MCP TECHNOLOGICAL INNOVATION AND RESOURCES Special Issue: Immunopeptidomics

Precision Neoantigen Discovery Using Large-scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation

Authors

Rachel Marty Pyke, Dattatreya Mellacheruvu, Steven Dea, Charles W. Abbott, Simo V. Zhang, Nick A. Phillips, Jason Harris, Gabor Bartha, Sejal Desai, Rena McClory, John West, Michael P. Snyder, Richard Chen, and Sean Michael Boyle

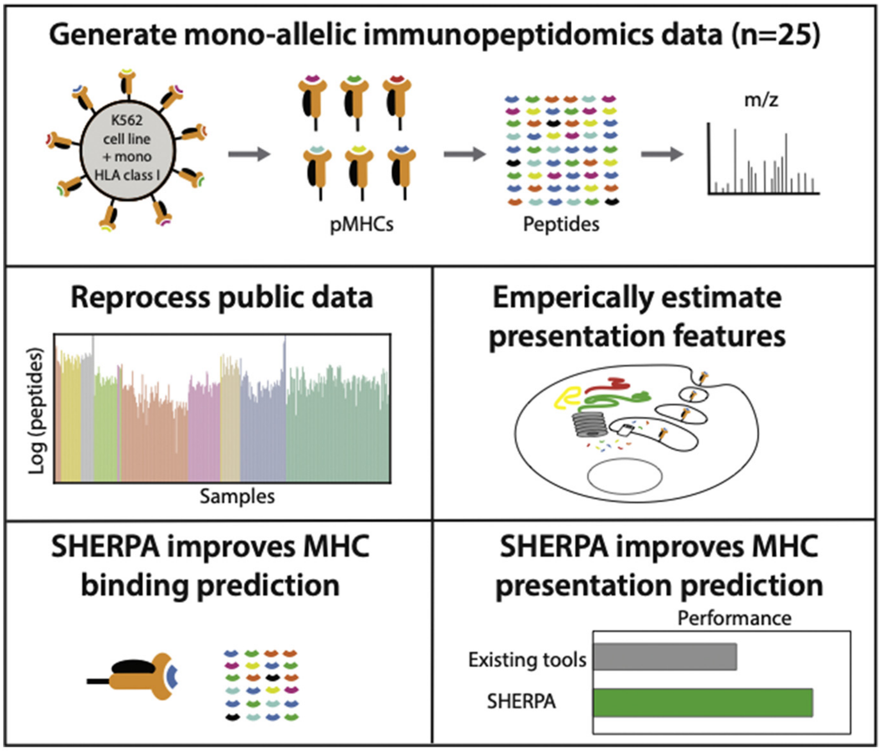
Correspondence

sean.boyle@personalis.com

In Brief

Accurately identifying neoantigens is critical for many clinical applications. We generated immunopeptidomics data from 25 stably transfected monoallelic cell lines. Then, we systematically reprocessed a large corpus of public data to improve major histocompatibility complex (MHC) binding pocket diversity and to empirically learn the rules of antigen presentation. In applying these datasets, we trained SHERPA, an MHC binding and presentation prediction algorithm. SHERPA improves performance compared with existing tools by 1.44-fold in held-out monoallelic data and 1.11-fold for immunogenic epitopes.

Graphical Abstract


Generate mono-allelic immunopeptidomics data (n=25)
K562 cell line + mono HLA class I
pMHCs
Peptides
m/z
Reprocess public data
Log (peptides)
Samples
Emperically estimate presentation features
SHERPA improves MHC binding prediction
SHERPA improves MHC presentation prediction
Performance
Existing tools
SHERPA

Highlights

- Generated 25 stably transfected monoallelic cell lines and applied immunopeptidomics.
- Harmonized 512 public immunopeptidomic samples through systematic reprocessing.
- Developed pan-allele MHC-binding algorithm (SHERPA) utilizing 167 human HLA alleles.
- SHERPA demonstrates up to 1.44-fold increased precision over competing algorithms.

2021, Mol Cell Proteomics *20*, 100111

https://doi.org/10.1016/j.mcpro.2021.100111

Precision Neoantigen Discovery Using Large-scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation

Rachel Marty Pyke[1,‡], Dattatreya Mellacheruvu[1,‡], Steven Dea[1], Charles W. Abbott[1], Simo V. Zhang[1], Nick A. Phillips[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,*§]

Major histocompatibility complex (MHC)-bound peptides that originate from tumor-specific genetic alterations, known as neoantigens, are an important class of anticancer therapeutic targets. Accurately predicting peptide presentation by MHC complexes is a key aspect of discovering therapeutically relevant neoantigens. Technological improvements in mass-spectrometry-based immunopeptidomics and advanced modeling techniques have vastly improved MHC presentation prediction over the past two decades. However, improvement in the sensitivity and specificity of prediction algorithms is needed for clinical applications such as the development of personalized cancer vaccines, the discovery of biomarkers for response to checkpoint blockade, and the quantification of autoimmune risk in gene therapies. Toward this end, we generated allele-specific immunopeptidomics data using 25 monoallelic cell lines and created *Systematic HLA Epitope Ranking Pan Algorithm* (SHERPA), a pan-allelic MHC-peptide algorithm for predicting MHC-peptide binding and presentation. In contrast to previously published large-scale monoallelic data, we used an HLA-null K562 parental cell line and a stable transfection of HLA alleles to better emulate native presentation. Our dataset includes five previously unprofiled alleles that expand MHC-binding pocket diversity in the training data and extend allelic coverage in under profiled populations. To improve generalizability, SHERPA systematically integrates 128 monoallelic and 384 multiallelic samples with publicly available immunoproteomics data and binding assay data. Using this dataset, we developed two features that empirically estimate the propensities of genes and specific regions within gene bodies to engender immunopeptides to represent antigen processing. Using a composite model constructed with gradient boosting decision trees, multiallelic deconvolution, and 2.15 million peptides encompassing 167 alleles, we achieved a 1.44-fold improvement of positive predictive value compared with existing tools when evaluated on independent monoallelic datasets and a 1.15-fold improvement when evaluating on tumor samples. With a high degree of accuracy, SHERPA has the potential to enable precision neoantigen discovery for future clinical applications.

The concerted efforts of innate and adaptive immunity help maintain homeostasis and fight pathogen attacks. The innate immune system reacts quickly and is largely nonspecific while the adaptive immune system is highly specific, typically takes a longer time to develop, and is long-lasting. Within the adaptive immune system, T cells survey the health of a cell by examining major histocompatibility complex (MHC) protein complexes on the cell surface. All nucleated cells express MHC on their surface, and those cells presenting non-self and aberrant peptides are identified and eliminated. Although several factors are needed to mount an immunogenic response from CD8+ T cells, MHC Class I presentation of relevant peptides is a gatekeeping step (1–3).

Early *in vitro* and *in vivo* experiments evaluating binding characteristics of various peptide MHC pairs indicated that presented peptides had allele-specific motifs. *In vitro* experiments primarily attempted to measure binding affinity of specific peptides and their cognate MHC complexes using competitive binding assays in a hypothesis-driven manner (4). Advances in liquid chromatography and mass spectrometry (LC-MS/MS) heralded a new era of large-scale immunopeptidomics and the ability to learn

From the [1]Personalis, Inc, Menlo Park, California, USA; [2]Department of Genetics, Stanford University, Palo Alto, California, USA
‡These authors contributed equally to this work.
§These authors jointly supervised this work.
*For correspondence: Sean Michael Boyle, sean.boyle@personalis.com.

https://doi.org/10.1016/j.mcpro.2021.100111

the rules of binding and presentation in an unsupervised way (5). However, the assignment of peptides to their cognate alleles was problematic (6). Further advances in using genetically engineered cell lines expressing a single allele of interest, often at high copy number, helped circumvent this problem and produce unambiguous HLA peptidomes (7).

Algorithms to model MHC binding have progressed along with these technological advances. Early approaches relied exclusively on MHC-peptide binding affinity data (8). As immunopeptidomics data have become increasingly available, more groups have incorporated these data into prediction models. Several recent approaches employ multiallelic immunopeptidomics data, including clustering-based deconvolution (9, 10), iterative assignment (11), and direct modeling (12). Other efforts capitalize on the unambiguous nature of monoallelic data (7, 13, 14). Furthermore, some algorithms aim to model MHC-peptide binding alone (10, 15, 16), whereas others extend to modeling antigen processing and surface presentation as well (7, 12, 13, 17). Although there is disagreement on the best way to model MHC presentation, there is clear consensus that advances in mass spectrometry have enabled scalable data generation and significantly improved MHC-peptide binding and presentation prediction algorithms.

With the goal of creating an accurate MHC-peptide binding and prediction algorithm, we generated a high-quality dataset comprising 25 monoallelic cell lines. All of the monoallelic cell lines are distinct from previously published datasets due to their stable transfection and background cell line (K562). We also profiled five of the alleles that have not previously been profiled using monoallelic immunopeptidomics technologies, expanding the known MHC-binding pocket diversity and the allelic representation in underprofiled Asian, African, and Middle Eastern populations; the lack of information for underrepresented groups is an area of high current interest (18). Further, we combined our dataset with curated publicly available binding affinity data and monoallelic immunopeptidomics data to create a robust MHC-binding prediction algorithm. In order to capture the diverse facets of antigen processing and presentation, we systematically reprocessed and deconvoluted publicly available multiallelic data from several tissue types with different transcript expression profiles using our binding model trained on monoallelic data and expanded our training data to encompass 167 alleles. We also employed this large dataset to empirically learn patterns in antigen processing at the protein and peptide level. We applied gradient boosted decision trees to train our prediction models, which we call SHERPA (*Systematic HLA Epitope Ranking Pan Algorithm*). We evaluated SHERPA on held-out monoallelic and independently generated multiallelic tumor datasets to demonstrate that our integrated approach for MHC-peptide presentation prediction has a 1.44 and 1.15 fold improvement in performance, respectively, over the best existing tools.

EXPERIMENTAL PROCEDURES

Immunopeptidomics Using Monoallelic Cell Lines

Experimental Design and Statistical Rationale—In this study, all peptides studied were derived from the MHC-I immunopeptidome. Experimental work was performed on two sample types: monoallelic cell lines and tumor tissue. For the monoallelic cell lines, 28 cell lines of 5 × 10^9 cells were processed. Three biological replicates were assessed for a single cell line. One cell line was an HLA-null line used as a negative control. Peptides from eight of the cell lines were processed a single time. Peptides from the other 20 of the cell lines were divided and processed on the mass spectrometer twice (11 with CID only, nine with both CID and EThcD). For the tumor tissues, 12 tissues were processed without controls or replicates. In addition to the experimental work, publicly available peptides were also analyzed. A Wald test and two-sided T tests were used in this study.

Cell Culture—We generated monoallelic cell lines by stably transfecting K562 parental cells with a single allele of interest using Jump-In technology (Thermo Fisher Scientific). Optimized plasmids comprising the sequences for HLA, beta 2 microglobulin (B2M), and IRES promoter were synthesized for each allele (GeneArt). Cells were screened for plasmid integration and expanded to 500 M cells through various passages. The cells were pelleted once surface expression of target alleles was confirmed using flow cytometry (using W6/32 antibody). Transfection experiments were performed by Thermo Fisher Scientific.

Immunoprecipitation of Peptide-MHC Complexes—Pelleted cells were resuspended in octylthioglucoside lysis buffer, and the cell lysate was incubated overnight with W6/32 antibody immobilized on Protein A sepharose. After washing the resin, MHC-bound peptides were eluted using 0.1 M acetic acid, 0.1% TFA. The success of immunoprecipitation was verified by confirming the depletion of MHC Class I complex in post-IP samples using ELISA (enzyme-linked immunosorbent assay). Immunoprecipitations were performed by Cayman Chemical.

Peptide Sequencing Using LC-MS/MS

Eluted peptides were desalted using solid-phase extraction (SPE; Empore C18), loaded on a column, eluted using 80/20 acetonitrile/water (0.1% TFA), lyophilized, and stored. Samples were reconstituted in 0.1% TFA before they were analyzed using liquid chromatography–mass spectrometry (LC/MS/MS). Chromatographic separation was performed using a 2 h gradient on a Waters NanoAcquity system. Peptides were analyzed using a Thermo Fisher Fusion Lumos mass spectrometer in data-dependent mode (MS1: Orbitrap at 60,000 FWHM resolution, m/z range: 300–800; isolation window: 1.6 Da; fragmentation: EThcD and CID; MS2: Orbitrap at 15,000 FWHM; cycle time: 3 s). MS experiments were performed by MS Bioworks, LLC.

Peptide Identification—Peptides were identified using PEAKS software (PEAKS Studio 10.0 build 20190129) (19) using the default two-step identification workflow, where the first step performs *de novo* sequencing to identify mass tags, and the second step performs a database search on a subset of putative proteins identified using *de novo* mass tags. The workflow was run with the following settings. Protein database: Swissprot proteome database (20,402 entries; dated 03-20-2019); precursor mass tolerance: 10 ppm; fragment mass tolerance: 0.02 Da; enzyme specificity: none; fixed modifications: carbamidomethylation of cysteine (+57.0215); variable modifications: oxidation of methionine (+15.9949), N-terminal acetylation (+42.0106). Peptide-to-spectrum matches (PSMs) were filtered at 1% FDR, estimated using decoy sequences.

Postprocessing and Quality Control—Peptides identified at 1% false discovery rate (FDR) were filtered further to remove spurious peptides using the following constraints: spurious peptides were defined as polymeric peptides, peptides from highly chimeric spectra ($n > 2$), and identifications with less than ten fragment ions (Supplementary Code). Background contaminants, profiled using a mock transfection (GFP), were filtered out. Samples were manually inspected for the presence of motif signatures (using Gibbs Clustering software (9)) and peptide yields (heuristic cutoff of 500 peptides per sample).

Extraction and Processing of Public Datasets

Extraction and Processing of Publicly Available Mono- and Multi-allelic Data—Publicly available immunopeptidomics data were identified after an exhaustive literature search. Raw data (.raw files) were downloaded and systematically processed similar to in-house data. HLA types to sample mappings were extracted from corresponding publications. Peptide identifications from samples that passed rigorous quality control criteria (as described above) were aggregated.

Processing of in Vitro *Binding Affinity Data*—Raw HLA-peptidome data were downloaded from the Immune Epitope Database (IEDB) (date: 03-09-2020) and filtered to identify entries corresponding to *in vitro* binding assays. The "Object Type" column was filtered to exclude any nonlinear peptides, and the "Units" column was filtered to exclude any non-nM entries. Further, only four digit MHC Class I peptides of length 8 to 11 were retained. The peptides were derived from the following "Method/Technique" categories: "purified MHC/competitive/radioactivity," "purifiedMHC/direct/fluorescence," "purified MHC/competitive/fluorescence," "cellularMHC/competitive/fluorescence," "cellular MHC/direct/fluorescence," "cellular MHC/competitive/radioactivity," "binding assay," and "lysate MHC/direct/radioactivity." Ligands with IC50 values less than 500 nm were identified as binders.

Measurement and Analysis of Transcript Expression

Sequencing of Monoallelic Cell Lines—Representative samples of monoallelic K562 cell lines generated in-house were sequenced in triplicate using our in-house commercial platform ImmunoID NeXT with 200 million paired end reads (150 base pair) of sequencing for RNA from tumor samples. Reads were aligned in accordance with Personalis Cancer RNA pipeline and transcript per million (TPM) values were extracted. Another cell line, B721.221, with large amounts of publicly available monoallelic data, was sourced from the American Type Culture Collection (ATCC) and profiled in triplicate using ImmunoID NeXT similarly to minimize technological variance.

Generation of Transcriptome for External Immunopepdomics Datasets—For one of the largest publicly available monoallelic datasets generated using B721.221 cell (MSV000080527; MSV000084172) lines, we regenerated the transcriptome data on ImmunoID NeXT (7, 13), as described above. For the rest of the samples in the expanded dataset, we imputed TPM values with our internal database by taking the median values from all samples with a matching tissue type.

Evaluation of Differential Transcriptome Abundances Across Cell Lines and Tissue Types—To evaluate the differences in expression between K562 and B721.221, we performed differential gene expression analysis using the Deseq2 package (20), which takes in raw transcript counts from all replicates of a sample set to determine differentially expressed genes. We used a threshold of an absolute value $\log_2$ fold change of greater than or equal to 2 in conjunction with an adjusted p-value threshold of less than or equal to 0.01 to identify differential gene expression. We then visualized the differentially expressed genes in a volcano plot to show cell line gene expression differences, using the Bioinfokit tool by Renesh Bedre (https://github.com/reneshbedre/bioinfokit/tree/v0.9). We then took the identified genes that were enriched in K562 and the genes that were depleted in K562 as compared with B721.221 and calculated the Gene Ontology (GO) enrichment terms for both gene sets using http://geneontology.org/. To evaluate the differences in expression between the tissue types, we restricted our analysis to the tissues with unique expression profiles ($n = 30$). Then, we identified the 5000 genes with the highest standard deviation between the expression profiles and visualized those genes as rows and the unique expression profiles as columns in a clustered heatmap.

Exploratory Analysis of Immunopeptidomics Data

Relationship Between Expression and Peptide Presentation—Peptides identified in K562 monoallelic samples were assigned a transcript expression value (TPM) by mapping each peptide to one or more proteins, and proteins to one or more transcripts. The highest TPM was assigned as the representative value when there are multiple mappings. A set of random peptides ($n > 50$ million) were also generated and assigned TPMs as described. TPMs were binned to deciles, and enrichment of presented peptides in each bin was calculated as the log ratio of number of presented peptides to background random peptides.

Relationship Between Cleavage Specificities and Peptide Presentation—To understand the cleavage preferences of the proteasome in our monoallelic cell lines, we mapped each peptide to a protein (as described above) and identified the right and left flanking sequences (five amino acids) of the peptides. A set of random peptides ($n > 50$ million) were also generated and assigned flanking sequences as described to serve as a null dataset. To calculate the enrichment or depletion of amino acids at each position in the flanking regions, we subtracted the observed amino acid frequency from the random amino acid frequency and divided by the random amino acid frequency. Of note, the C- and N-termini of the protein were also considered in addition to the individual amino acids. We visualized the enrichment and depletion as a heatmap.

Clustering Alleles by Binding Pocket Similarity—The binding pocket was represented by a pseudo-sequence of 34 amino acids as described previously (8). The distance between alleles was calculated by taking the sum of the BLOSUM62 scores for the amino acids in each allele at each position across the pseudo-sequence. All of the unique alleles with in-house or public monoallelic immunopeptidomics data were compared with one another and visualized as a heatmap.

Estimating Allelic Coverage in Underrepresented Ethnic Populations—Allele frequencies from several ethnic populations were downloaded from http://www.allelefrequencies.net/ on 05-22-2020. The frequencies of the five novel alleles in underrepresented ethnic populations were taken directly from the data on the website. To estimate the allelic coverage of all of the unique alleles in the whole data across ethnic populations, we focused on the data derived from the National Marrow Donor Program (NMDP, $n = 18$ populations). For each population, we performed a Monte Carlo simulation to generate 10,000 synthetic individuals with the allelic frequencies for that population specified by NMDP. Then, we calculated the percentage of the alleles in the synthetic cohort that were represented in the expanded dataset (in-house monoallelic, public monoallelic, IEDB, and public multiallelic).

Generation of Peptide Motifs—We generated the motifs for each allele using the weblogo software (version: 3.7.1). All motifs were based on a specific length of peptide (8-, 9-, 10-, or 11-mers). Motifs for public datasets were based on the peptides derived from our internal processing pipeline.

Generation of Position-specific Frequencies of Amino Acids of the Binding Pocket—To assess the representativeness of our expanded dataset to the full space of known alleles, we calculated the frequency of amino acid at each position of the binding pocket pseudo-sequence for both all of the alleles in the IMGT (the international

ImMunoGeneTics information system; http://www.imgt.org/) and our expanded dataset. We visualized these frequencies as stacked bar plots.

Feature Engineering and Building Prediction Models

Data Splitting for Training and Evaluation—In order to rigorously evaluate our performance on novel peptides, we ensured that no peptides with overlapping 8-mer cores were present across the training, validation, and testing datasets. We took the unique set of peptides across our monoallelic immunopeptidomics, multiallelic immunopeptidomics, and IEDB datasets. Then, we grouped peptides that contained any identical 8-mer substrings ("nested" peptides) and placed each peptide group in one of ten different subsets. Once we had roughly equal peptide numbers in each subset, we assigned one subset to be for validation (~10%), one subset to be for testing (~10%), and the final eight subsets to be for training (~80%). Moreover, we fully held out 32 multiallelic samples from the training dataset (described in detail in the "Benchmarking and evaluation of prediction models" section) and evaluated the models on a subset of the peptides (~10%) that were fully held out from training (as described above). All publicly available multiallelic data (except for the 32 held-out samples) was used for feature engineering of the gene propensity and hotspot scores because systematically holding out proteins or peptides would skew the scores. Only the ~80% training subset of the multiallelic data was used for deconvolution.

Specifications of Prediction Models—We trained two types of prediction algorithms, one that models *binding* and another that models *presentation*. The binding algorithm models MHC peptide interaction affinities and takes as inputs the HLA-binding pocket, the amino acid sequence, and the length of peptide ligands. The presentation algorithm jointly models both antigen processing and peptide-MHC (pMHC) binding, whose inputs include: the HLA-binding pocket, amino acid sequence and length of peptide ligands, proteasomal processing footprints manifested in the left and right flanking regions of peptide ligands, abundance of source proteins engendering peptide ligands as measured by gene expression, and two features that model the propensity of antigen processing and presentation. Features corresponding to these inputs are generated as follows:

1. *HLA-binding pocket (B):* The binding pocket is represented by a pseudo-sequence of amino acids as described previously (8). Briefly, 34 positions on the protein sequence of the HLA that are a distance of 4 Å or lesser in crystallographic structures were selected from the full protein to serve as the pseudo-sequence. The amino acid sequence is encoded using a BLOSUM62 substitution matrix, where each amino acid is represented by a 20-dimensional vector constituting the relative weights of amino acid substitutions. We chose BLOSUM62 encoding as opposed to a one-hot encoding approach under the assumption that substitutions between amino acids with evolutionary similarity have a lower impact on epitope-binding changes than substitutions between amino acids that are very dissimilar.
2. *Peptide ligand (P):* The amino acid sequence of the peptide ligand is encoded using a BLOSUM62 substitution matrix, where each amino acid is represented by a 20-dimensional vector constituting the relative weights of amino acid substitutions. Since HLA ligands could be of variable lengths (8- to 11-mers), we have adopted a middle-padding approach described previously and adjusted all peptides to a length of 11 amino acids by inserting blanks in the middle (12). Briefly, for every peptide with fewer than 11 amino acids, we assign one to three 20-dimensional vector(s) of zeros to the center of the peptide encoding to reach the maximal 11-mer length. We take this approach to have a pan-length algorithm that keeps the peptide anchors in consistent columns of the matrix.
3. *Peptide length (L):* The number of amino acids in the peptide ligand is designated as the peptide length.
4. *Left and right flanking regions (F):* Peptide sequences (5-mers) to the left and right of the peptide ligand in the source protein are used as the left and right flanking regions respectively. Multimappers are resolved by assigning the protein with the highest transcript expression to the peptide. These left and right flanking 5-mers are encoded using BLOSUM62 substitution matrix as described above.
5. *Abundance of source protein (T):* Peptide ligands are redundantly assigned to all source proteins and then to transcripts. The transcript with the highest expression (calculated as the TPM) is chosen. Both the transcript and the TPM are then assigned to the peptide ligand.
6. *Gene propensity score (G): P*ublicly available multiallelic data was used to estimate gene propensity. Peptide ligands from each sample are mapped to source proteins and then to transcripts redundantly. The number of peptides mapping to each transcript-associated protein were determined. To calculate the expected number of peptides for each transcript-associated protein, the number of transcripts per million for each protein across all multiallelic data sources are determined and normalized by the protein length and the number of peptides derived from each sample. Then, the number of observed and expected peptides for each protein are summed across all of the multiallelic datasets. Finally, the observed values were divided by the expected values to derive a gene propensity score for each protein (gene). Proteins without any observed transcripts across all samples are (potentially pseudo genes) given a score of −3 to deprioritize them.
7. *Hotspot score (H):* Publicly available multiallelic data were used to estimate the hotspot score. Peptide ligands from each sample are mapped to source proteins. When peptides were able to map to multiple proteins, all potential mappings are used. The number of peptides overlapping a particular amino acid represents the hotspot score of that amino acid. To assign a hotspot score for a peptide, the peptide is mapped to its source protein and the amino acid hotspot score spanning the peptide is averaged.

We evaluated the contribution of each feature to the XGBoost model using the "gain" metric, which measures the relative contribution of a feature to the main model by aggregating the individual contributions of the feature in each tree. Higher values indicate greater importance.

Generating Negative Examples (Nonbinders)—Immunopeptidomics experiments only generate peptides that successfully bind to and are presented by MHC molecules. Thus, we synthetically generated negative examples instead of experimentally identifying them. We generated 20 negative examples for every positive example in our training and validation datasets. To generate a negative example, we randomly selected a protein from the Swissprot proteome (downloaded on 03-20-2019) and then randomly selected a peptide from within that protein. Peptides were selected to have length 8, 9, 10, or 11 with equal probability. Flanking regions were assigned based on the true flanking regions around the selected peptides. A gene expression value (TPM) was assigned by randomly selecting a transcript from the transcriptome of the associated positive example. The gene propensity score and hotspot score were assigned based on the protein and position in the protein of the peptide selected for the negative example.

Training the Prediction Algorithm—All models were trained using a gradient boosted decision tree algorithm implemented using an open-source package XGBoost (21). All numeric and encoded features were provided as a vectorized input feature vector for training the algorithm.

Optimal training parameters were selected based on a subset of samples using sequential model-based optimization, implemented using the open-source package HyperOpt (22). The resulting training parameters used for the final training were as follows: loss function—binary logistic; max depth—10; eta—0.01; subsample—0.7; early stopping rounds—5; min child weight—0.5; max delta step—1; tree method—hist; number of estimators—500. Large training sets were subset and processed in parallel based on available compute resources and processed on a high-performance cluster.

Calibrating Raw Scores Using Percent Rank Values—A set of 500,000 peptides were randomly selected from the human proteome. Once we trained a model, we calculated the predictions across the set of random peptides with every allele. Then, we ranked the random peptides according to their raw prediction probability (output of XGBoost) for each allele. For each new peptide predicted, the assigned rank is the percentage of the random set that is predicted to bind or to be presented with a better raw score than the new peptide. The ranks range from 0 to 100 with lower scores meaning better bound or presented peptides. The ranks are recalculated for each model and allele combination.

Applying Prediction Models to Multiallelic Cell Line and Patient Samples—The MONO-binding model was used to generate binding ranks for the peptides in the public multiallelic samples (including tumors, tissues, and cell lines) with immunopeptidomics data. Only peptides with 8, 9, 10, and 11 amino acids were considered. In total, 10% of multiallelic data was held out for testing purposes. Predictions were made for all HLA alleles (up to six) of each sample. Samples without HLA typing were excluded from this analysis.

Deconvolution of Multiallelic Data—Once predictions were made for all HLA alleles of each multiallelic sample with immunopeptidomics data, the data was deconvoluted to decide which allele–peptide pairs should be included in the training dataset for the final model. First, we excluded all allele–peptide pairs with a predicted binding rank of ≥ 0.5 to exclude all peptides that do not bind to any of the designated alleles. Second, if there were multiple alleles predicted to bind to a specific peptide, we selected the allele–peptide pair with the lowest rank (best binder) and excluded all other pairs. Then, we removed any duplicate allele–peptide pairs. Finally, we generated 20 negative examples (as described above) for every new positive example derived from the multiallelic data.

Training Composite Models—We trained a total of three prediction models for our composite model. In addition, we also trained five more models to help us better understand features that contribute to optimal performance. The prediction models in our composite model were trained as follows:

1. *MONO-Binding*: Trained using the IEDB, in-house monoallelic immunopeptidomics and public monoallelic immunopeptidomics data with B, P, and L as features.
2. *SHERPA-Binding*: Trained using the IEDB, in-house monoallelic immunopeptidomics, public monoallelic immunopeptidomics, and deconvoluted multiallelic immunopeptidomics data with B, P, and L as features.
3. *SHERPA-Presentation*: Trained using the in-house and public monoallelic immunopeptidomics data with SHERPA-binding, F, T, G, and Has features.

The additional prediction models were trained as follows:

1. *PUBLIC-Binding*: Trained using the public monoallelic immunopeptidomics data with B, P, and L as features.
2. *MONO-Binding-LOO*: 126 allele-specific models trained using the IEDB, in-house monoallelic immunopeptidomics, and public monoallelic immunopeptidomics data with B, P, L, IEDB-binding, INHOUSE-binding, and PUBLIC-binding as features. Each allele-specific model was trained without peptides in the training dataset from the respective alleles.
3. *SHERPA-Binding+F*: Trained using the IEDB, in-house monoallelic immunopeptidomics, public monoallelic immunopeptidomics, and deconvoluted multiallelic immunopeptidomics data with SHERPA-binding and F as features.
4. *SHERPA-Binding+FT*: Trained using the IEDB, in-house monoallelic immunopeptidomics, public monoallelic immunopeptidomics, and deconvoluted multiallelic immunopeptidomics data with SHERPA-binding, F and T as features.
5. *SHERPA-Binding+FTG*: Trained using the IEDB, in-house monoallelic immunopeptidomics, public monoallelic immunopeptidomics, and deconvoluted multiallelic immunopeptidomics data with SHERPA-binding, F, T, and G as features.

Benchmarking and Evaluation of Prediction Models

Generation of the Monoallelic Held-out Test Data—For the monoallelic immunopeptidomics datasets (in-house and public), ~10% of the positive examples were held out of the training and validation datasets. Each positive example was supplemented with 999 negative examples to reflect the positive-to-negative ratio accepted by the field (5, 23–26). For the IEDB data, ~10% of the positive and negative data was withheld from the training dataset. No supplementary negative examples were added, so the ratios in the test dataset reflect the positive-to-negative peptide ratio in IEDB. All monoallelic validation figures rely on the monoallelic immunopeptidomics test data except for one comparison which uses the IEDB test data.

Calculation of Evaluation Metrics—Three metrics were used to evaluate the performance of the prediction models. The metrics are as follows:

1. *Positive predictive value (PPV)*: PPVs were calculated by making predictions on the entire test dataset and calculating the percentage of peptides in the top X% of predictions that are positive examples, with X representing the portion of the dataset that are positive examples. Each PPV was calculated individually for the peptides of each allele and combined using a median giving a single metric. Of note, the positive-to-negative ratios of the monoallelic immunopeptidomics and IEDB test datasets vary, so the interpretations of the plots are different.
2. *Precision–recall curves*: Precision–recall curves were generated by calculating the precision and the recall for every possible cutoff and plotting them as a single line.
3. *Fraction of observed peptides predicted by model*: This metric was used for all multiallelic tumor validation analyses. First, a single score is selected to represent each peptide observed with immunopeptidomics by making predictions on all of the patient's HLA alleles and selecting the best (lowest) rank among the predictions. Then, the score is calculated by determining the percentage of observed peptides that are given a rank of ≤ 0.1.

Leave-one-out Pan-allelic Analysis—To evaluate the pan-allelic performance of the MONO-binding model used for model-based deconvolution, we trained 126 independent models with the same features as the MONO-binding model. For each model, we excluded the peptides from a specific allele from the set of peptides used to train the MONO-binding model. To generate the predicted motif for each allele, we predicted the binding rank for 500,000 random peptides for a given allele with the model for which the allele had been excluded from training. Then, we generated the motif for peptides with the top percentile of binding ranks. Motifs were only visualized for alleles with at least 50 positive peptides in the training data. To generate the precision–recall curve, we used all 126 models to predict the binding ranks of the monoallelic immunopeptidomics data (in-house and public) that was excluded from training and validation (10% test dataset). The predictions for each allele were made with the model that excluded that allele from training.

Generating Validation Data Using Tissue Samples—A total of 12 fresh-frozen tumor samples (five colorectal and seven lung) and matched adjacent normal fresh-frozen samples were selected for patient validation. These samples were purchased from a biobank and were collected under IRB-approved protocols and abide by the Declaration of Helsinki principles. Each tumor sample was divided into two pieces. Immunoprecipitation of MHC complexes followed by LC-MS-MS (as described above) was performed in a portion of each tumor sample to yield immunopeptidomics data. DNA and RNA were extracted from the remaining tumor sample, and DNA was extracted from the adjacent normal sample for analysis with ImmunoID NeXT (as described above). The RNA extracted from the tumor sample was used to yield transcriptomic data.

Extraction of Held-out Validation Datasets from External Multiallelic Samples—In total, ~10% of multiallelic immunopeptidomics data with nonoverlapping "nested" peptides with the training and validation datasets was withheld from deconvolution and training to serve as a test dataset. We used ten samples from each of two specific datasets to validate our internal tumor immunopeptidomics performance (PXD007635, PXD009602) (27, 28). Only samples with HLA-A, -B, and -C typing at four-digit resolution were used for the analysis.

Immunogenicity Evaluation—In order for a peptide to incite an immunogenic response, it must be presented on the cellular surface by an MHC allele. Thus, we evaluated the ability of the algorithms to positively identify immunogenic peptides. We used the dataset described in Chowell *et al.* (29) for this analysis and focused exclusively on the immunogenic peptides. Then, we evaluated the percentage of immunogenic peptides that were predicted by the various models at ≤0.1 percentile rank.

Running Comparison Prediction Algorithms—NetMHCpan-4.1-BA, NetMHCpan-4.1-EL, and MHCFlurry-2.0-BA were used as comparison prediction algorithms. All algorithms were run according to default settings. Percentile rank outputs were used for each analysis.

RESULTS

Generation of Monoallelic Immunopeptidomics Data

In order to generate high-quality data, we used an HLA-null K562 parental cell line and engineered monoallelic cell lines that expressed a single HLA allele of interest (Fig. 1*A*). Expanded cells with good surface expression of target alleles were used for immunoprecipitation of peptides associated with MHC complexes, followed by elution of ligands and peptide sequencing using LC-MS/MS. Our stable transfection protocol, which ensures single-site integration of the target allele, minimizes expression biases and enables modeling of antigen processing in closer-to-native condition. Moreover, our K562 parental cell line provides a different background than the majority of previously published monoallelic data (B721.221 parental cell line), providing a novel and standardized system to study peptide presentation.

We generated data for 25 alleles (supplemental Table S1) and identified an average of 1575 unique peptides per allele (range: 905–2712) (Fig. 1*B*). We first evaluated the allele-wise length distribution and recapitulated previously described observations, with HLA-A alleles presenting longer peptides than HLA-B alleles (two-sided *t* test, $p = 0.0066$, supplemental Fig. S1) (13). We also compared the motifs in our raw data with those published previously. Whereas most of the motifs are concordant, we also observed some differences (supplemental Fig. S2). As an example, a previously published dataset (7, 13) for HLA-B*35:01 shows a tryptic digest like lysine/arginine signature for 8, 10, and 11-mers on the C-terminal end of the peptide, whereas our data shows a motif that is consistent with the ones observed with 9-mers (Fig. 1*C*). We observe similar patterns in HLA-A*29:02, HLA-A*30:02, and HLA-B*57:01 (supplemental Fig. S3), suggesting that there may be contamination in the publicly available datasets. Since MHC-binding algorithms learn allele-specific motifs, such subtle differences are critical to create accurate prediction models. We then evaluated the relationship between transcript abundance and peptide presentation (Fig. 1*D*). We observed a positive correlation between the propensity of a transcript to engender presented peptides and its expression level, represented by the TPM. We also evaluated the proteasomal cleavage signatures of presented peptides by examining the left- and right-flanking amino acids (Fig. 1*E*). The relative enrichment of amino acids compared with a random background set indicates an overrepresentation of lysine and arginine on the C-terminal end of the peptide. This observation is in concordance with previously described tryptic and chemotryptic enzymatic activity of proteasome cleavage (7, 13). Of note, we observe an enrichment of peptides originating from either the C-terminal or N-terminal of the source protein, indicating preferential proteasomal processing of peptides requiring relatively lesser enzymatic cleavage (two-sided *t* test, $p = 3.77\text{e-}32$, supplemental Fig. S4). Enrichment was not observed in the N-terminal end of the peptide in other monoallelic datasets (7, 13), possibly due to overexpression of MHC with transient transfection protocols leading to a lack of discrimination among presented peptides.

Novel Monoallelic Cell Lines Enhance Diversity of Training Data and Reveal Subtle Binding Preferences

Beyond the added value of our unique parental cell line and a stable transfection approach, our dataset includes five alleles that had not previously been profiled using monoallelic cell lines. We selected these five alleles on the basis of two criteria: (1) the uniqueness of the allele-binding pocket compared with previously profiled alleles (Fig. 2*A*) and (2) the population frequency of the allele in ethnic populations that are underrepresented in current datasets (Fig. 2*B*). For example, we clustered alleles by their binding pocket similarity and selected HLA-A02:52 from the HLA-A02 subcluster due to its high frequency in Iranian Kurdish populations (7% frequency, Fig. 2*B*). Though several HLA-A02 alleles have been profiled by previous groups (light gray), we found small differences in the motif from closely clustered alleles despite very similar binding pockets. For example, unlike surrounding HLA02 alleles, we observed the presence of tyrosine in the second anchor position (C-terminal end of peptide) for HLA-A02:52. Moreover, although HLA-A02:52 clusters tightly with

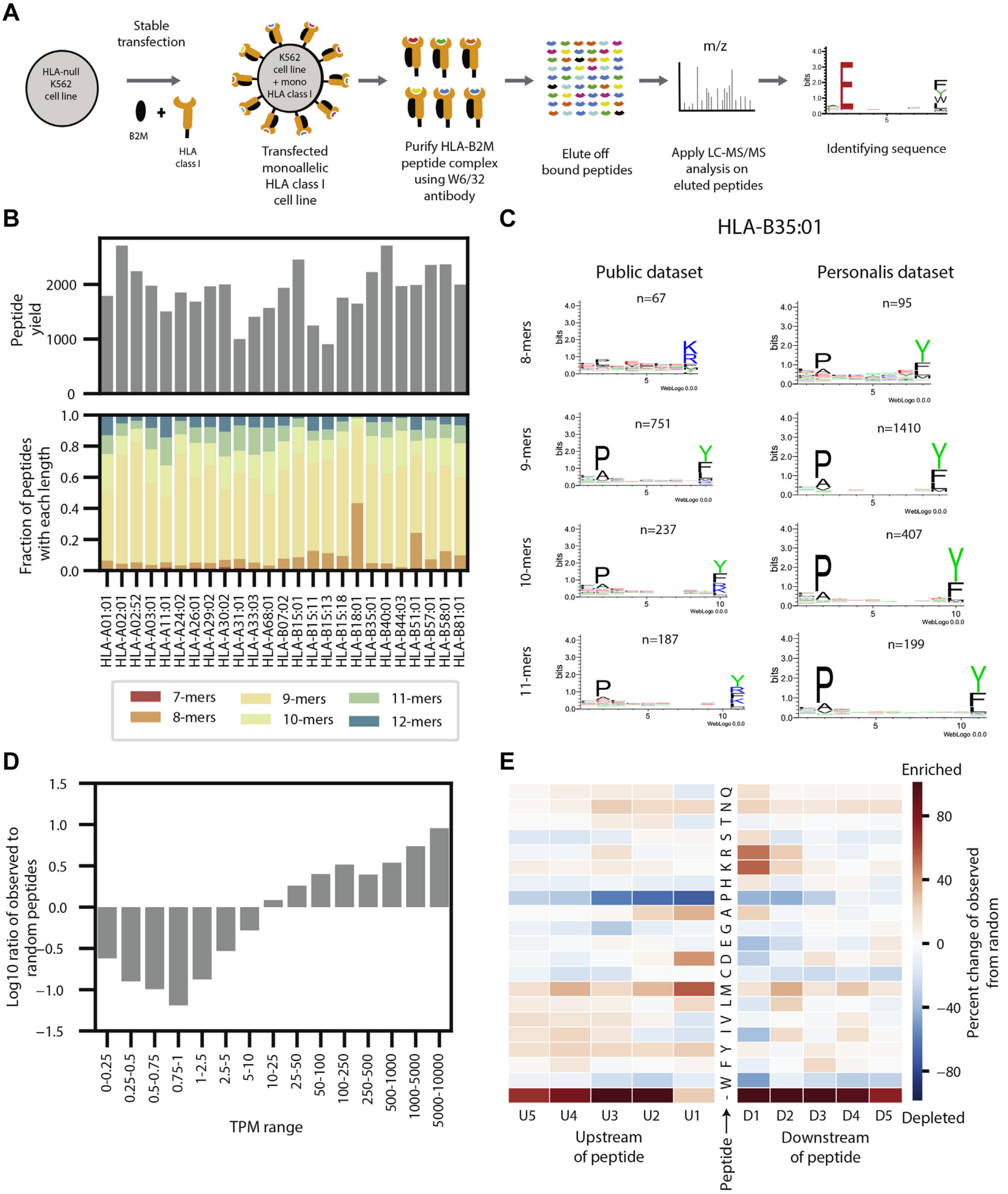


FIG. 1. **Generation and overview of the monoallelic data.** *A*, a schematic of the experimental procedure to generate the monoallelic training data. An HLA allele and B2M were stably transfected into an HLA-null K562 parental cell line. The MHC-peptide complex was purified using a w6/32 antibody and the peptides were gently eluted off the complexes. The peptides were sequenced with LC-MS/MS and identified with a database search. *B*, bar plots showing the peptide yields and distribution of peptide lengths for each of the 25 monoallelic cell lines. *C*, a

HLA-A02:07, HLA-A02:07 has a strong preference for aspartic acid at position 3, whereas HLA-A02:52 has no preference. Similarly, we profiled three HLA-B15 alleles (HLA-B15:11, HLA-B15:13, and HLA-B15:18) that are frequent across several Asian populations, including the second most frequent allele in a Beijing population at over 12% (HLA-B15:11). Despite inclusion in the same allele subgroup and having binding pockets that cluster together, the three alleles have very distinct associated peptide motifs with each having different preferences at each anchor position. These three alleles highlight the value of our novel alleles in the creation of a pan-allelic model, which relies on learning how amino acid changes impact binding preference. Another novel allele, HLA-B81:01, falls outside of a major binding pocket cluster, but its peptide motif shows strong similarities to surrounding alleles (Fig. 2*A*). Of note, HLA-B81:01 has moderate population frequency across over a dozen African populations (2–6% frequencies). Taken together, our novel alleles enhance both binding pocket diversity and population representation in a manner that is not represented in previously published training data.

Large-scale Data Integration Enhances the Representativeness of HLA Peptidome

Given the large amount of publicly available immunopeptidomics and binding affinity data from diverse tissue types and cell lines, we reasoned that a systematic integration of all available high-quality datasets would improve the representation of training data and enable feature engineering. Toward this end, we downloaded a large corpus of publicly available raw data (supplemental Table S2), comprising 2.15 million peptides from 512 experiments and covering 167 distinct alleles. A majority of the datasets are multiallelic (384 experiments from 15 projects) (12, 23, 27, 28, 30–39), but there are also a large number of monoallelic experiments (128 experiments from six projects) (7, 13, 40–44) and alleles with binding assay data (n = 90 alleles) (26) (Fig. 3*A*). As anticipated, the tissue types in the multiallelic datasets have diverse expression profiles (Fig. 3*B*). Furthermore, we observed significant differences in the gene expression profiles of transcripts from B721.221 cell lines (used in a large-scale monoallelic dataset generated from B721.221 cell line with alleles from HLA-A, -B, and -C) (7, 13) and K562 cell lines (used in our in-house data generated) (Fig. 3*C*). In total, 21% of transcripts display significant differences in expression between B721.221 and K562 cell lines, and these variations impact key pathways that change the functionality of the cells (supplemental Table S3). Given the vast differences in underlying proteomes and antigen processing machinery of various tissue types, incorporation of these diverse data is particularly important for modeling presentation effectively.

We evaluated the representation of the MHC alleles in our expanded dataset on two parameters—population coverage and structural diversity of the binding pocket. We assessed population coverage using a simulation study that utilizes allele frequencies from the US National Marrow Donor Program as documented in the allele frequency net database (45). The alleles within our extended dataset represent an average coverage of 96% of observed alleles across various ethnic populations in the United States (Fig. 3*D*). Our training data spans a wide range of allele-binding pocket clusters that have vastly different motif characteristics, indicating a robust representation at the macro level. Since it is possible that relatively small differences in binding pocket sequences might lead to substantial differences in the presented motif, possibly due to the varied importance of amino acids in the binding pocket (46), we evaluated the representation at each position in the binding pocket. We observed a high degree of concordance in the distribution of frequencies between all known alleles and alleles in our expanded training dataset (Fig. 3*E*). Of note, this concordance remains high even when we exclude alleles only present in multiallelic samples (supplemental Fig. S5, *A* and *B*). By expanding our in-house dataset using publicly available data, we not only enhance the breadth of our training data, in terms of both tissue of origin and allelic diversity, but also minimize potential biases introduced by data acquisition techniques, such as chromatography and peptide fragmentation approaches. Additionally, the expanded scale and scope help minimize out-of-distribution modeling errors.

Modeling Peptide-MHC Presentation Using Novel Features Encapsulating Antigen Processing

The unbiased and *ex vivo* nature of mass-spectrometry-based immunopeptidomics data uniquely lends itself to model both MHC-peptide binding and presentation. Modeling *binding* entails learning the motif preferences of HLA alleles and their cognate peptides, whereas modeling *presentation* includes learning propensities of upstream antigen processing in addition to binding motifs (Fig. 4*A*). Several features that capture various aspects of antigen processing and the biophysical characteristics of MHC-peptide binding have been

comparison between motifs of peptides generated from our monoallelic cell line with HLA-B35:01 and a publicly available dataset for the same allele. Motifs are shown for peptides of length 8, 9, 10, and 11. See supplemental Fig. S1 for motifs from all 25 cells. See supplemental Fig. S2 for comparisons with other public datasets. *D*, a bar plot showing the distribution of the ratio of observed peptides from the monoallelic cell lines compared with random expectation across several TPM ranges. Values are shown with a log10 transformation. *E*, a heatmap showing the enrichment and depletion of five amino acids upstream and downstream of the peptides identified from the monoallelic cell lines compared with a random expectation. *Red* denotes the enrichment of amino acids and *blue* denotes the depletion of them. The C- and N-termini of the protein are denoted with “-.”

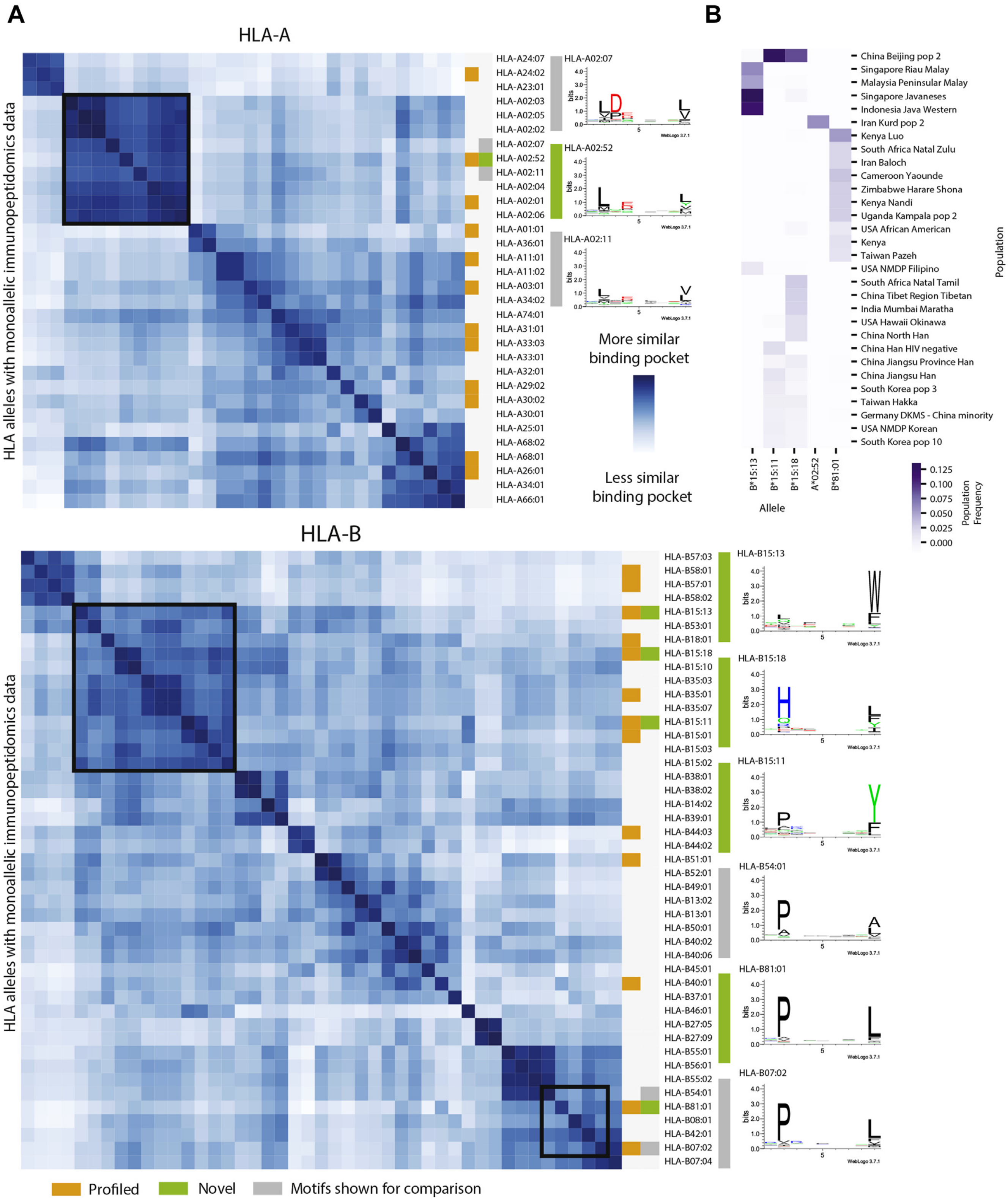


FIG. 2. **Binding pocket diversity and population frequencies of novel alleles.** *A*, heatmaps for HLA-A and -B that represent the binding pocket similarity between alleles with monoallelic immunopeptidomics data. *Dark blue squares* represent alleles that have very similar binding pockets while *white squares* represent alleles with divergent binding pockets. The 25 alleles profiled with our monoallelic system are denoted in *orange*. The five alleles that have not previously been profiled are denoted in *green*. Motifs for these novel alleles are shown alongside motifs for

discussed in the literature, the most prominent of these are peptide sequence, binding pocket sequence, expression of source protein, cleavage patterns, and various propensities for presentation that are not captured by expression (12). We use a combination of widely applied and novel features to model MHC-peptide binding and presentation.

We modeled binding using three features: the amino acid sequence of the peptide ligand (P), the binding pocket pseudo-sequence of the presenting allele (B), and peptide length (L). In addition to these three binding-specific features, we integrated four additional features into our presentation model that encapsulate antigen processing. First, as evidence from our monoallelic immunopeptidomics data, the abundance of the source protein has a strong influence on peptide presentation (Fig. 1*D*). Accordingly, we used a transformed value of transcript expression (T) as a surrogate for protein expression. Second, our data also showed the strong preference of amino acids flanking the presented peptides for proteasomal cleavage; therefore, we incorporated the five amino acids upstream and downstream of each peptide (F) as an additional feature (Fig. 1*E*). It has been discussed previously that certain proteins or regions within proteins have a higher propensity for MHC presentation (23, 33, 47). We have noticed similar trends in the large set of public data that we had downloaded and reprocessed systematically (supplemental Table S2). We evaluated the dependence of the number of peptides presented (adjusted for transcript length) as a function of transcript expression of its cognate gene. As expected, we observed a positive correlation in the median values, but we also observed several outliers that do not follow this trend (data points highlighted in red boxes, Fig. 4*B*). We modeled this skew (propensity) and developed a gene propensity score (G) as our third feature (supplemental Fig. S6*A*, see Experimental Procedures). Similarly, we noticed positional preferences within proteins when we compared observed and expected peptides from various locations within a protein. An anecdotal example of such a skew in Actin Beta gene (ACTB) is shown in Figure 4*C*. Briefly, position-specific coverage of observed peptides was contrasted with expected coverage based on predictions for the 30 most frequent alleles in our dataset. Importantly, we found a lack of correlation between predicted and observed peptide profiles across the entire proteome (supplemental Fig. S6*B*). We modeled this phenomenon using position-specific coverage of immunopeptides in our large dataset and developed a hotspot score (H) as our final feature (see Experimental Procedures). An ensemble of these propensity scores and other features described above facilitate the modeling of antigen processing and surface presentation.

Systematically Incorporating Binding Affinity, Monoallelic and Multiallelic Data into a Composite Model

Both *in vitro* binding assays and monoallelic immunopeptidomics experiments produce allele-specific data, but assigning a peptide to its cognate allele, also known as deconvolution, is a key challenge in modeling multiallelic data. Several approaches have been applied to perform deconvolution (10–12). We utilized our comprehensive binding affinity and monoallelic datasets to develop a model-based deconvolution approach to convert multiallelic data into pseudo monoallelic data (Fig. 5*A*, supplemental Table S4). First, we generated a binding prediction model using exclusively monoallelic data (in-house monoallelic data, public monoallelic data, and *in vitro* binding assay data extracted from IEDB). Unlike the binding affinity data from IEDB that generates both positive and negative examples, immunopeptidomics data only generates positive examples, so we generated negative examples using a large set of random peptides from the human proteome at a 1:20 ratio. A large proportion of negative examples helps minimize the risk of learning random or systematic biases in synthetically generated negative examples. This monoallelic model (MONO-binding) is able to accurately predict motifs for alleles excluded from the training dataset (Fig. 5*B*, supplemental Fig. S7). Given our confidence in the model trained with allele-specific data (MONO-binding), we applied it to each of the multiallelic datasets to assign each peptide in a sample to its cognate allele and generate pseudo monoallelic data. From 289 multiallelic sources representing 118 unique HLA alleles, we began with over 1.15 million peptides and successfully mapped about 700,000 peptides to alleles (~60% mapping rate). Then, the monoallelic and pseudo monoallelic data were merged to train a comprehensive binding model (SHERPA-binding). Finally, SHERPA-binding was used as a primary model (p-model) or feature for our SHERPA-presentation model. This final model incorporates all presentation features but is trained exclusively on monoallelic data to avoid overfitting to the multiallelic datasets that generated the gene propensity and hotspot features. As expected, when we examined the feature importance for each of the models, we observe the peptide anchor residues and diverse binding pocket amino acids have the greatest influence over the SHERPA-binding model, and SHERPA-binding has the greatest influence over the SHERPA-presentation model (supplemental Fig. S8). For all models, prediction probabilities were calibrated into percentile ranks to remove allele-specific biases as described earlier (15).

Benchmarking the Performance of SHERPA Binding and Presentation Models

To evaluate our models, we first tested their performance on ~10% held-out monoallelic data (positive examples), mixed

related alleles in *gray*. *Black boxes* denote the cluster of alleles containing the newly profiled alleles. *B*, a heatmap showing the frequencies of the five novel alleles in several populations of diverse world ethnicities. *Dark purple* denotes high population frequencies of the alleles and *light purple* denotes low population frequencies.

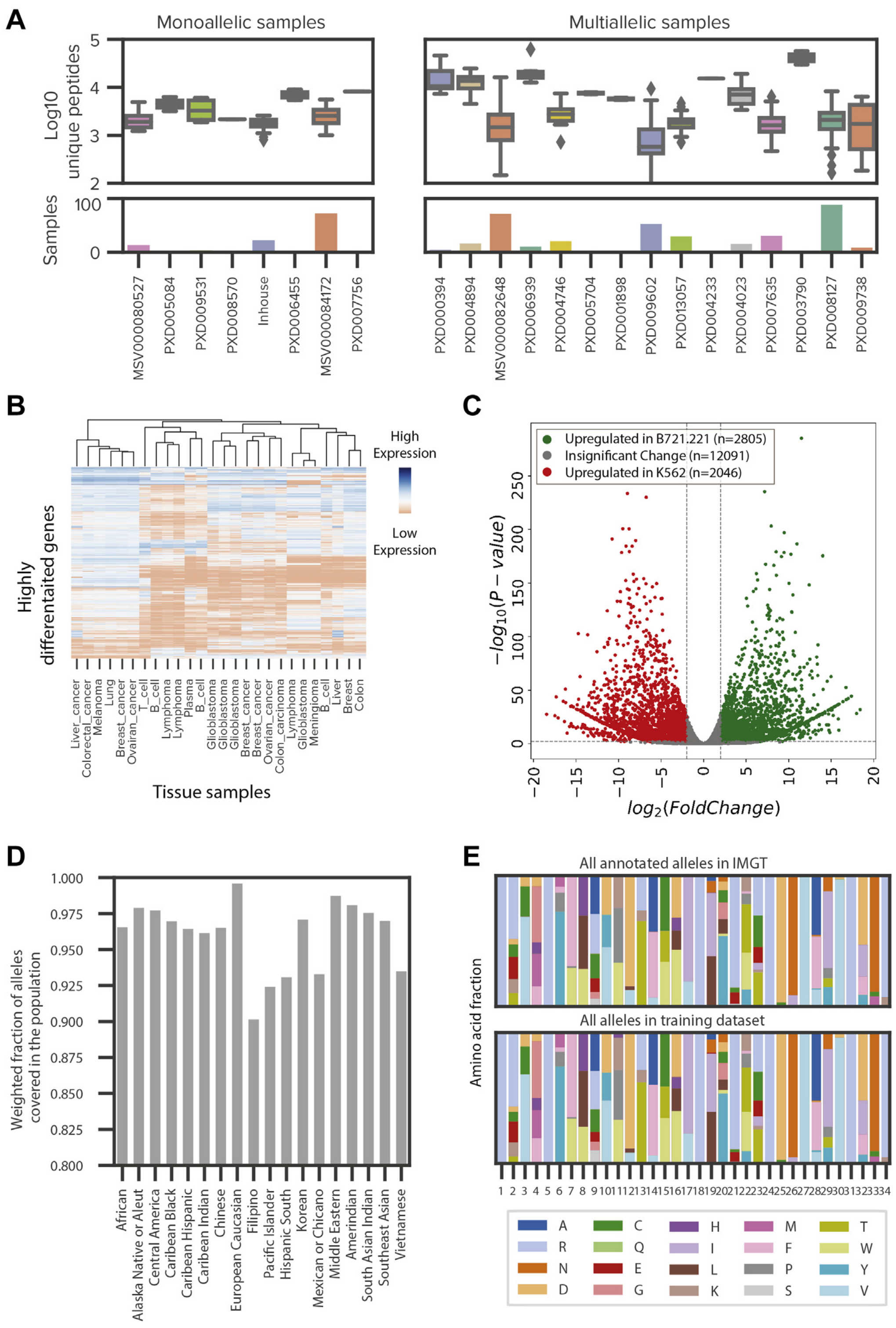


FIG. 3. **Systematic expansion of HLA ligandome through the incorporation of publicly available data.** *A*, box plots representing the number of unique peptides per sample from monoallelic and multiallelic immunopeptidomics samples that were reprocessed through our pipeline. Bar plot showing the number of samples for each project. Samples are colored according to their project. Peptide yields are log10 transformed. See supplemental Table S2 for additional details. *B*, a heatmap of expression values (TPM) of highly differentiated genes between tissue and tumor types of publicly available multiallelic immunopeptidomics data. Low expression is shown with *red*, and high expression is shown with *blue*. *C*, a volcano plot denoting the differential gene expression between the monoallelic parental cell lines, B721.221 and K562. Gene transcripts with significant upregulation in B721.221 compared with K562 are shown in *green* while gene transcripts with significant

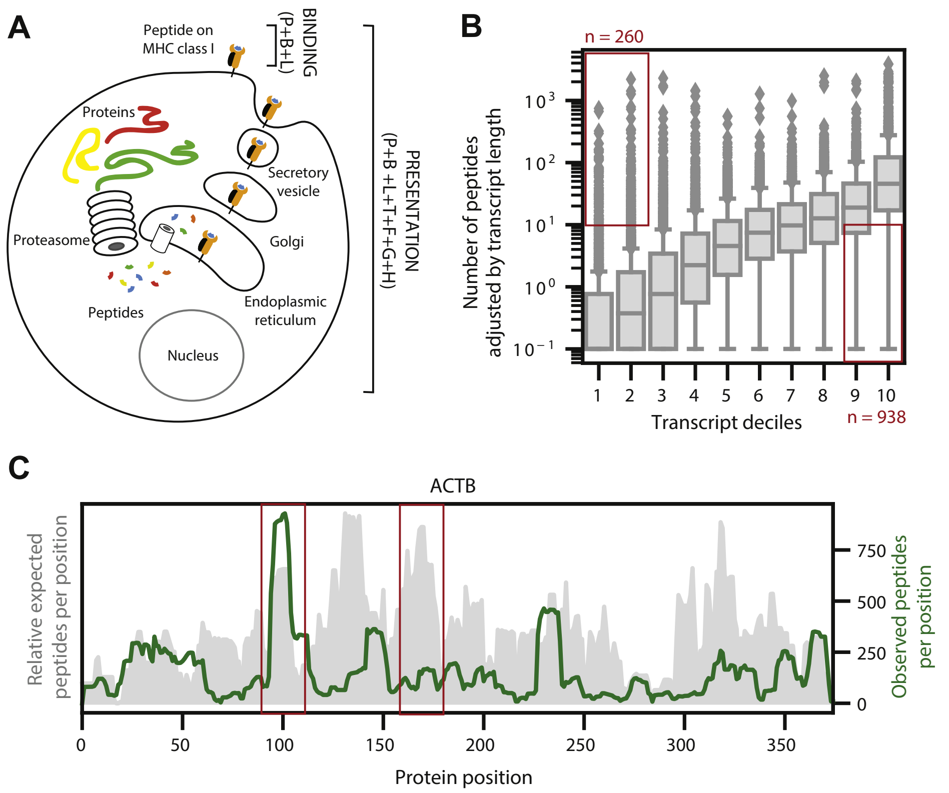


FIG. 4. **Modeling binding and presentation.** *A*, a schematic showing the difference between MHC binding and MHC presentation. MHC binding involves the ability of an MHC allele to bind to a paired peptide and is modeled with the peptide (P), allele-binding pocket (B), and peptide length (L). MHC presentation involves all steps in the antigen processing pathway in addition to MHC binding and is modeled with the peptide (P), allele-binding pocket (B), peptide length (L), gene expression (T), flanking regions around the peptide (F), propensity of the gene to engender peptides (G), and propensity of the region within the gene to engender peptides (H). *B*, boxplots representing the distribution of peptides per transcript observed in the reprocessed multiallelic immunopeptidomics data across transcript deciles. The peptides observed are normalized by transcript length. *Red boxes* denote the transcripts that generate many observed peptides despite low expression levels and transcripts that generate few observed peptides despite high expression levels. *C*, the distributions of expected and observed peptides from across the ACTB protein. Expected peptides, shown in *gray*, are generated by summing the number of frequent alleles predicted to bind each peptide (Rank <2 by netMHCpan4.0). The 30 most frequent alleles in the reprocessed multiallelic immunopeptidomics dataset were used for the analysis. Observed peptides are measured from the reprocessed multiallelic immunopeptidomics data and are shown in *green*.

with negative examples in a 1:999 ratio to mimic true ratio (5, 23–26). We chose this prevalence to reflect the underlying characteristics of peptide presentation (5, 23–26). The MONO-binding model significantly outperformed NetMHCpan-4.1-BA, NetMHCpan-4.1-EL, and MHCFlurry-2.0-BA (0.53, 0.34, 0.44, and 0.40, respectively, Fig. 5*C*, supplemental Data 1*A*). When we restricted the input data to be only the publicly available monoallelic data (PUBLIC-binding), SHERPA still outperformed the other three models, suggesting that the XGBoost modeling approach or our strict data curation added significant value (supplemental Fig. S9*A*). The binding model trained on the full dataset (SHERPA-binding), comprising both monoallelic and pseudo monoallelic data, had the same PPV compared with MONO-binding, a model trained on monoallelic data alone (0.53 and 0.53, respectively; Fig. 5*C*). SHERPA-presentation has a better PPV compared with SHERPA-binding, attesting to the utility of presentation-specific features. We further evaluated the contribution of each of our features and confirmed that each feature has an additive effect on performance (Fig. 5*D*). Flanking regions had

upregulation in K562 compared with B721.221 are shown in *red*. Gene transcripts with no significant up- or downregulation are shown in *gray*. *D*, a bar plot denoting the weighted fraction of alleles in 18 ethnicity populations from the National Marrow Donor Program within the expanded training dataset, including monoallelic cell lines profiled in house, public monoallelic data, public multiallelic data, and binding assay data from IEDB. *E*, two stacked bar plots showing the frequencies of amino acids at each position in the pseudo binding pocket for all annotated alleles in IMGT (*top*) and all alleles from the expanded training dataset, including monoallelic cell lines profiled in house, public monoallelic data, public multiallelic data, and binding assay data from IEDB.

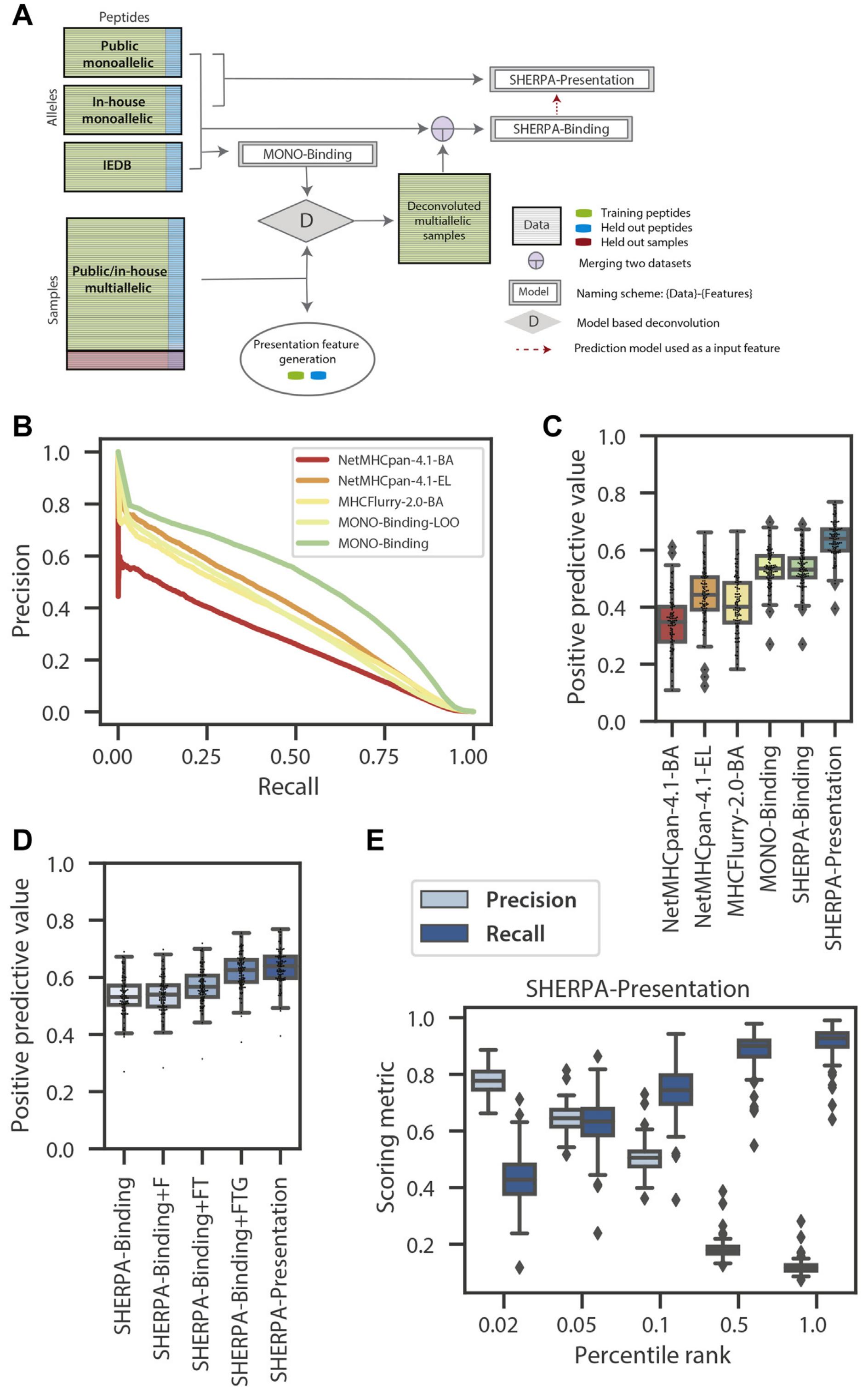


Fig. 5. **Overview of composite modeling approach and model performance.** *A*, a schematic of the composite modeling approach. Inhouse monoallelic immunopeptidomics data, public monoallelic immunopeptidomics data, and IEDB data are used to train MONO-binding. MONO-binding is used to deconvolute the multiallelic immunopeptidomics data to create pseudo monoallelic data. All monoallelic and pseudo monoallelic data is combined to train the SHERPA-binding model. The SHERPA-binding model is used as a feature along with other presentation features to train the SHERPA-presentation model on monoallelic immunopeptidomics data. *B*, a precision–recall curve demonstrating the predicted pan-performance on unseen alleles (MONO-binding-LOO) compared with MONO-binding and NetMHCpan4.1-BA, NetMHCpan-4.1-EL, MHCFlurry-2.0-BA. A model was trained for each allele with the data for that allele excluded from the training dataset. The MONO-binding-

a relatively minor impact on performance (~0.01 PPV gain compared with SHERPA-binding). We found the largest performance gain with the addition of abundance (TPM, ~0.03 increase in PPV above SHERPA-Binding+F) and gene propensity (~0.05 increase in PPV above SHERPA-Binding+FT), but we also observed significant gain with the addition of the hotspot score (~0.01 PPV above SHERPA-Binding+FTG). For an orthogonal allele-specific validation, we assessed the performance of the models using the held out IEDB data. Of the comparison models, NetMHCpan-4.1-BA had the best performance. MONO-binding and SHERPA-binding performed similarly to NetMHCpan-4.1-EL and MHCFlurry-2.0-BA (supplemental Fig. S9*B*, supplemental Data 1*B*). Finally, we evaluated the precision and recall at various rank values for both our multiallelic binding and presentation models. We observed a consistently and significantly higher precision and recall compared with other models at all rank values (Fig. 5*E*, supplemental Fig. S9, *C*–*E*). Using this analysis, we empirically determined a rank threshold of 0.1 to be our definition of a "binding" or "presented" peptide.

Experimental Validation of SHERPA Using Tissue Samples

To understand the utility of our prediction models in a patient sample, we performed immunopeptidomics on tumor samples from seven lung cancer patients and five colorectal cancer patients. In these experiments, we observed robust peptide yields (4043 median, supplemental Fig. S10*A*, supplemental Table S5, supplemental Data 1*C*). As expected, we did not observe any neoantigens from these patients in the immunopeptidomics data, as data-dependent acquisition (DDA)-based IP-MS protocols lack the required sensitivity of detection. A total of 46 alleles were represented across the 12 patients, and two of the alleles were outside of our training dataset. Thus, testing our performance on these samples provided an opportunity to assess our prediction accuracy on trained alleles, as well as the pan-allelic prediction capability on untrained alleles. To establish a prediction score for each peptide across all six potential patient alleles (patient-centric rank), we used the rank from the allele with the lowest rank to represent how well a peptide is bound or presented by that patient. We observed a strong performance of our models, which recovered 1.15 times more experimentally observed peptides among those that were predicted to be bound or presented than the next best model (MHCFlurry-2.0-BA) (Fig. 6*A*). Further, the multiallelic models outperformed the exclusively allele-specific models, highlighting the importance of the multiallelic deconvolution approach. Of note, we observed one sample with significantly lower recall than those from the other patients across all models. We explored that patient further and found that the patient has human leukocyte antigen (HLA) loss of heterozygosity across two HLA genes (-A and -B; supplemental Fig. S10*B*), highlighting the importance of considering a comprehensive view of the patient HLA features during neoantigen prediction. Although all models used a percentile ranking approach, we found that MHCFlurry-2.0-BA predicts an average of 30 percentile rank for negative peptides (compared with the expected ~50 percentile rank for all other models), suggesting that the sensitivity metric used in this analysis may be overrepresenting MHCFlurry's performance (supplemental Fig. S10, *C* and *D*). To further corroborate our results, we performed the same analysis in the ~10% of the peptides from tissue data from ovarian cancer and colorectal cancer studies that had been held out from multiallelic training (27, 28). We found similar trends across both of these external datasets, on data that were not part of our training data, with the SHERPA-presentation model showing the best results, showing 1.03 to 1.22 times more experimentally observed peptides predicted than MHCFlurry-2.0-BA (Fig. 6, *B* and *C*).

Immunogenicity

Although the SHERPA models do not specifically predict the likelihood that a peptide will be immunogenic, MHC presentation is a gatekeeping step of an immunogenic response. Thus, one would expect that all immunogenic peptides are successfully bound and presented on the cell surface. As an additional evaluation, we used the dataset of immunogenic peptides described by Chowell *et al.* (29) to evaluate the ability of the models to predict these epitopes. Across the entire dataset, we observe the best performance by the MONO-binding model and the SHERPA-binding model (Fig. 6*D*, supplemental Data 1*D*). We see variable but similar performance across the specific alleles. Of note, we observe a 1.11-fold improvement over the next best model for HLAA*02:01, the most studied allele (Fig. 6*E*). Together, this evidence suggests that the performance of the SHERPA models is generalizable and resilient to data sources.

DISCUSSION

Many personalized cancer immunotherapies require precise neoantigen identification. Although advances in next-generation sequencing technologies have enabled large-

LOO curve represents the predictions from each of the models on the test data of the allele excluded from the training data. *C* and *D*, boxplots denoting the distributions of positive predictive values (top 0.1%) across alleles within the monoallelic immunopeptidomics held-out test data. Distributions are shown for (*C*) NetMHCpan4.1-BA, NetMHCpan-4.1-EL, MHCFlurry-2.0-BA, MONO-binding, SHERPA-binding and SHERPA-presentation, and (*D*) SHERPA-binding, SHERPA-binding+F, SHERPA-binding+FT, SHERPA-binding+TTG, and SHERPA-presentation. *E*, boxplots showing the distribution of precision and recall values across alleles in the monoallelic immunopeptidomics data for SHERPA-presentation across several percentile rank thresholds. A percentile rank of 0.1 is selected as the optimal threshold.

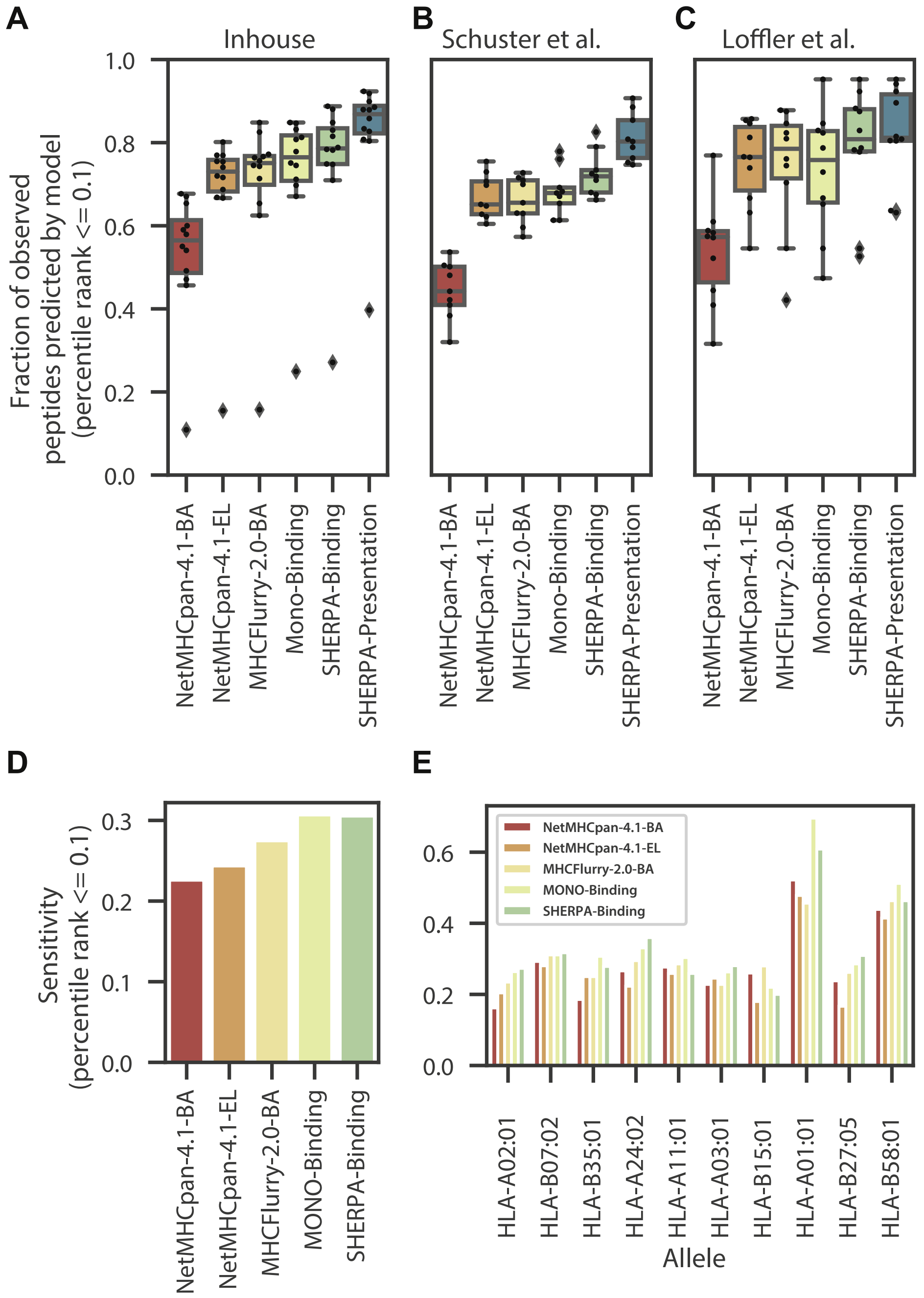


FIG. 6. **Performance of SHERPA on tissue samples and immunogenic epitopes.** Boxplots showing the distribution of prediction performance across (*A*) tumors profiled with immunopeptidomics in-house (lung and colorectal, *left*), (*B*) by Schuster *et al*. (ovarian, *middle*) and (*C*) Loffler *et al.* (colorectal, *right*). Performance is defined as the fraction of peptides observed with immunopeptidomics that are predicted to bind in the top 0.1% of all peptides percentile rank ≤0.1. Performance is shown for the following models: NetMHCpan4.1-BA, NetMHCpan-4.1-EL, MHCFlurry-2.0-BA, MONO-binding, SHERPA-binding, and SHERPA-presentation. *D* and *E*, bar plots showing the sensitivity of NetMHCpan4.1-BA, NetMHCpan-4.1-EL, MHCFlurry-2.0-BA, MONO-binding, and SHERPA-binding on the Chowell *et al.* immunogenicity dataset: (*D*) performance across all epitopes and (*E*) performance across high frequency alleles.

scale survey of putative neoantigens, algorithmic neoantigen ranking continues to have limited accuracy. Our work seeks to improve precision neoantigen discovery by addressing both the training data and modeling approach. Since prediction models are highly dependent upon their training data, we generated immunopeptidomics data using monoallelic cell lines that unambiguously map a peptide to an HLA allele. To ensure generalizability of SHERPA beyond our model system, we further expanded the HLA-peptidome by incorporating publicly available monoallelic immunopeptidomics data from different parental cell lines, allele-specific binding array data, and multiallelic tissue immunopeptidomics data. Moreover, we developed a modeling approach that integrates the data types strategically and mines the breadth of data to model antigen processing across a variety of tissue types.

While a combination of several factors leads to marked improvements in performance of SHERPA, we believe that three aspects set it apart. First, SHERPA employs the largest training data of any neoantigen prediction tool to our knowledge. Peptides from 167 unique human alleles and antigen processing from 30 unique expression profile backgrounds ensure the representativeness of our training data. Second, we designed features that mine extensive tissue-based immunopeptidomics data to model presentation and demonstrate ~0.06 improvement in our monoallelic PPV assessment with their inclusion. Third, SHERPA's composite modeling approach for integrating multiple data types reduces inherent biases caused by a single data source and allows for robust multiallelic deconvolution.

Our work is timely because it addresses two critical needs in the field. First, the Human ImmunoPeptidome Project outlined the need for large cohort immunopeptidomics studies in their effort to comprehensively map the human immunopeptidome (48). With this project, we contribute a new set of monoallelic immunopeptidomics data on a different cell line background than previous large studies and a previous set of lung and colorectal tumor tissues. Second, the scientific community has an urgent need for representing diverse ethnic populations in medical research and immunogenomics (18, 49). We addressed this issue by profiling five novel HLA alleles that are frequent in underrepresented Asian, African, and Middle Eastern populations.

Though our work shows great improvement, both SHERPA and the field of neoantigen prediction continue to face some key challenges. We integrated different data types (binding array and immunopeptidomics) to decrease bias and increase generalizability, whereas the vast majority of previous pMHC data was produced using very similar protocols. Accordingly, developing newer approaches to affinity purification other than the commonly used W6/32 antibody immunoprecipitation, applying newer data acquisition strategies such as data-independent acquisition, and generating immunopeptidomics data under different cell states are avenues for future improvement. In addition, gene expression is an imperfect surrogate for protein abundance. Other methods, such as ribosome footprinting, may provide a better estimate of protein levels. Moreover, some of the features used in the SHERPA-presentation model are exclusively designed to the human proteome without holding out specific proteins or protein subregions, so the results of the SHERPA-presentation model should be interpreted in light of the inclusion of all proteins and protein regions in the feature generation process. The SHERPA-binding-FT model is a more accurate representation of performance on nonhuman candidates. Finally, SHERPA does not specifically predict the likelihood that a ligand will be immunogenic; however, SHERPA captures a higher percentage of immunogenic epitopes than other methods, suggesting that peptides predicted by SHERPA will be more likely to be immunogenic.

In summary, we demonstrate that SHERPA is an important method that balances the tension between the clarity of monoallelic motifs and the generalizability of tissue data while consistently outperforming previous models. Its improved performance is expected to be useful in neoantigen prediction for immunotherapy.

DATA AVAILABILITY

The MS proteomics data have been deposited to the ProteomeXchange Consortium *via* the PRIDE partner repository with the dataset identifier PXD023064 (50). Evaluation data can be downloaded with the following links: https://doi.org/10.6084/m9.figshare.14916552.v1; https://doi.org/10.6084/m9.figshare.14916549.v1; https://doi.org/10.6084/m9.figshare.14916555.v1; https://doi.org/10.6084/m9.figshare.14916558.v1.

Supplemental data—This article contains supplemental data.

Acknowledgments—The cell line transfections were conducted at Thermo Fisher Scientific. The immunopeptidomics experiments were conducted at Cayman Chemicals and MS Bioworks, LLC. We thank Josette Northcott for her help in interpretation of the flow cytometry data. We thank Eric Levy, Euan Ashley, Russ B. Altman, and Atul Butte for their helpful feedback on the article. Personalis Inc provided the funding for this project.

Author contributions—R. M. P., D. M., R. C., and S. M. B. conceptualization; R. M. P., D. M., and S. D. data curation; R. M. P. and D. M. formal analysis; J. W. and R. C. funding acquisition; R. M. P., D. M., and S. M. B. methodology; S. M. B. project administration; C. W. A. resources; R. M. P., D. M., S. D., S. V. Z., N. P., J. H., and G. B. software; R. C. and S. M. B. supervision; R. M. P., D. M., and S. D. validation; R. M. P. and D. M. visualization; R. M. P. and D. M. writing—original draft; S. D., C. W. A., Sejal Desai, R. M., M. P. S., R. C., and S. M. B. writing—review and editing.

Conflict of interest—R. M. P., D. M., S. D., C. W. A., S. V. Z., N. P., J. H., G. B., Sejal Desai, R. M., J. W., R. C., and S. M. B are full- time employees of Personalis and owners of Personalis stock. M. P. S. co-founded Personalis and owns Personalis stock.

Abbreviations—The abbreviations used are: ATCC, American Type Culture Collection; B, binding pocket (model feature); ELISA, enzyme-linked immunosorbent assay; F, flanking regions (model feature); FDR, false discovery rate; G, gene propensity (model feature); GFP, green fluorescent protein; H, hotspot score (model feature); HLA, human leukocyte antigen; IEDB, Immune Epitope Database and Analysis Resource; IMGT, International ImMunoGeneTics Information System; L, peptide length (model feature); LC-MS/MS, liquid chromatography with tandem mass spectrometry; LOO, leave one out model; MHC, major histocompatibility complex; NMDP, National Marrow Donor Program; P, peptide (model feature); P-models, primary models; pMHC, major histocompatibility complex-peptide; SHERPA, Systematic HLA Epitope Ranking Pan Algorithm; T, protein abundance as measured by TPM (model feature); TPM, transcripts per million.

Received December 22, 2020, and in revised from, May 7, 2021
Published, MCPRO Papers in Press, June 12, 2021, https://doi.org/10.1016/j.mcpro.2021.100111

REFERENCES

1. Wells, D. K., van Buuren, M. M., Dang, K. K., Hubbard-Lucey, V. M., Sheehan, K. C. F., Campbell, K. M., Lamb, A., Ward, J. P., Sidney, J., Blazquez, A. B., Rech, A. J., Zaretsky, J. M., Comin-Anduix, B., Ng, A. H. C., Chour, W., *et al*. (2020) Key parameters of tumor epitope immunogenicity revealed through a Consortium approach improve neoantigen prediction. *Cell* **183**, 818–834
2. Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., Modrusan, Z., Mellman, I., Lill, J. R., and Delamarre, L. (2014) Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576
3. Schumacher, T. N., and Schreiber, R. D. (2015) Neoantigens in cancer immunotherapy. *Science* **348**, 69–74
4. Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., Oseroff, C., Yuan, L., Ruppert, J., Sidney, J., del Guercio, M. F., Southwood, S., Kubo, R. T., Chesnut, R. W., *et al*. (1994) The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* **153**, 5586–5592
5. Hunt, D., Henderson, R., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A., Appella, E., and Engelhard, V. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263
6. Andreatta, M., Lund, O., and Nielsen, M. (2013) Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* **29**, 8–14
7. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326
8. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., and Buus, S. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* **2**, e796
9. Andreatta, M., Alvarez, B., and Nielsen, M. (2017) GibbsCluster: Unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45**, W458–W463
10. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G., and Gfeller, D. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* **13**, e1005725
11. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020) 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 Res.* **48**, W449–W454
12. Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., Ojo, N. C., Caldwell, K., Abhyankar, J., Boucher, T., Hart, M. G., *et al*. (2018) Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* **37**, 55–63
13. Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., *et al*. (2020) A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209
14. O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018) MHCflurry: Open-source class I MHC binding affinity prediction. *Cell Syst.* **7**, 129–132
15. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368
16. Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I. K. A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., Riemer, A. B., Velculescu, V. E., Anagnostou, V., Pagel, K. A., and Karchin, R. (2020) High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunol. Res.* **8**, 396–408
17. O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020) MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* **11**, 418–419
18. Diversifying clinical trials. *Nat. Med.* **24**, (2018), 1779
19. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012) PEAKS DB:de NovoSequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587
20. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550
21. Chen, T., and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785
22. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015) Hyperopt: A Python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **8**, 014008
23. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015) Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14**, 658–673
24. Rammensee, H. G., Friede, T., and Stevanovíc, S. (1995) MHC ligands and peptide motifs: First listing. *Immunogenetics* **41**, 178–228
25. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanović, S. (1999) SYFPEITHI: Database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213–219
26. Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A., and Peters, B. (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–D412
27. Schuster, H., Peper, J. K., Bösmüller, H. C., Röhle, K., Backert, L., Bilich, T., Ney, B., Löffler, M. W., Kowalewski, D. J., Trautwein, N., Rabsteyn, A., Engler, T., Braun, S., Haen, S. P., Walz, J. S., *et al*. (2017) The immunopeptidomic landscape of ovarian carcinomas. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9942–E9951
28. Löffler, M. W., Kowalewski, D. J., Backert, L., Bernhardt, J., Adam, P., Schuster, H., Dengler, F., Backes, D., Kopp, H. G., Beckert, S., Wagner, S., Königsrainer, I., Kohlbacher, O., Kanz, L., Königsrainer, A., *et al*. (2018) Mapping the HLA ligandome of colorectal cancer reveals an imprint of malignant cell transformation. *Cancer Res.* **78**, 4627–4641
29. Chowell, D., Krishna, S., Becker, P. D., Cocita, C., Shu, J., Tan, X., Greenberg, P. D., Klavinskis, L. S., Blattman, J. N., and Anderson, K. S. (2015) TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1754–E1762
30. Khodadoust, M. S., Olsson, N., Wagar, L. E., Haabeth, O. A., Chen, B., Swaminathan, K., Rawson, K., Liu, C. L., Steiner, D., Lund, P., Rao, S., Zhang, L., Marceau, C., Stehr, H., Newman, A. M., *et al*. (2017) Antigen

presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature* **543**, 723–727

31. Ternette, N., Olde Nordkamp, M. J. M., Müller, J., Anderson, A. P., Nicastri, A., Hill, A. V. S., Kessler, B. M., and Li, D. (2018) Immunopeptidomic profiling of HLA-A2-positive triple negative breast cancer identifies potential immunotherapy target antigens. *Proteomics* **18**, e1700465
32. Shraibman, B., Barnea, E., Kadosh, D. M., Haimovich, Y., Slobodin, G., Rosner, I., López-Larrea, C., Hilf, N., Kuttruff, S., Song, C., Britten, C., Castle, J., Kreiter, S., Frenzel, K., Tatagiba, M., *et al*. (2019) Identification of tumor antigens among the HLA peptidomes of glioblastoma tumors and plasma. *Mol. Cell. Proteomics* **18**, 1255–1268
33. Pearson, H., Daouda, T., Granados, D. P., Durette, C., Bonneil, E., Courcelles, M., Rodenbrock, A., Laverdure, J. P., Côté, C., Mader, S., Lemieux, S., Thibault, P., and Perreault, C. (2016) MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126**, 4690–4701
34. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., Busch, D. H., Peschel, C., *et al*. (2016) Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404
35. Chong, C., Marino, F., Pak, H., Racle, J., Daniel, R. T., Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2018) High-throughput and sensitive immunopeptidomics platform reveals profound interferonγ-mediated remodeling of the human leukocyte antigen (HLA) ligandome. *Mol. Cell. Proteomics* **17**, 533–548
36. Löffler, M. W., Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., Schroeder, C. M., Trautwein, N., Hilke, F. J., Zinser, R. S., Mühlenbruch, L., Kowalewski, D. J., Schuster, H., Sturm, M., Matthes, J., Riess, O., *et al*. (2019) Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med.* **11**, 28
37. Laumont, C. M., Daouda, T., Laverdure, J. P., Bonneil, É., Caron-Lizotte, O., Hardy, M. P., Granados, D. P., Durette, C., Lemieux, S., Thibault, P., and Perreault, C. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238
38. Shraibman, B., Kadosh, D. M., Barnea, E., and Admon, A. (2016) Human leukocyte antigen (HLA) peptides derived from tumor antigens induced by inhibition of DNA methylation for development of drug-facilitated immunotherapy. *Mol. Cell. Proteomics* **15**, 3058–3070
39. Mommen, G. P., Frese, C. K., Meiring, H. D., van Gaans-van den Brink, J., de Jong, A. P., van Els, C. A., and Heck, A. J. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4507–4512
40. Illing, P. T., Pymm, P., Croft, N. P., Hilton, H. G., Jojic, V., Han, A. S., Mendoza, J. L., Mifsud, N. A., Dudek, N. L., McCluskey, J., Parham, P., Rossjohn, J., Vivian, J. P., and Purcell, A. W. (2018) HLA-B57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome. *Nat. Commun.* **9**, 4693
41. Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H. G., and Stevanović, S. (2017) Unveiling the peptide motifs of HLA-C and HLA-G from naturally presented peptides and generation of binding prediction matrices. *J. Immunol.* **199**, 2639–2651
42. Kaur, G., Gras, S., Mobbs, J. I., Vivian, J. P., Cortes, A., Barber, T., Kuttikkatte, S. B., Jensen, L. T., Attfield, K. E., Dendrou, C. A., Carrington, M., McVean, G., Purcell, A. W., Rossjohn, J., and Fugger, L. (2017) Structural and regulatory diversity shape HLA-C protein expression levels. *Nat. Commun.* **8**, 15924
43. Alpízar, A., Marino, F., Ramos-Fernández, A., Lombardía, M., Jeko, A., Pazos, F., Paradela, A., Santiago, C., Heck, A. J. R., and Marcilla, M. (2017) A molecular basis for the presentation of phosphorylated peptides by HLA-B antigens. *Mol. Cell. Proteomics* **16**, 181–193
44. Yair-Sabag, S., Tedeschi, V., Vitulano, C., Barnea, E., Glaser, F., Melamed Kadosh, D., Taurog, J. D., Fiorillo, M. T., Sorrentino, R., and Admon, A. (2018) The peptide repertoire of HLA-B27 may include ligands with lysine at P2 anchor position. *Proteomics* **18**, e1700249
45. Santos, E. J. M. D., McCabe, A., Gonzalez-Galarza, F. F., Jones, A. R., Middleton, D., and Middleton, D. (2016) Allele Frequencies Net Database: Improvements for storage of individual genotypes and analysis of existing data. *Hum. Immunol.* **77**, 238–248
46. Johansen, T. E., McCullough, K., Catipovic, B., Su, X. M., Amzel, M., and Schneck, J. P. (1997) Peptide binding to MHC class I is determined by individual pockets in the binding groove. *Scand. J. Immunol.* **46**, 137–146
47. Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017) "Hotspots" of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front. Immunol.* **8**, 1367
48. Vizcaíno, J. A., Kubiniok, P., Kovalchik, K. A., Ma, Q., Duquette, J. D., Mongrain, I., Deutsch, E. W., Peters, B., Sette, A., Sirois, I., and Caron, E. (2020) The human immunopeptidome project: A roadmap to predict and treat immune diseases. *Mol. Cell. Proteomics* **19**, 31–49
49. Peng, K., Safonova, Y., Shugay, M., Popejoy, A. B., Rodriguez, O. L., Breden, F., Brodin, P., Burkhardt, A. M., Bustamante, C., Cao-Lormeau, V. M., Corcoran, M. M., Duffy, D., Fuentes-Guajardo, M., Fujita, R., Greiff, V., *et al*. (2021) Diversity in immunogenomics: the value and the challenge. *Nat Methods* **18**, 588–591
50. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., *et al*. (2019) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450