cell reactivity to fifty-nine Mycobacterium tuberculosis antigens in diverse populations from five continents. *Tuberculosis (Edinb).* 95: 713–21.

17. Lindestam Arlehamn, C. S., A. Gerasimova, F. Mele, R. Henderson, J. Swann, J. A. Greenbaum, Y. Kim, J. Sidney, E. A. James, R. Taplitz, D. M. McKinney, W. W. Kwok, H. Grey, F. Sallusto, B. Peters, and A. Sette. 2013. Memory T cells in latent Mycobacterium tuberculosis infection are directed against three antigenic islands and largely contained in a CXCR3+CCR6+ Th1 subset. *PLoS Pathog.* 9: e1003130.

18. Oseroff, C., J. Sidney, V. Tripple, H. Grey, R. Wood, D. H. Broide, J. Greenbaum, R. Kolla, B. Peters, A. Pomés, and A. Sette. 2012. Analysis of T cell responses to the major allergens from German cockroach: epitope specificity and relationship to IgE production. *J. Immunol.* 189: 679–88.

19. Dillon, M. B. C., V. Schulten, C. Oseroff, S. Paul, L. M. Dullanty, A. Frazier, X. Belles, M.-D. Piulachs, C. Visness, L. Bacharier, G. R. Bloomberg, P. Busse, J. Sidney, B. Peters, and A. Sette. 2015. Different Bla-g T cell antigens dominate responses in asthma versus rhinitis subjects. *Clin. Exp. Allergy* 45: 1856–67.

20. The UniProt Consortium. 2014. UniProt: a hub for protein information. *Nucleic Acids Res.* 43: D204–212.

21. The Human Microbiome Project Consortium, -. 2012. A framework for human

microbiome research. *Nature* 486: 215–21.

22. The Human Microbiome Project Consortium, -. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207–14.

23. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89: 10915–10919.

24. Calis, J. J. A., R. J. de Boer, and C. Keşmir. 2012. Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLoS Comput. Biol.* 8: e1002412.

25. Haase, K., S. Raffegerst, D. J. Schendel, and D. Frishman. 2015. Expitope: a web server for epitope expression. *Bioinformatics* 31: 1854–6.

Figure Legends

Figure 1 – **The relationship between number of amino acid exchanges and the BLOSUM score**. Twenty thousand peptides were randomly selected from UniProt proteins and compared to the human gut microbiome in terms of BLOSUM score. The figure shows the average and 5^{th} percentile – 95^{th} percentile range of the generated scores as a function of the number of amino acid substitutions between the peptides.

Figure 2 – Similarity of epitopes and non-epitopes to the human proteome. Each panel shows the cumulative distribution of similarity scores for epitopes (blue line) and non-epitopes (red-line). The different panels depict peptides from A) *B. pertussis*, B) *M. tuberculosis* and C) cockroach.

Figure 3 – Similarity of epitopes and non-epitopes to proteins encoded by microbes found in healthy human gut. Each panel shows the cumulative distribution of similarity scores for epitopes (blue line) and non-epitopes (red-line). The different panels depict peptides from A) *B. pertussis*, B) MTB and C) cockroach.

Tables

Dataset	#positive peptides	#negative peptides	#intermediate peptides	#donors tested
M. tuberculosis	79	523	148	61
B. pertussis	142	300	206	31
Cockroach	59	437	170	90

Table 1: Peptide immunogenicity datasets

Dataset	median similarity score, positive peptides (Standard deviation)	median similarity score, negative peptides (Standard deviation)
M. tuberculosis	0.507 (0.048)	0.515 (0.048)
B. pertussis	0.485 (0.041)	0.493 (0.047)
Cockroach	0.513 (0.094)	0.534 (0.107)

Table 2: Median similarity scores of epitopes and non-epitopes to the human proteome.

Dataset	median similarity score, positive peptides (Standard deviation)	median similarity score, negative peptides (Standard deviation)
M. tuberculosis	0.584 (0.044)	0.590 (0.059)
B. pertussis	0.558 (0.052)	0.571 (0.053)
Cockroach	0.580 (0.046)	0.599 (0.073)

Table 3: Median similarity scores of epitopes and non-epitopes to the human gut microbiome.

Model	Parameter fit (standard deviation)			Distance
	offset	β microbiome	β human	
both parameters	0.78 (0.02)	0.66 (0.04)	0.39 (0.03)	0.1458
microbiome only	0.69 (0.02)	0.85 (0.04)	N/A	0.1459
human only	0.50 (0.02)	N/A	0.6 (0.03)	0.1471

Table 4: Linear regression combining microbiome and human proteome scores to predict immunogenicity.