Precision neoantigen discovery using novel algorithms and expanded HLA-ligandome datasets

Dattatreya Mellacheruvu*, Rachel Marty Pyke*, Charles Abbott, Steven Dea, Nick Phillips, Sejal Desai, Rena McClory, John West, Richard Chen and Sean Michael Boyle

Personalis, Inc. | 1330 O'Brien Dr., Menlo Park, CA 94025

* co-authors

Abstract #57

Contact:

Datta.Mellacheruvu@personalis.com

Rachel.Pyke@personalis.com

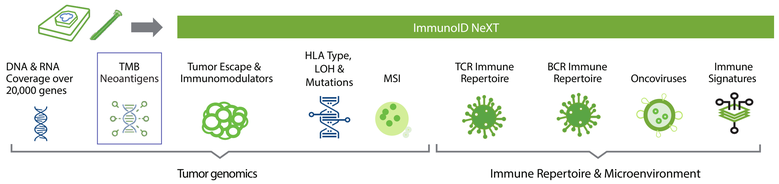
Sean.Boyle@personalis.com

I. Introduction and background

Personalis Immunopeptidomics + Personalis Machine Learning Model = Personalis Neoantigen Presentation Prediction

Technologies for neoantigen discovery are critical for developing personalized cancer vaccines and neoantigen-based biomarkers. Precision neoantigen discovery entails comprehensive detection of tumor-specific genomic variants and accurate prediction of MHC presentation of epitopes originating from such variants. Our ImmunoID NeXT™ Platform enables a comprehensive survey of putative neoantigens by combining highly sensitive and exome scale DNA and RNA sequencing with advanced analytics. Here, we present **S**ystematic **HLA** **E**pitope **R**anking **P**an **A**lgorithm (SHERPA™), our pan-predictive machine learning model for predicting MHC class I presentation and identifying potentially immunogenic patient-specific neoantigens.

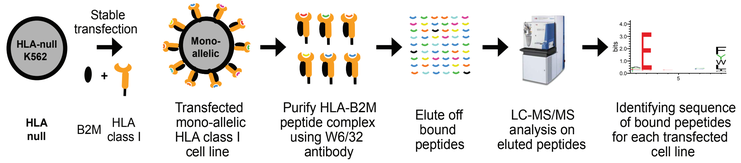
II. Overview of NeoantigenID and ImmunoID NeXT



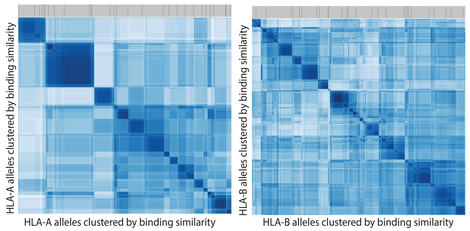
The upstream detection of cancer-specific somatic alterations is enabled by ImmunoID NeXT, our platform to analyze both a tumor and its microenvironment from a single tumor sample and a paired normal. Highly sensitive and accurate detection of putative neoantigens, and evaluation of their expression using transcriptome data, is a critical aspect of precision neoantigen discovery.

III. In-house mono-allelic immunopeptidomics data

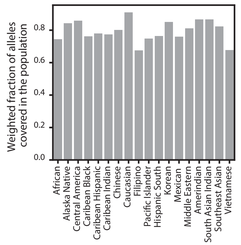
A High-quality in-house immunopeptidomics data using mono-allelic cell lines



B Allelic diversity in immunopeptidomic data



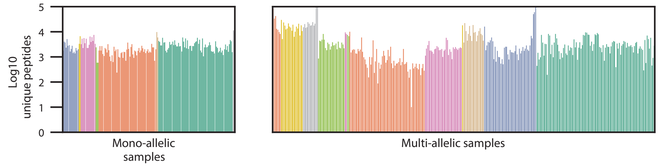
C Population coverage



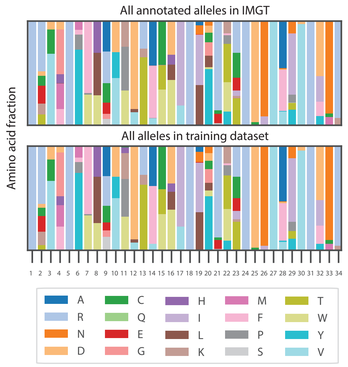
To train our algorithms, we generated high-quality and unambiguous training data using approximately 60 genetically engineered mono-allelic K562 cell lines (A). Briefly, MHC-peptide complexes were immunoprecipitated using W6/32 antibody followed by peptide elution and sequencing using tandem mass spectrometry. Our alleles, visualized on a clustered heatmap of all known IMGT/HLA alleles based on binding pocket similarity, effectively capture binding pocket diversity (B). The population coverage of our mono-allelic dataset, estimated using allele frequencies from Allele Frequencies Net Database, is robust across several ethnic world populations (C).

IV. Expanding the scale and scope of our HLA-ligandome

A Peptide identifications from the expanded set of high-quality mono- and multi-allelic samples

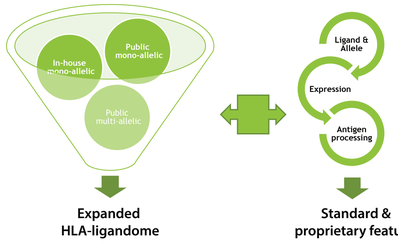


B Binding pocket representation



Recent advances in immuno-oncology, personalized immunotherapies in particular, and improvements in mass spectrometry-based immunopeptidomics led to a renewed scientific interest in neoantigens. As a result, vast amounts of HLA-ligandome data is being generated. We reasoned that integrating high-quality publicly available data, generated on diverse tissues and cell lines, will improve the performance of our prediction algorithms. Accordingly, we systematically curated both mono- and multi-allelic data from several publicly availble projects (n=14), appended it to in-house data and created an expanded HLA-ligandome. The scale of this expanded dataset, measured in terms of unique peptide counts, is significantly higher compared to any single dataset (A). The scope, measured in terms of number of alleles (data not shown) and binding pocket representation, is also improved. Briefly, we observed a very high degree of concordance in the diversity of amino-acid representation of the binding pocket between alleles in our expanded dataset and the complete set of alleles in IPD-IMGT/HLA Database (B).

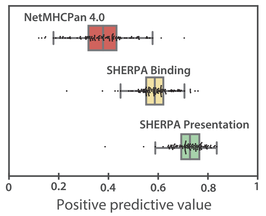
V. Modeling MHC-peptide binding and presentation



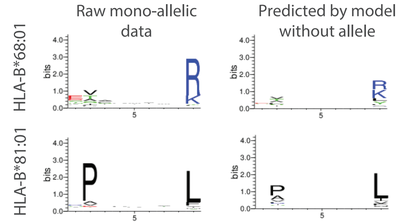
We modeled both MHC-peptide binding and presentation using our expanded HLA ligandome. Presentation includes binding and the upstream antigen processing. Our binding model comprises standard features, e.g. peptide sequences of ligands and binding sequences of their cognate alleles. We developed novel and proprietary features to model antigen processing, that take into account the expression of source protein. We trained pan-allelic prediction models, for both binding and presentation, using proprietary algorithms that benefit from the availability of large training data sets.

VI. Evaluating the performance of SHERPA

A Performance on 10% held-out data



B Pan-allelic prediction capabilities of SHERPA



We evaluated the performance of our expanded and composite models using 10% held-out data, mixed with decoys in a 1:999 ratio. The positive predictive value in the top 0.1% predicted ligands by our models, both binding and presentation, is significantly higher compared to NetMHCPan 4.0, a state-of-the-art publicly available tool (A). We also evaluated the pan-allelic performance of our presentation model using a leave-one-out approach, and observed a high-degree of concordance between motifs in raw data and motifs predicted by models that do not include that allele (B). Motifs were generated using WebLogo.

VII. Summary and conclusions

We have generated high-quality and unambiguous immunopeptidomics data using mono-allelic cell lines. Additionally, we expanded the scale and scope of our training data using curated public data. Our pan-allelic prediction algorithms, that model both MHC-peptide binding and presentation using a combination of standard and novel proprietary features, have a significantly higher positive predictive value compared to NetMHCPan 4.0. Taken together, the high sensitivity and specificity of SHERPA enables precision neoantigen discovery, with applications to the development of personalized immunotherapies and biomarker discovery.

Personalis®