Comprehensive and accurate prediction of presented neoantigens using ImmunoID NeXT and advanced machine learning algorithms

Dattatreya Mellacheruvu*, Rachel Marty Pyke*, Charles Abbott, Nick Phillips, Rena McClory, John West, Richard Chen and Sean Michael Boyle (*co-first authors)
Personalis, Inc. | 1330 O'Brien Dr., Menlo Park, CA 94025

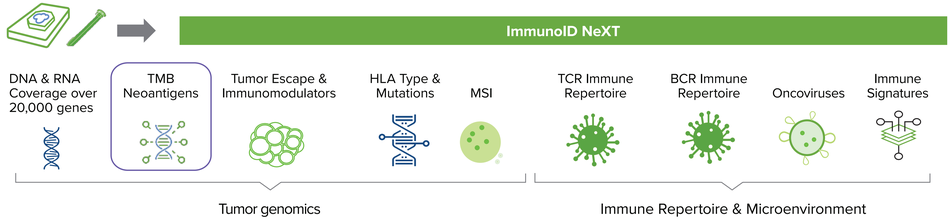
Contact: Datta.Mellacheruvu@personalis.com
Rachel.Pyke@personalis.com
Sean.Boyle@personalis.com

Background

Comprehensive detection of potential neoantigens and accurate prediction of their MHC presentation are critical prerequisites for selecting neoepitopes that can be used for creating personalized cancer vaccines. However, prediction models developed using in-vitro MHC-peptide binding assays cannot model upstream presentation machinery, such as proteasome cleavage and peptide loading. Advances in immuno-affinity purification followed by mass spectrometry (IP-MS) have enabled direct detection of MHC-bound peptides and can therefore be used for modelling native MHC-peptide presentation. Furthermore, genetically engineered cell lines that express a single HLA allele enable unambiguous HLA-peptide assignment. Here, we present an overview of our MHC presentation prediction framework based on a large collection of such mono-allelic cell lines and discuss its utility in conjunction with ImmunoID NeXT, our commercially available exome scale DNA and RNA sequencing and analytics platform specifically designed to enable the development of immunotherapies.

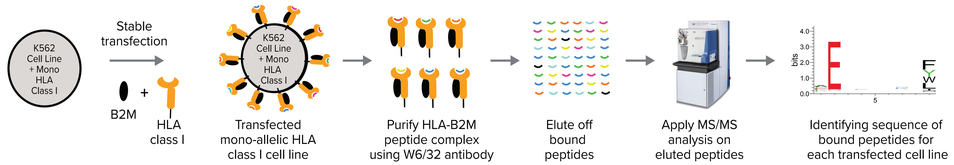
ImmunoID NeXT platform and neoantigen prediction

The ImmunoID NeXT platform provides joint tumor genomics and immune profiling from a single tumor/normal sample. Through the identification of somatic mutations (SNVs, indels and fusions), HLA types, RNA expression values and several other readouts, the platform provides a broad set of features for identifying and ranking potentially immunogenic peptides.



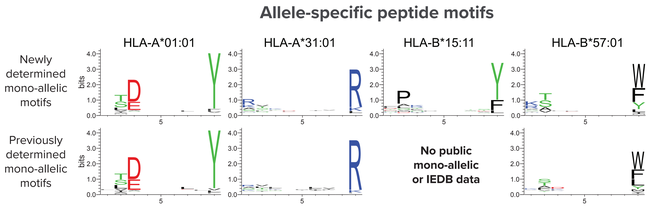
Immunopeptidomics with mono-allelic HLA class I cell lines

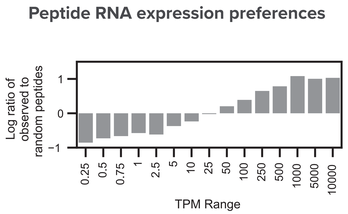
We selected nearly 60 HLA class I alleles that exist in high frequencies across multiple populations. For each of these alleles, we generated a mono-allelic cell with a stable transfection of the allele into K-562 null-HLA parental cells. Cells were grown, screened for surface expression, lysed and immuno-affinity purified using a column coated with HLA class I (W6/32) antibody. Peptides were gently eluted and analyzed using LC-MS/MS. Peptide-to-spectrum assignment was performed and filtered at 1% false discovery rate.

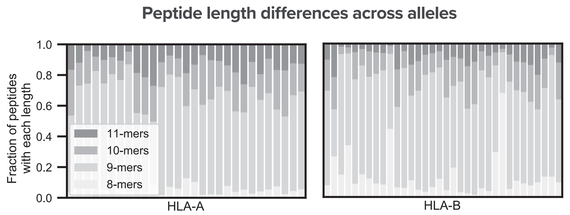


High quality immunopeptidomic data

Our immunopeptidomics data is of high quality and comprehensive. For alleles with public mono-allelic data, our new motifs agree with published motifs and often have even stronger definition at anchor positions. Furthermore, we profile 40 new alleles without public mono-allelic data, several of which do not have data in IEDB. Across our data, we find a strong preference for peptides with high RNA expression and several allele-specific peptide length preferences.

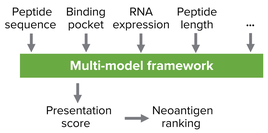


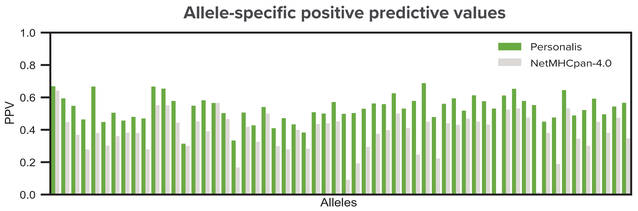


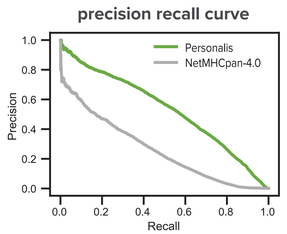


A machine learning model with high accuracy

Our prediction framework is based on multiple modelling algorithms, including a multi-layer neural network, and uses proprietary and standard features such as peptide sequence, peptide length, binding pocket sequence and abundance (measured by RNA expression). We created a pan-allelic model and evaluated it on an independent hold-out dataset. Our pan-allelic model had superior performance compared to the public gold standard NetMHCpan-4.0, with a higher precision across a range of recall (sensitivity) values. Furthermore, we show higher 1% positive predictive value (PPV) than NetMHCpan-4.0 across every allele.

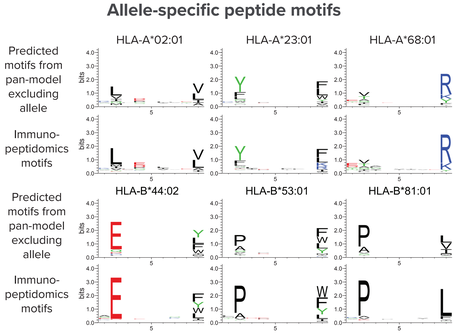






Predictions generalize to unseen alleles

To assess the ability of our pan-allelic model to accurately predict neoantigens for alleles without training data, we retrained our prediction algorithm several times, excluding data from a different allele with each training. We then compared the predicted motifs of the excluded allele with our actual immunopeptidomics data for that allele. Here, we show that our pan-allelic model is capable of learning motif patterns for unseen alleles.



****Motifs generated by WebLogo*

Conclusion

We present here our pipeline for generating high quality mono-allelic data and our MHC presentation prediction model that has high accuracy and generalizes to unseen HLA alleles. This module, which will be integrated into our ImmunoID NeXT platform, enables effective discovery of neoepitopes that is critical for developing personalized cancer vaccines.

Personalis®
Pioneering Genome-Guided Medicine