Accurate modeling of antigen processing and MHC peptide presentation using large-scale immunopeptidomes and a novel machine learning framework

Rachel Marty Pyke[1]*, Dattatreya Mellacheruvu[1]*, Steven Dea[1], Charles Abbott[1], Nick Phillips[1], Sejal Desai[1], Rena McClory[1], Steven Ketelaars[2], Pia Kvistborg[2], John West[1], Richard Chen[1] and Sean Michael Boyle[1]. *co-first authors

[1]Personalis, [2]Netherlands Cancer Institute

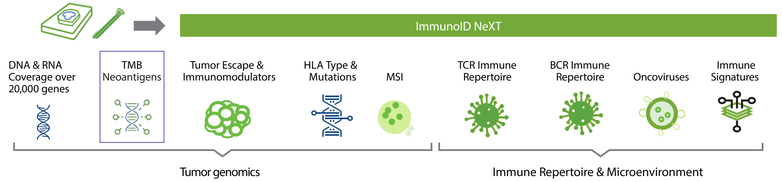
Contact:
Rachel.Pyke@personalis.com
Sean.Boyle@personalis.com

#1898

I. Introduction

Neoantigens, which are antigens specific to cancer cells, can be harnessed to develop personalized cancer vaccines and prognostic biomarkers for checkpoint blockade inhibition. Next generation sequencing technologies have enabled comprehensive profiling of putative neoantigens by interrogating the tumor exome and transcriptome, but accurate prediction of peptides presented by MHC complexes remains a significant challenge. Here, we present **S**ystematic **HL**A **E**pitope **R**anking **P**an **A**lgorithm (SHERPA ™) that addresses this critical need. SHERPA comprises highly sensitive, accurate and pan-allelic MHC-peptide (MHCp) binding and presentation prediction models, that were built using a multi-pronged strategy.

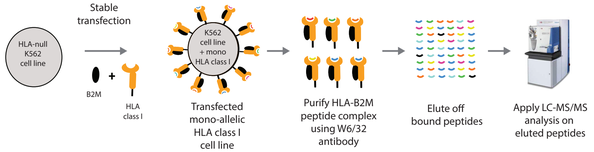
II. Comprehensive neoantigen profiling using ImmunoID NeXT



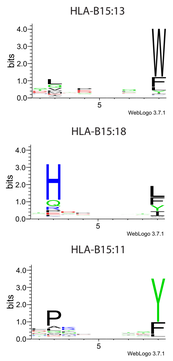
Our immuno-oncology platform (ImmunoID NeXT) enables researchers to analyze both a tumor and its microenvironment from a single tumor sample. In-depth interrogation of tumor and normal samples and identification of tumor-specific genomic events allows us to comprehensively profile the landscape of potential neoantigens, a critical aspect of precision neoantigen discovery.

III. Mono-allelic immunopeptidomics

A. Generation of immunopeptiomics training data



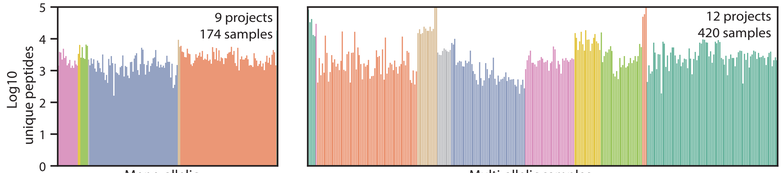
B. Motif examples



To provide high-quality data to our machine learning model, we generated a large-scale HLA ligandome using approximately 75 stably and transiently transfected mono-allelic K562 cell lines. MHC-peptide complexes were immunoprecipitated using W6/32 antibodies followed by peptide elution and peptide sequencing using tandem mass spectrometry. We profiled several novel alleles to increase the allelic coverage in under represented populations and to improve our prediction capability of unseen alleles. Three of these novel alleles included HLA-B*15:13, HLA-B*15:18 and HLA-B*15:11 (pictured to the right), which are a part of the same sub-type but have very different motifs.

IV. Overview of training data and prediction models

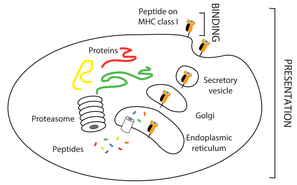
A. Integration of generated and publicly available data



We expanded the scale and scope of our in-house dataset using a large, systematically reprocessed and curated repository of publicly available mono- and multi-allelic datasets resulting in >180 alleles and >1.6 million peptides. Integrating data from diverse cell line and tissue types improved the generalizability of our models, a critically important aspect when applying our models to patient samples.

V. Modeling binding and presentation

A. Binding and presentation models

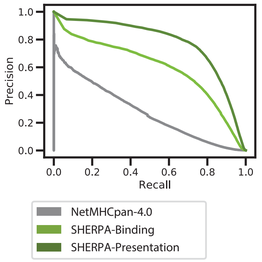


Briefly, MHCp binding was modeled using the amino acid sequences of the ligand and the binding pocket of the cognate allele. MHCp presentation, which encompasses in vivo antigen processing, was modeled using multiple features including the expression level of the source protein, proteasomal cleavage, and two novel features representing presentation propensities of genes and regions within gene bodies. We implemented a model-based deconvolution of multi-allelic datasets to generate pseudo mono-allelic data, and developed an integrative machine learning architecture to model our expanded HLA-ligandome.

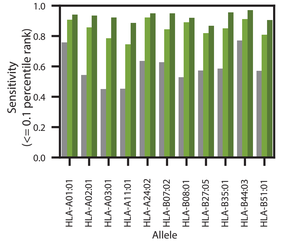
VI. Performance of SHERPA on held-out mono-allelic data

The performance of SHERPA was first evaluated using 10% of the held-out immunopeptidomics data mixed with synthetic negative examples in a 1:999 ratio. SHERPA models have higher precision over all recall values compared to NetMHCPan-4.0, a state-of-the-art publicly available tool and significantly higher sensitivites across the most frequent alleles.

A. Positive predictive value (0.1%)



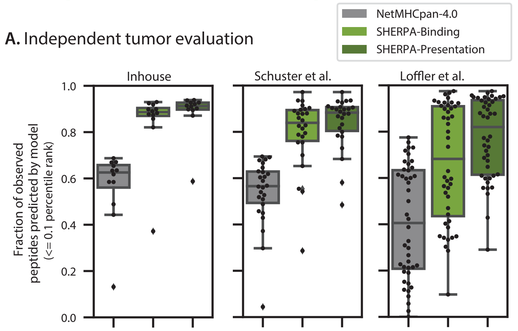
B. High frequency alleles



VII. Performance of SHERPA on independent tumor samples

SHERPA has pan-allelic prediction capabilities, so we further evaluated the performance of SHERPA on tumor samples with some alleles not present in our dataset. We performed both ImmunoID NeXT and immnopeptidomics on the same tissue samples. Then, we calculated patient-specific scores for each antigen by aggregating prediction scores across all HLA alleles in the sample. On 12 tissue samples profiled in-house, the SHERPA presentation model had a consistently high recall (90%) compared to NetMHCpan 4.0 (63%). The same trend holds true on external immunopeptidomics datasets from tumor samples.

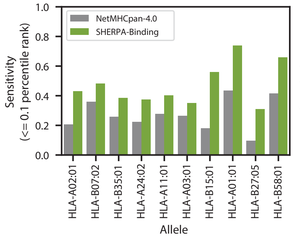
A. Independent tumor evaluation



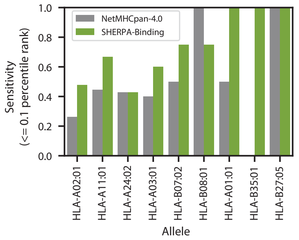
VIII. Performance of SHERPA on immunogenic epitopes

To elicit an immunogenic response, an epitope must be both presented on the cellular surface and recognized by a CD8+ T cell. Though SHERPA does not specifically predict immunogenicity, we evaluated the performance of SHERPA on two datasets of immunogenic peptides. SHERPA-Binding is able to recover a higher fraction of eptiopes than NetMHCpan-4.0 at the same percentile rank. Of note, SHERPA recovers over twice as many epitopes as NetMHCpan-4.0 for HLA-A*02:01. Alleles are ordered according to the number of immunogenic peptides in the dataset.

A. Chowell et al dataset



B. EBV/FLU verified dataset



*Unable to evaluate SHERPA-Presentation due to insufficient features provided by the datasets.

IX. Summary and concluding remarks

In conclusion, we present SHERPA, a machine learning-based prediction model for neoantigen discovery. To power SHERPA, we created a high-quality immunopeptidomics data from genetically engineered monoallelic cell lines and integrated large-scale publicly available data. SHERPA has consistently higher performance in comparison to the widely accepted and state-of-the-art publicly available tool on held-out mono-allelic data, tumor samples and immunogenic epitopes.

Personalis