Mono-allelic immunopeptidomics data from 109 MHC alleles reveals variability in binding preferences and improves neoantigen prediction algorithm

#5640

Rachel Marty Pyke[1], Steven Dea[1], Hima Anbunathan[1], Charles W. Abbott[1], Neeraja Ravi[1], Jason Harris[1], Gabor Bartha[1], Sejal Desai[1], Rena McClory[1], John West[1], Michael P. Snyder[2], Richard Chen[1] and Sean Michael Boyle[1]
Affiliations: [1]Personalis, Inc. Menlo Park, CA; [2]Stanford University, Palo Alto, CA

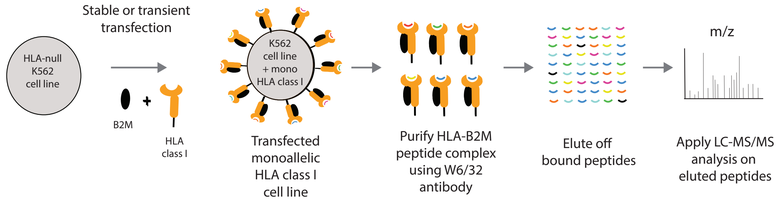
Contact:
Rachel.Pyke@personalis.com
Sean.Boyle@personalis.com

I. Introduction

Sequence variability in the major histocompatibility complex (MHC) leads to the presentation of diverse neoantigens to T cells. Understanding this diversity is a critical component of improving neoantigen-based biomarkers and designing effective personalized cancer vaccines. Previously, we published data from 25 mono-allelic cell lines and built an associated MHC class I, pan-allelic binding prediction algorithm (SHERPA™)[1]. Here, we profile an additional 84 MHC alleles including 37 that have never previously been profiled with mono-allelic immunopeptidomics, improve neoantigen presentation prediction of the SHERPA algorithm and explore the impact of MHC variability on peptide binding.

II. Immunopeptidomics data generation

1. Generation of immunopeptiomics training data

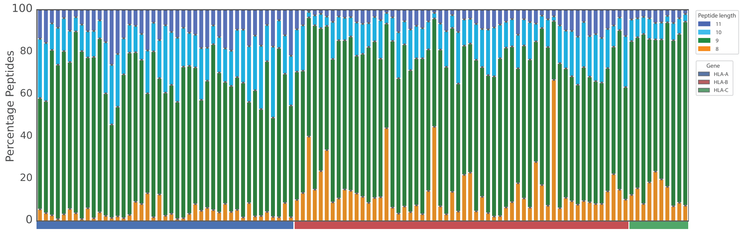


To generate the data, we stably and transiently transfected a total of 109 different MHC alleles (43 HLA-A, 56 -B and 10 -C alleles) into independent K562 HLA-null cell lines, immunoprecipitated intact MHC complexes using a W6/32 antibody, and profiled the bound peptides using LC/MS-MS (Figure 1).

III. Immunopeptidomics data overview

We recovered a median of 1430 peptides per allele, with yields from the transient transfections being consistently higher than the stable transfections. However, we also observed a substantially stronger preference for peptides originating from either terminus of the protein in the stable than the transient transfections, suggesting that the larger quantity of MHC in transient transfections may be altering the profile of the peptides presented on the cellular surface. Moreover, in accordance with tryptic and chemotrypic enzymatic activity of proteasomal cleavage, we observed an enrichment of lysine and arginine on the C-terminal end of the peptides across all alleles. We also show that peptides associated with HLA-A alleles were the longest, followed by HLA-B and HLA-C respectively (Figure 2).

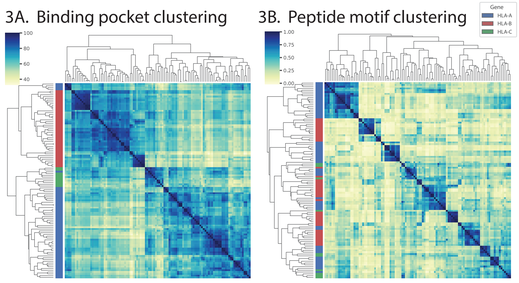
2. Peptide length distribution across all alleles



III.Immunopeptidomics data overview (continued)

3A. Binding pocket clustering

3B. Peptide motif clustering

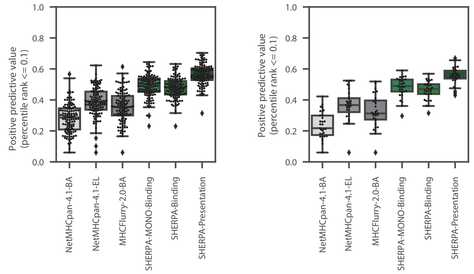


We clustered the 109 profiled MHC alleles by similarity in their 34-mer amino acid binding pocket (Figure 3A). As expected, we see that the alleles cluster strongly within their parental gene (HLA-A/B/C) with a few exceptions. On the other hand, we observe much higher mixture between parental genes when we cluster MHC motifs derived from the allele-specific peptides (Figure 3B). These heatmaps highlight that peptide motifs are shared across MHC genes and suggest that amino acids within the MHC binding pocket have variable importances to peptide binding.

IV. SHERPA model and performance

4A. All alleles

4B. Unpublished alleles

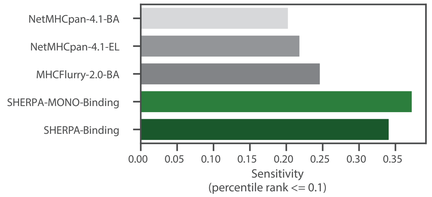


In addition to the 109 mono-allelic cell lines used for the SHERPA-MONO-Binding algorithm, SHERPA-Binding increases generalizability by systematically integrating an additional 104 mono-allelic and 384 multi-allelic samples with publicly available immunopeptidomics data and binding assay data. The 186 alleles in the resulting training dataset have an average allelic coverage of 98% across 18 different ethnicities represented in the United States. We evaluated our updated performance on 10% of the mono-allelic immunopeptidomics data that was held-out from training.

The positive predictive value (PPV) of SHERPA was markedly higher than either NetMHCPan 4.1[2] or MHCFlurry-2.0[3] (1.45 and 1.58-fold increase, respectively) (Figure 4A), with even further gains when only the 37 previously unprofiled alleles were considered (1.51 and 1.79-fold increase, respectively) (Figure 4B).

V. Immunogenicity performance

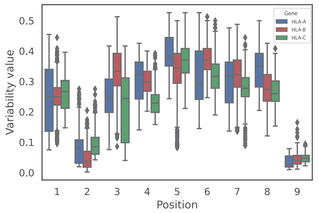
5. Immunogenicity



Though SHERPA's performance on mono-allelic immunopeptidomics data provides strong evidence of cell surface presentation, we wanted to further evaluate SHERPA's utility for cancer applications through its ability to recognize peptides that elicit an immune response. Thus, we evaluated the ability of the algorithms to identify immunogenic epitopes using a published dataset[4]. SHERPA was able to detect 1.38-fold more immunogenic epitopes than either other method (Figure 5).

VI. Model interpretability

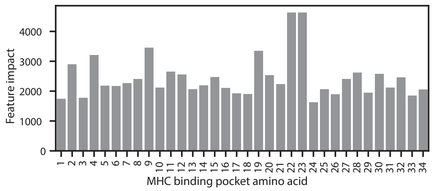
6A. Peptide residue variability



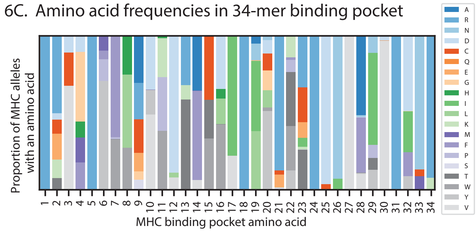
Using SHERPA's pan-allelic capability, we explored MHC binding trends across the entire space of observed MHC alleles. From a large set of random peptides, we identified 500 peptides per MHC allele that were predicted to bind the most strongly. We observed that nearly all alleles have a strong anchor residue in the ninth position, but the positions of the secondary anchor residue vary by gene. HLA-B showed a stronger preference for the second position while HLA-A exhibited more variability across the first, second and third positions (Figure 6A).

Finally, we performed predictions with SHERPA across millions of synthetic binding pockets and peptides to elucidate the impact of MHC variability on peptide diversity. We generated a feature impact score for each MHC binding pocket residue, identifying positions 22 and 23 of the binding pocket to be the most influential (Figure 6B).

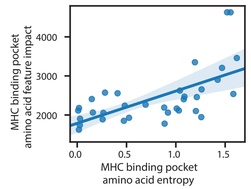
6B. Influence of amino acids in binding pocket



6C. Amino acid frequencies in 34-mer binding pocket



6D. Correlation between amino acid influence and entropy



Interestingly, we found a strong correlation between binding pocket positions that highly influence peptide binding and those that are highly diverse across the space of all MHC alleles, suggesting that influential residues experience the strongest divergent evolutionary pressure (Figures 6C & 6D).

VII. Conclusions

In conclusion, we profiled 109 mono-allelic cell lines, showed key trends in MHC-associated peptides, improved the SHERPA neoantigen prediction model and demonstrated the variable importance of binding pocket positions to peptide binding.

VIII. References

1. Pyke, R. M. et al. Precision Neoantigen Discovery Using Large-scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation. Mol. Cell. Proteomics 20, 100111 (2021).
2. 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 Research vol. 48 W449–W454 (2020).
3. O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. Cell Systems vol. 11 418–419 (2020).
4. Chowell, D. et al. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. Proc. Natl. Acad. Sci. U. S. A. 112, E1754–62 (2015).

Personalis