# SHERPA™ Bioinformatics Engine

Systematic HLA Epitope Ranking Pan Algorithm

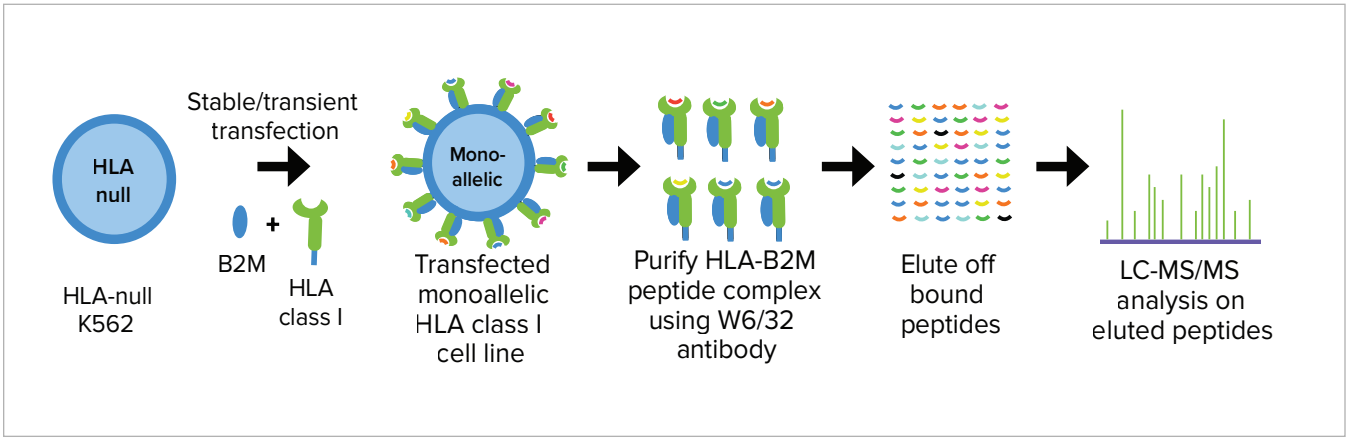
## Introduction

Technologies for neoantigen discovery are critical for developing personalized cancer vaccines and neoantigen-based biomarkers. Precision neoantigen discovery entails the comprehensive detection of tumor-specific genomic variants and accurate prediction of MHC presentation of epitopes originating from such variants. Personalis' ImmunoID NeXT™ enables a comprehensive survey of putative neoantigens by combining highly sensitive exome scale DNA and RNA sequencing with advanced analytics. Personalis developed the Systematic HLA Epitope Ranking Pan Algorithm (SHERPA) to predict MHC class I binding and presentation and to identify potentially Immunogenic patient-specific neoantigens[1].

### Generating Monoallelic Immunopeptidomics Training Data

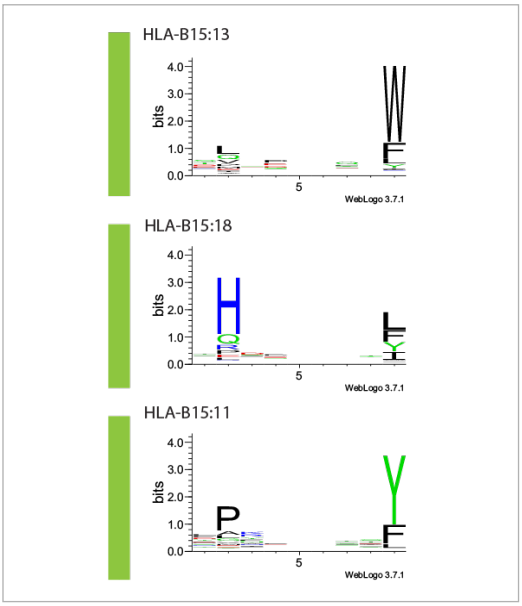
To provide high-quality data to the machine learning model, a large-scale HLA ligandome was generated using approximately 70 stably and transiently transfected mono-allelic K562 cell lines. MHC-peptide complexes were immunoprecipitated using W6/32 antibodies followed by peptide elution and peptide sequencing using tandem mass spectrometry (**Figure 1**). Several novel alleles were profiled to increase the allelic coverage in underrepresented populations and to improve the prediction capability of alleles for which no direct training data is available. Three of the included novel alleles: HLA-B*15:13, HLA-B*15:18, and HLA-B*15:11 (**Figure 2**), which are part of the same subtype, were found to have very different motifs.

**Figure 1: Overview of SHERPA Training Data Set**




Personalis®

**Figure 2: Motif Examples**



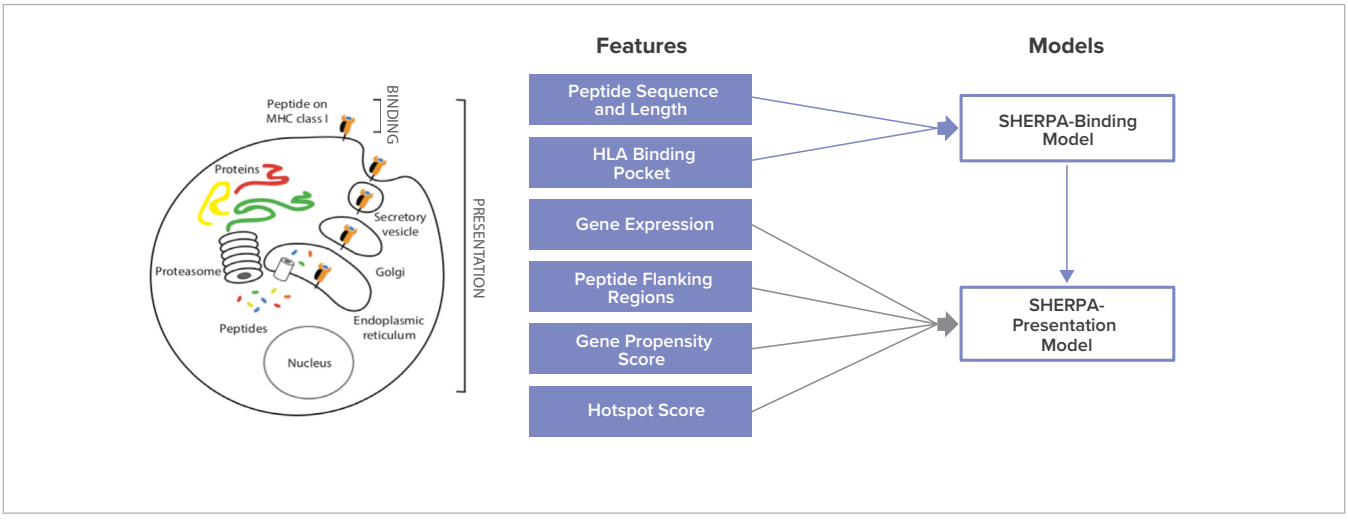

## Integration of Publicly Available Data Into SHERPA Training Data Set

The scale and scope of the in-house dataset was expanded using a large, systematically reprocessed and curated repository of >500 publicly available mono- and multi-allelic samples as well as binding affinity data from the Immune Epitope Database (IEDB). This combined approach resulted in one of the largest training datasets, consisting of 180 unique human alleles and >1.4 million positive peptides. Integrating data from diverse cell lines and tissue types improved the generalizability of SHERPA, a critically important aspect when applying these models to patient samples.

## Modeling Binding and Presentation

Briefly, pMHC binding was modeled using the amino acid sequences of the ligand and the binding pocket of the cognate allele. pMHC 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 (**Figure 3**). A model-based deconvolution of multi-allelic datasets were implemented to generate pseudo mono-allelic data and to develop an integrative machine learning architecture to model the expanded HLA-ligandome.

**Figure 3: SHERPA-Binding and SHERPA-Presentation Prediction Models**

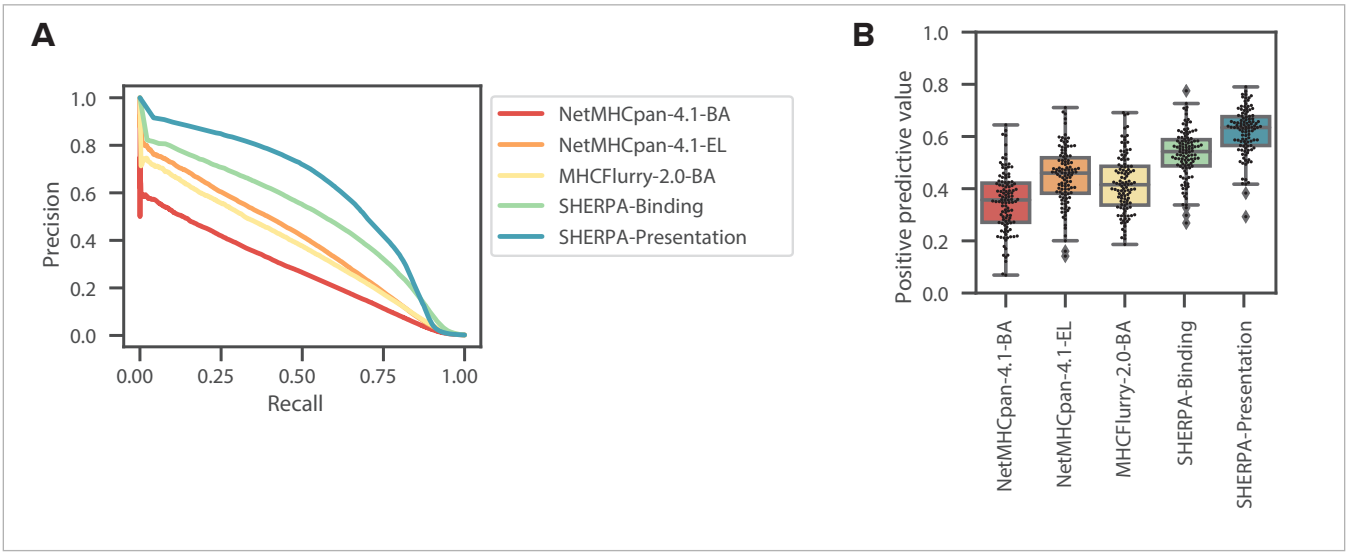


## Performance of SHERPA on Held-Out Monoallelic Data

The performance of SHERPA was evaluated on a ~10% held-out mono-allelic data set, mixed with negative examples in a 1:999 ratio (commonly assumed prevalence). The precision-recall curves demonstrate that SHERPA models have consistently higher precision at all recall values compared to other publicly available prediction algorithms (**Figure 4A**). Both SHERPA models also have better positive predictive values (PPV) across all alleles compared to publicly available prediction tools (**Figure 4B**). SHERPA-Presentation has a better PPV compared to SHERPA-Binding model, attesting to the utility of presentation-specific features.

Boxplots in **Figure 4B** denote the distributions of positive predictive values (top 0.1%) across alleles within the mono-allelic immunopeptidomics held-out test data. Distributions are shown to compare SHERPA with other publicly available models.

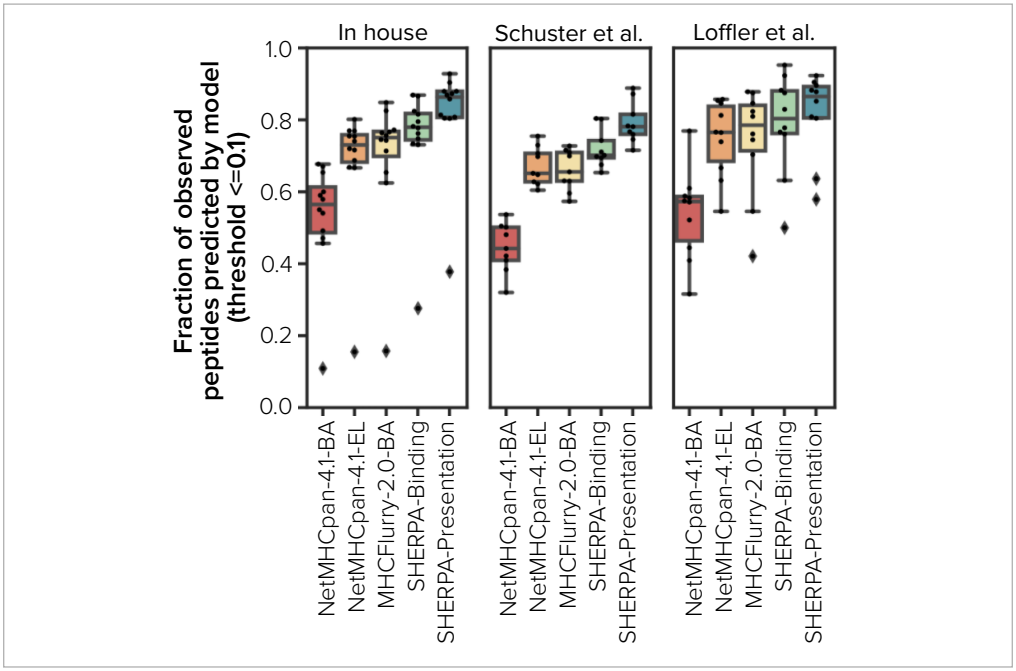
**Figure 4A and 4B: SHERPA Enables Superior Neoantigen Presentation Prediction**




## Performance of SHERPA on Independent Tissue Samples

SHERPA has pan-allelic prediction capabilities, so the performance of SHERPA was further evaluated on tumor samples with some alleles not present in the dataset. Both ImmunoID NeXT and immunopeptidomics were performed on the same tissue samples. Then, the patient-specific scores for each antigen were calculated 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 higher recall compared to NetMHCpan 4.1 and MHCFlurry 2.0. The same trend holds true on external immunopeptidomics datasets from tumor samples (**Figure 5**).

**Figure 5: SHERPA Models Demonstrate High Recall For Peptides From Independent Tissue Samples**




## Performance of SHERPA on Immunogenic Epitopes

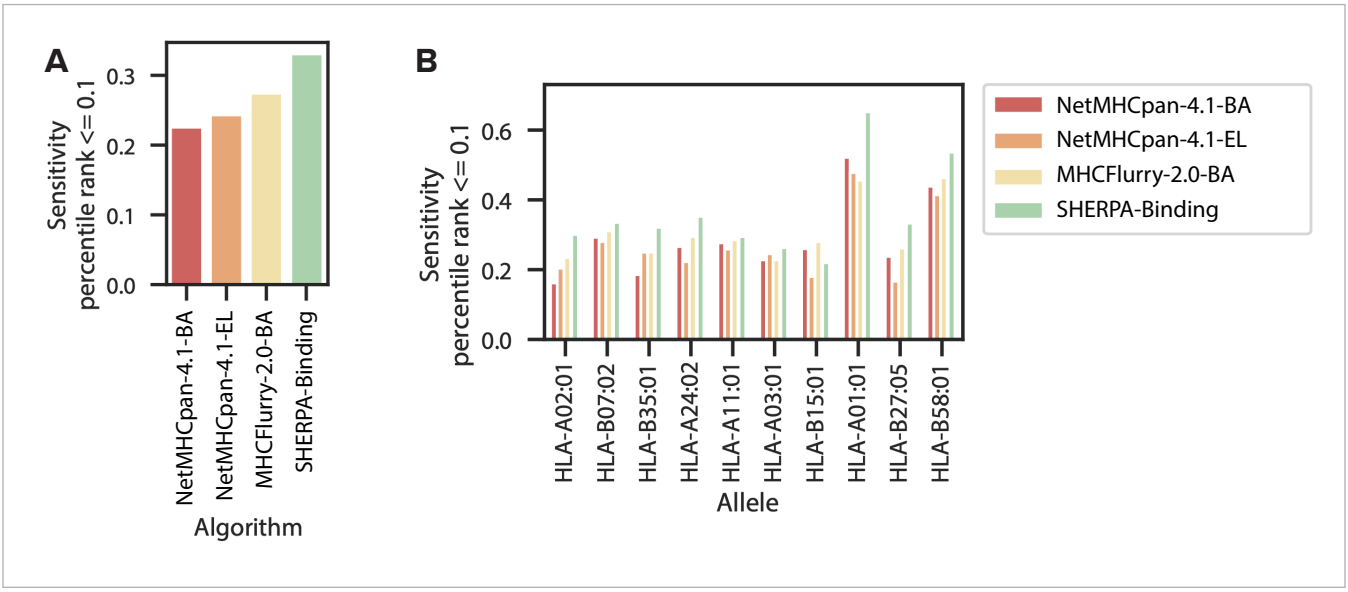
To elicit an immunogenic response, an epitope must be both presented on the cell surface and recognized by a CD8+ T cell. Though SHERPA does not specifically predict immunogenicity, the performance of SHERPA was evaluated on a dataset of >2,200 previously characterized immunogenic peptides from 75 alleles[2]. The SHERPA-Binding model was able to recover a higher fraction of epitopes than NetMHCpan-4.1 and MHCFlurry-2.0-BA at the same percentile rank (**Figure 6A**). Similar performance was observed across the high frequency alleles (**Figure 6B**).

## Developing Biomarkers for Immunotherapy Using SHERPA

Applying SHERPA to a cohort of 48 unresectable, stage III/IV melanoma patients treated with anti-PD-1 therapy, a clear difference was observed in the neoantigen burden between responders and non-responders (**Figure 7A**). Also, neoantigen burden is a powerful prognostic biomarker to stratify patients by progression free survival (**Figure 7B**).

While elevated measures of neoantigen burden predict in part which patients will benefit from immunotherapy, additional resistance mechanisms arising from the antigen presentation machinery

**Figure 6A and 6B: SHERPA Outperforms Other Algorithms in Identifying Immunogenic Peptides**




may further modulate immune response by diminishing capacity for neoantigen presentation. By accounting for these escape mechanisms that have been previously associated with reduced response to immunotherapy and combining them into a composite neoantigen prediction score (NEOPS™), a fuller representation of tumor antigen presentation to the immune system was captured, thereby increasing the predictive strength of this biomarker (**Figure 8A** and **8B**)[3].

**Figure 7A and 7B: Neoantigen Burden as a Powerful Prognostic Biomarker**

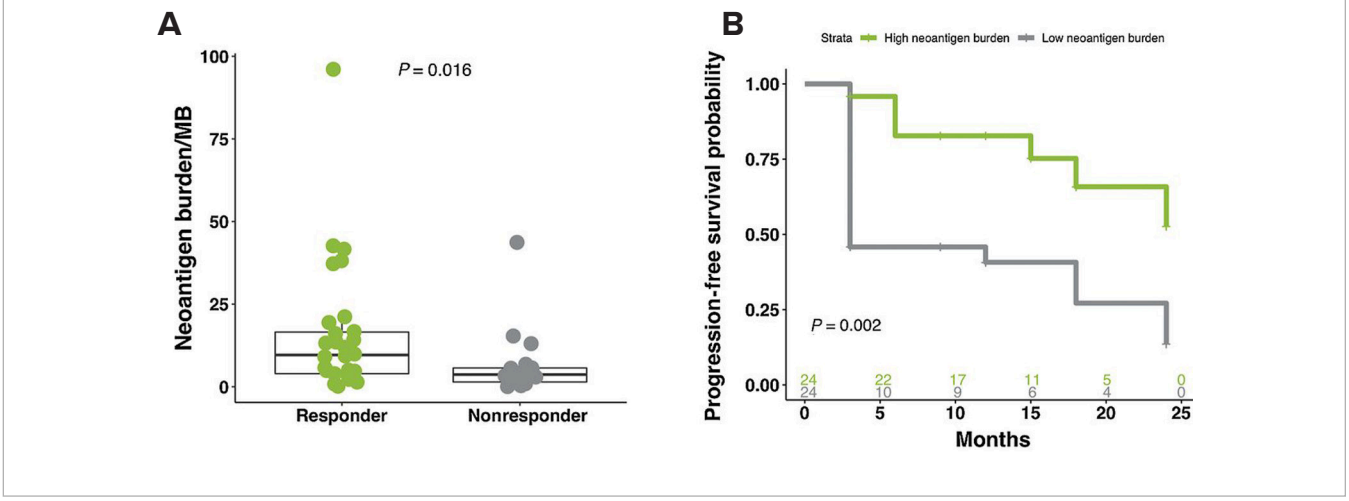
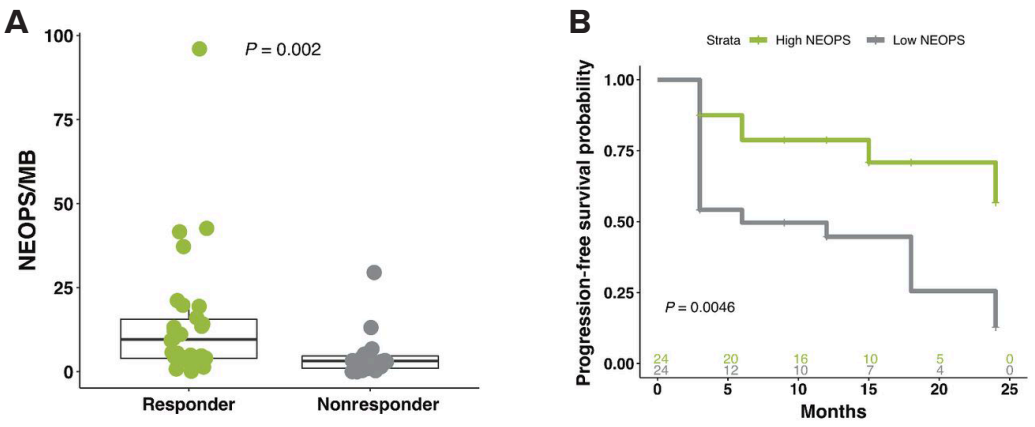


**Figure 8A and 8B: Neoantigen Prediction Score (NEOPS™) Increases the Predictive Strength of this Biomarker**




## Summary

In conclusion, high-quality immunopeptidomics data from genetically engineered monoallelic cell lines were generated and were integrated with large-scale publicly available data to develop SHERPA, a machine learning-based prediction model for neoantigen discovery. SHERPA has consistently higher performance in comparison to the widely accepted and publicly available tools on held-out mono-allelic data, tumor samples and immunogenic epitopes. Further, the pan-prediction capabilities of SHERPA enable accurate prediction of neoantigens from unseen alleles in patient samples and allow the development of personalized cancer therapy and neoantigen-based biomarkers that are predictive of response to immunotherapy.

## References


1. Rachel Marty Pyke, Dattatreya Mellacheruvu, Steven Dea, et al. 2021, Mol Cell Proteomics 20, 100111, DOI:https://doi.org/10.1016/j.mcpro.2021.100111
2. D. Chowell, S. Krishna, P.D. Becker, C. Cocita, et al. Proc. Natl. Acad. Sci. U. S. A., 112 (2015), pp. E1754-E1762
3. Charles W. Abbott, Sean M. Boyle, Rachel Marty Pyke, et al. Clin Cancer Res 2021;27:4265-4276.

## Get in Touch

To learn more about how we can help accelerate your personalized cancer vaccine or neoantigen-based adoptive cell therapy development program, contact us at **info@personalis.com.**

## Sales Contact

United States
**info@personalis.com**

Europe
**europe@personalis.com**

Other Countries
**info@personalis.com**

## Personalis, Inc.

1330 O'Brien Drive, Menlo Park, CA 94025
+1 855-GENOME4 (436-6634) | +1 650-752-1300
**www.personalis.com**



For Research Use Only. Not for use in diagnostic procedures.

Personalis®

DOC-000572(B)