Supplementary Methods

Extended methods surrounding our approach for neoantigen prediction (SHERPA; used to generate our NBS values) and calling HLA allele-specific loss of heterozygosity (DASH) are detailed below. These data represent core inputs for the generation of NEOPS, which incorporates both of these values.

Neoantigen prediction

We developed SHERPA, an algorithm for predicting pan-allelic MHC-peptide binding and presentation using allele-specific immunopeptidomics data generated using mono-allelic cell lines. To model native presentation we employed HLA-null K562 parental cells with stably transfected HLA alleles. SHERPA combines data from 128 mono-allelic and 384 multi-allelic samples with publicly available immunoproteomics data and binding assay data to improve generalizability and reproducibility. Details of our methodology are described below, as well as in the manuscript associated with the work (preprint version available through bioRxiv, Pyke et al, 2021)

Immunopeptidomics using mono-allelic cell lines

Experimental design

All peptides under evaluation were derived from the MHC-I immunopeptidome, with work performed on both tumor tissue and mono-allelic cell lines. For tumor tissue, 12 samples were processed without controls or replicates. Approximately 5x10^9 cells from 28 different mono-allelic cell lines were processed. Three biological replicates were required for each cell line. An HLA-null line was used as a negative control. Peptides from 8 of the cell lines were processed a single time, and peptides from the other 20 cell lines were divided and processed on the mass spectrometer twice (11 with CID only, 9 with both CID and EThcD). Publicly available peptides were also analyzed to compliment data obtained from cell line studies. Wald test and two-sided T tests were used to determine significance.

Cell culture

Mono-allelic cell lines were generated using Jump-In™ technology (Thermo Fisher Scientific) to stably transfect K562 parental cells with a single allele of interest. Optimized plasmids containing sequences for HLA, beta 2 microglobulin (B2M) and IRES promoter were synthesized for each allele (GeneArt). Cultured cells with confirmed plasmid integration were expanded to 500M cells. Cells were pelleted following confirmation of surface expression of target alleles using flow cytometry (using W6/32 antibody). Transfection experiments were conducted by Thermo Fisher Scientific.

Immunoprecipitation of peptide-MHC complexes

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

Peptide sequencing using LC-MS/MS

Peptides were first desalted using solid phase extraction (SPE; Empore C18), then loaded on a column and eluted using 80/20 acetonitrile/water (0.1% TFA). Samples were then lyophilized prior to storage. Samples were reconstituted in 0.1% TFA prior to analysis using liquid chromatography mass spectrometry (LC/MS/MS). A Waters NanoAcquity system with a 2h gradient was used for chromatographic separation. Peptides were then analyzed using a ThermoFisher Fusion Lumos 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: 3s). MS Bioworks, LLC performed all mass spectrometry experiments.

Peptide identification

PEAKS (PEAKS Studio 10.0 build 20190129) (Zhang et al. 2012) was used to identify peptides using the default two-step identification workflow: mass tags are first identified using de novo sequencing, followed by a second database search step where a subset of putative proteins are identified using mass tags identified in step one. The following settings were used in the workflow - Protein database: Swissprot proteome database (20,402 entries; dated 03-20-2019); Fragment mass tolerance: 0.02 Da; precursor mass tolerance: 10 ppm; Enzyme specificity: none; Variable modifications: oxidation of Methionine (+15.9949), N-terminal acetylation (+42.0106); Fixed modifications: carbamidomethylation of cysteine (+57.0215). Peptide-to-spectrum matches (PSMs) were filtered at 1% false discovery rate (FDR), estimated using decoy sequences.

Post processing and quality control

Spurious peptides identified at 1% FDR were further filtered using the following criteria: spurious peptides were defined as those from highly chimeric spectra (n>2), polymeric peptides, as well as identifications with fewer than 10 fragment ions. A mock transfection (GFP) was used to profile and remove background contaminants. Motif signatures were identified manually using Gibbs Clustering software (Andreatta, Alvarez, and Nielsen 2017) and peptide yields (cutoff of 500 peptides per sample).

Extraction and processing of public data sets

Extraction and processing of publicly available mono- and multi- allelic data

Raw data (.raw files) was downloaded from publicly-available sources, and HLA type-to-sample mappings were obtained from the respective publications. Data was then processed using a similar approach to in-house generated data. Peptide identifications from samples that passed QC as detailed above were aggregated.

in vitro binding affinity data

Raw HLA-peptidome data was downloaded from the Immune Epitope Database (IEDB) (date: 03-09-2020). We then filtered the data for entries corresponding to *in vitro* binding assays. Non-nM entries and any non-linear peptides were excluded. Additionally, only four digit MHC class I peptides of length 8-11 were included. The peptides were derived from the following 'Method/Technique' categories: 'purified MHC/competitive/fluorescence', 'purified MHC/competitive/radioactivity', 'purifiedMHC/direct/fluorescence', 'cellular MHC/direct/fluorescence', 'cellularMHC/competitive/fluorescence', 'cellular

MHC/competitive/radioactivity', 'lysate MHC/direct/radioactivity' and 'binding assay'. An IC50 value >500nm was used to classify ligands as binders.

Measurement and analysis of transcript expression

Sequencing of mono-allelic cell lines

Representative samples of mono-allelic K562 cell lines were sequenced in triplicate using ImmunoID NeXT™ according to the same approach outlined in the primary methods section. Reads were aligned using STAR, and transcript per million (TPM) values were determined. Data from the B721.221 cell line, was obtained from the American Type Culture Collection (ATCC) and assayed in triplicate using ImmunoID NeXT to minimize platform-related variance.

Generation of transcriptomic data for external immunopepdomics data

Transcriptomic data was re-created using ImmunoID NeXT for publicly available mono-allelic datasets generated with B721.221 cell lines (MSV000080527; MSV000084172) (Abelin et al. 2017; Sarkizova et al. 2020), following the same protocol as described previously. All additional samples in the expanded dataset use imputed TPM values obtained from our internal database that were generated by using the median value of all samples from the same tissue type.

Building prediction models and feature engineering

Data splitting for training and evaluation

Performance was assessed on novel peptides. First we ensured that no peptides had overlapping 8-mer cores in either the training, testing or validation datasets. We then used the unique set of peptides from the IEDB, immunopeptidomics, mono-allelic immunopeptidomics, and multi-allelic datasets. Next, peptides containing any identical 8-mer substrings ("nested" peptides) were grouped, with each peptide group assigned to one of ten different subsets. After approximately equal peptide numbers of peptides had been added to each subset, we selected one subset for validation (~10%), one for testing (~10%) and assigned the remaining eight subsets to training (~80%). 32 multi-allelic samples were held out from the training dataset and assessed using the ~10% test subset from the fully held out samples. feature engineering of the gene propensity and hotspot scores was conducted using all publicly available multi-allelic data (except for the 32 held out samples) to avoid potential bias that could arise from systematically holding out proteins or peptides. Deconvolution was performed using the training subset of the multi-allelic data.

Prediction model specifications

Two prediction algorithms were trained: one that models peptide *binding* and a second which models peptide *presentation*. The binding algorithm uses the following inputs to model MHC peptide interaction affinities: amino acid sequence and length of peptide ligands and HLA binding pocket information. The presentation algorithm simultaneously models antigen processing *and* peptide-MHC (pMHC) binding, using the following information as inputs: amino acid sequence and length of peptide ligands, HLA binding pocket information, proteasomal processing footprints identified in the right and left flanking regions of peptide ligands, source protein abundance (captured as gene expression), as well as two additional features which model propensity for antigen processing and presentation. Features corresponding to these inputs are generated as described:

1. *HLA binding pocket (B):* A pseudo sequence of amino acids was used to represent the binding pocket as previously described (Nielsen et al. 2007). Briefly, 34 positions on the protein sequence of the HLA that are a distance of 4 Å or lesser in crystollagraphic structures were selected from the full protein to serve as the psuedo sequence. a BLOSUM62 substitution matrix was used to encode the amino acid sequence, with each amino acid represented by a 20-dimensional vector constituting the relative weights of amino acid substitutions. BLOSUM62 encoding was selected under the expectation that substitutions between evolutionarily similar amino acids will have a lower impact on epitope binding changes when compared to substitutions between dissimilar amino acids.

2. *Peptide ligand (P):* The peptide ligand amino acid sequence is encoded using a BLOSUM62 substitution matrix, with each amino acid represented by a 20-dimensional vector constituting the relative weights of amino acid substitutions. To account for the variable length of HLA ligands (8- to 11-mers), we employed a middle-padding approach which inserts blanks into the middle of the to adjust all peptides to a length of 11 amino acids as described previously (Bulik-Sullivan et al. 2018). Briefly, for every peptide with less than 11 amino acids, we assign one to three 20-dimensional vector(s) of zeros to the middle of the peptide encoding to reach the full 11-mer length. This approach was used to ensure a pan-length algorithm that maintains peptide anchors in consistent columns of the matrix.

3. *Peptide length (L):* Peptide length is defined as the number of amino acids in the peptide ligand.

4. *Left and right flanking regions (F):* 5-mer peptide sequences to the left and right of the peptide ligand in the source protein are set as the left and right flanking regions respectively. In the event of multi-mapping the protein with the highest transcript expression is assigned to the peptide. Left and right flanking 5-mers are encoded using BLOSUM62 substitution matrix.

5. *Hotspot score (H):* The hotspot score was estimated using publicly available multi-allelic data. Peptide ligands from each sample are mapped to source proteins. In the event of mapping of peptides to multiple proteins, all potential mappings are used. The hotspot score of each amino acid was then calculated as the number of peptides overlapping each particular amino acid. For peptides, hotspot scores were determined by mapping the peptide to its source protein then taking the average of the hotspot score spanning the peptide.

6. *Gene propensity score (G)*: Gene propensity was estimated using publicly available multi-allelic data. Peptide ligands from each sample are redundantly mapped to source proteins then to transcripts. The number of peptides mapping to each transcript-associated protein was determined. The expected number peptides for each transcript-associated protein was calculated using the TPM for each protein across all multi-allelic data sources, then normalizing by protein length and the number of peptides derived from each sample. The number of expected and observed peptides for each protein are then summed across all multi-allelic datasets. Lastly, observed values were divided by expected values to

calculate a gene propensity score for each protein (gene). To deprioritize potential pseudogenes a score of -3 was given to all proteins without any observed transcripts across all samples.

7. *Abundance of source protein (T):* Peptide ligands are redundantly assigned to all source proteins, followed by transcripts. The most highly expressed transcript (calculated as the TPM) is chosen. Both the TPM and transcript are then assigned to the peptide ligand.

The contribution of each feature to the XGBoost model was then evaluated using the 'gain' metric, which measures the relative contribution of each feature by aggregating individual contributions of the feature in each tree. Increased values indicate greater contribution.

Generation of non-binding negative examples

We synthetically generated non-binding negative examples as our immunopeptidomics experiments only generated peptides that were successfully bound to and presented by MHC molecules. For every positive example in our training and validation datasets we generated 20 negative examples. Negative examples were generated by randomly selecting a protein from the Swissprot proteome (downloaded on 03-20-2019), then randomly selecting a peptide from within that protein. Peptides were selected to have equal probability of being length 8, 9, 10 or 11. Flanking regions were the true flanking regions around the selected peptides. Gene expression values (TPM) were randomly assigned by selecting a transcript from the the associated positive example. Gene propensity and hotspot scores were assigned based on the protein and peptide position within the protein selected as the negative example.

Training the prediction algorithm

XGBoost, a gradient boosted decision tree algorithm was used to train all models (Chen and Guestrin 2016). To train the algorithm all encoded and numeric features were provided as a vectorized input feature vector. Sequential model-based optimization was performed to select an optimal set of training parameters using HyperOpt (Bergstra et al. 2015). The parameters used for final training were: 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.

Calibrating raw scores using percent rank values

A set of 500,000 peptides were selected at random from the human proteome. After training a model, the predictions across the set of random peptides were calculated with every allele. Next, raw prediction probability (output of XGBoost) for each allele was used to rank the random peptides. For each newly predicted peptide, the assigned rank equals the percentage of the random set that is predicted to bind or to be presented with a better raw score than the new peptide. Ranks range from 0-100, with lower scores indicating peptides with better binding or presentation. Ranks are re-calculated for each allele and model combination.

Applying prediction models to patient samples and multi-allelic cell lines

Binding ranks for peptides in the public multi-allelic samples (cell lines, tissues and tumors) with immunopeptidomics data were generated using the MONO-Binding model. Peptides were required to have either 8, 9, 10 or 11 amino acids to be considered. 10% of multi-allelic data was held out for testing. For each sample predictions were made for all HLA alleles (up to six). Samples that lacked HLA typing were excluded from analysis.

Deconvolution of multi-allelic data

Data for each multi-allelic sample was deconvoluted to determine which allele-peptide pairs to include in the training dataset for the final model. Allele-peptide pairs with a predicted binding rank of ≥ 0.5 were excluded, removing all peptides which fail to bind to any of the designated alleles. Second, in the case of multiple alleles predicted to bind to a specific peptide, the strongest binding allele-peptide pair (lowest rank) was selected, and all other pairs excluded. Next, duplicate allele-peptide pairs were removed. Lastly, for every new positive example derived from the multi-allelic data, 20 negative examples were generated (as described previously).

Training composite models

A total of three prediction models were trained for our composite model. Five additional models were trained to determine features that contribute to optimal performance. Prediction models in our composite model were trained as follows:

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

The five additional prediction models were trained as follows:

1. *PUBLIC-Binding*: Trained using public mono-allelic immunopeptidomics data with B, P, and L as features.
2. *MONO-Binding-LOO*: 126 allele-specific models trained using IEDB, public mono-allelic immunopeptidomics and in-house mono-allelic immunopeptidomics data with B, P, L, IEDB-Binding, PUBLIC-Binding and INHOUSE-Binding as features. For training each allele-specific model, peptides from the respective alleles were held out from the training dataset.
3. *SHERPA-Binding+F*: Trained using IEDB, public mono-allelic immunopeptidomics, in-house mono-allelic immunopeptidomics and deconvoluted multi-allelic immunopeptidomics data with SHERPA-Binding and F as features.
4. *SHERPA-Binding+FT*: Trained using IEDB, public mono-allelic immunopeptidomics, in-house mono-allelic immunopeptidomics and deconvoluted multi-allelic immunopeptidomics data with SHERPA-Binding, F and T as features.
5. *SHERPA-Binding+FTG*: Trained using the IEDB, public mono-allelic immunopeptidomics, in-house mono-allelic immunopeptidomics and deconvoluted multi-allelic immunopeptidomics data with SHERPA-Binding, F, T and G as features.

Benchmarking and evaluation of prediction models

Generation of the mono-allelic held-out test data

Approximately 10% of positive examples were held out of the training and validation datasets for the mono-allelic immunopeptidomics datasets (public and in-house). 999 negative examples were added for each positive example to reflect the positive to negative ratio commonly accepted in the field (Bassani-Sternberg et al. 2015; Hunt et al. 1992; H. G. Rammensee, Friede, and Stevanoviíc 1995; H. Rammensee et al. 1999; Vita et al. 2015). Approximately 10% of the negative and positive data was withheld from the training dataset for the IEDB data. To maintain the same positive to negative peptide ratio found in IEDB no supplementary negative examples were added to the test dataset.

Evaluation metrics

Performance of the prediction models was evaluated using three distinct metrics:

1. *Positive predictive value (PPV)*: PPVs were calculated using the full test dataset to make predictions. The percentage of peptides in the top X% of predictions that are positive examples was determined, with X representing the positive example portion of the dataset. Individual PPVs were calculated for the peptides of each allele, then combined using a median yielding a single value.
2. *Precision-recall curves*: Precision-recall curves were created by calculating the precision and the recall for all possible cutoffs, then plotting them as a single line.
3. *Fraction of observed peptides predicted by model*: This metric was employed for all multi-allelic tumor validation analyses. First, a single score representing the best (lowest) rank predictions is selected to represent each peptide observed with immunopeptidomics. The score is then determined by calculating the percentage of observed peptides with a rank of ≤ 0.1.

Leave-one-out pan-allelic analysis

Pan-allelic performance of the MONO-Binding model used for model-based deconvolution was evaluated by training 126 independent models with the same features as the MONO-Binding model. For each model, the peptides from a specific allele from the set of peptides used to train the MONO-Binding model were excluded. The predicted motif for each allele was generated by predicting the binding rank for 500,000 random peptides for a given allele using the model for which the allele had been excluded from training. Next, the motif for peptides with the top percentile of binding ranks were generated. For each allele a threshold of 50 positive peptides in the training data was set as a cutoff for visualization of the motif. Precision recall curves were created using all 126 models to predict binding ranks for the mono-allelic immunopeptidomics data (public and in-house) that was excluded from training and validation (10% test dataset). Predictions for each allele were made using the model which excluded that allele from training.

Generating validation data using tissue samples

For patient validation 12 paired fresh frozen tumor/normal cases (five colorectal and seven lung) were obtained from a commercial biobank. All patients were consented, and specimen collected under IRB-approved protocols. Tumor samples were divided into two pieces for further evaluation. The first portion of tumor was used for immunoprecipitation of MHC complexes followed by LC-MS-MS (as described above) to generate the immunopeptidomics data. DNA from the paired normal, and DNA/RNA from the remaining tumor tissue was extracted and analyzed with ImmunoID NeXT (as described above).

Extraction of held-out validation data sets from external multi-allelic samples

To generate the test dataset we withheld approximately 10% of multi-allelic immunopeptidomics data with non-overlapping 'nested' peptides from the training and validation datasets. 10 samples from two specific datasets were used to validate our internal tumor immunopeptidomics performance (datasets: PXD007635, PXD009602) (Schuster et al. 2017; Löffler et al. 2018). Only samples with 4-digit resolution typing of HLA-A, -B and -C were used for the analysis.

Immunogenicity evaluation

To elicit an immunogenic response, a peptide must be presented on the cellular surface by an MHC allele. As such, the ability of the algorithms to positively identify immunogenic peptides was evaluated using a publicly available dataset (Chowell et al. 2015). Percentage of immunogenic peptides that were predicted at <= 0.1 percentile rank was evaluated for the various models.

Running comparison prediction algorithms

NetMHCpan-4.1-BA, NetMHCpan-4.1-EL and MHCFlurry-2.0-BA prediction algorithms were used to benchmark comparative performance of SHERPA. Default settings were used when running all algorithms, and percentile rank outputs used for analysis.

HLA allele-specific deletion

HLA loss of heterozygosity (LOH) was detected using the machine-learning based algorithm DASH (Deletion of Allele-Specific HLAs) (a preprint version of the manuscript is available through bioRxiv, Pyke et al, 2021). Methodology surrounding the development of DASH are outlined below.

Patient data

The HLA allele-specific deletion algorithm was trained using paired tumor and adjacent normal or blood normal samples collected from 279 patients spanning 15 different tumor types (representing a subset of the 611 patients referenced in the following sections). These samples were purchased from various biobanks, and were collected under IRB-approved protocols from fully consented patients. For each case, paired fresh frozen or FFPE samples were profiled using ImmunoID NeXT as described elsewhere in the methods.

HLA typing and somatic mutations

NGS-based HLA-typing was performed using the paired blood or adjacent normal sample from each case. Polysolver (Shukla et al. 2015) was used to determine HLA types up to 6 digits using integrated tumor and normal data and default tool parameters.

Tumor copy number alterations, purity and ploidy estimates

For each sample allele-specific copy number alterations, tumor purity and ploidy were estimated using Sequenza (Favero et al. 2015) with default parameters.

HLA allele database

An imputation approach similar to (Shukla et al. 2015) was used to generate an HLA allele database which employs the multiple sequence alignment (MSA) format from IMGTv312 (Magadan et al. 2019). For exons in alleles and incompletely sequenced alleles the cDNA file was used to infer a reference allele with matching protein-level 4 digit nomenclature, or 2 digit when no 4 digit matches were identified. In the event of multiple 2 digit matches, the first allele listed in the MSA was used. Intronic regions of each allele were imputed following the same approach with the gDNA file. Full length genomic sequences were inferred for each allele by assembling introns from the gDNA imputation step and exons from the cDNA imputation set.

Feature assembly for DASH

Each individual's unique HLA alleles are used to create a subject-specific HLA reference. For HLA typing all reads that could map to the HLA region are collected using a 30 base pair seed, and mapped to the patient-specific HLA reference using BWA (Li and Durbin 2009). To preserve the highest quality coverage information reads were only considered for inclusion if they were an exact sequence match, and had soft clipping for <20% of their total length. If a somatic mutation was detected within the HLA alleles, stringency was reduced to allow single mismatch reads. Coverage was then calculated using Samtools (Li et al. 2009).

Positions of variance between alleles were determined by aligning patient-specific homologous alleles, allowing for the detection of both SNVs and indels. To ensure consistent weighting for SNVs and indels, only the first position of each indel was considered. Alleles were considered homozygous if fewer than five positions of difference between them were detected.

Six features were included; three allele-specific features extracted from the coverage data, two patient-specific features and one exome-scale feature:

1. Allele-specific coverage ratio: The ratio of tumor sample to normal sample coverage is calculated for each allele at each position of mismatch between the homologous alleles, then normalized by the exome-wide number of tumor reads divided by the exome-wide number of normal reads. This approach results in an allele-specific coverage ratio of one if there is no copy number variation, despite inter-run sequencing depth variability. For each bin, the median coverage ratio is then determined for each allele, and the smaller value between the two alleles is considered for that bin. The median value across all bins is then calculated, and used as the feature. Allele-specific coverage ratio has an expected value of 1 if there is no copy number variation, and a lower bound of 0. Lower allele-specific coverage ratios are interpreted as increased likelihood of LOH in an HLA gene.
2. Percentage coverage: Consistently observed lower coverage across all mismatch positions in an allele is more likely to be reflective of true HLA LOH. Accordingly, each allele is assigned a 0 or 1 for each bin depending on whether it has lower or higher coverage compared to its homologous allele. In the event allelic coverage of a bin cannot be determined (no mismatch sites), a value of 0.5 is assigned to each allele. The mean across all bins for each allele is then calculated, and the higher average assigned as the feature value. The feature value ranges from 0.5 to 1, with values closer to 1 representing increased probability of HLA LOH.
3. Adjusted b-allele frequency: The b-allele frequency (BAF) is calculated for the tumor and normal sample separately at each position of mismatch. The tumor BAF is then divided by the normal BAF to account for variability in probe capture of each specific allele. The ratios were collapsed into a single feature by first dividing the allele references into150 base pairs length bins, then calculating the absolute value of the median adjusted-BAF for each bin. The median value across all bins being is then used as the feature. This

feature has a lower bound of 0, with larger numbers indicating increased likelihood of LOH in the HLA gene.

4. Tumor purity: Tumor purity is obtained from Sequenza, with values ranging from 0.1 to 1, with 0.1 being the least pure tumor and 1 being the most pure tumor.
5. Tumor ploidy: Tumor ploidy is obtained from the Sequenza tumor ploidy estimation. Values are whole integers greater than or equal to one.
6. Deletion of flanking regions: As the majority of HLA LOH occurs due to large deletions, we developed a feature which captures deletions in the flanking regions of each gene, allowing us to leverage information from a greater number of variable sites. B-allele deletion calls from Sequenza are used to make a flanking region deletion call when a deletion is detected within 10,000 base pairs in either direction of the gene. This feature is binary, with 0 representing a deletion.

Training the DASH algorithm

279 patients across multiple tumor types were used to collect a set of 720 heterozygous genes. Scalable gold standard HLA LOH labeling is cost and labor prohibitive. As such, we visualized all features described for each heterozygous gene, then manually labeled each case of HLA LOH. To train DASH, 500 heterozygous genes were used for training and 220 heterozygous genes held out for testing the algorithm. An XGBoost algorithm was then used to learn how to detect HLA LOH in each pair of alleles using the features outlined above. A secondary check to ensure that the allele-specific ratio of the lost allele is < 0.98 and that adjusted-BAF is > 0.02 was performed to improve the specificity of DASH on low tumor purity samples. If the algorithm detected HLA LOH, the lower coverage allele was labeled as deleted. In the rare case of a bi-allelic deletion, where the algorithm detects HLA LOH and the allele with higher coverage has an allele-specific coverage ratio below 0.5 for at least 25% of the bins, both alleles are labeled as deleted.

Limit of detection analysis using cell lines

A paired tumor-normal lymphoblast cell line (NCI-H2009) was used to assess DASH' limit of detection across varying levels of tumor purity and clonality. In the NCI-H2009 cell line, HLA-A is homozygous while HLA-B*51:01 and HLA-C*15:02 alleles are both deleted. The tumor and normal cell lines were sequenced to 50x coverage and 30x coverage, respectively. To stimulate typical sequencing depths, normal data was downsampled to reflect 25x sequencing coverage. To create a model of decreasing tumor purity, we mixed increasing proportions of normal reads into decreasing proportions of tumor reads, with the combined tumor and normal reads always summing to an average of 35G sequencing coverage. All combinations of tumor and normal sub samples were derived from the same sequencing runs and performed in replicates of 10 using the seqkit library. Lower sub clonality was simulated by setting the proportion of tumor reads used in the mixture as the product of desired tumor purity and sub clonality. Tumor purity was then artificially increased to reflect the desired level of purity. We simulated samples without HLA LOH by including only normal reads and artificially increasing the estimated tumor purity to the desired level. Specificity was estimated using these runs.

Benchmarking performance against LOHHLA

DASH was benchmarked against a known and tested publicly available tool, LOHHLA (Loss of Heterozygosity in Human Leukocyte Antigen), which was used to call HLA LOH in both the test dataset and *in silico* cell line dilutions (McGranahan et al. 2017). LOHHLA was run using cutoffs specified in the manuscript: alternate allele copy number of < 0.5 and the p-value related to this allelic imbalance is < 0.01.

Allele-specific validation with digital PCR

Because each patient has a unique set of up to 6 HLA class I alleles, high specificity subject-specific primers and probes were designed for each patient. Due to the high degree of similarity between some homologous alleles, reliable probes and primers may not exist for all patients. Primers and probes were manually designed for eleven homologous allele pairs with DASH-predicted HLA LOH from one cell line and ten different patients to maximize descrimination between alleles. A probe targeting RNase P (RPP25) was included as a positive control, and water was used as a negative control. To facilitate multiplexing, HLA allele and RPP25 probes were assigned different fluorophores (FAM and VIC, respectively).

Primer and probe efficiency was assessed by performing dPCR in triplicate on DNA from normal and tumor samples (excluding one patient, which was performed in duplicate). Seven samples were fully independent, and the remaining three samples were from the training dataset. Lost and retained alleles are normalized by the control gene to account for sample input variation prior to analysis. Primer and probe design was deemed successful if the ratio of HLA allele copies to multiplexed RNaseP copies was 0.5 in the normal sample. This value was anticipated as the HLA allele is expected to be haploid, and RNaseP is expected to be diploid. For primer designs that met this requirement, allele:RNaseP ratio in tumor DNA is compared to allele:RNaseP ratio in the normal DNA using a one-sided T-test (null hypothesis: tumor ≥ normal) to determine if there has been a significant reduction in the tumor. Variability of dPCR measurements are assumed to follow a normal distribution. This test is performed for both the predicted retained, and lost allele. Allelic imbalance is identified as a significant difference between the predicted lost and retained alleles in the normal and tumor DNA. This validation approach focuses on specific gene sections and is not powered to identify small focal deletions in a small portion of the gene.

References

Abelin, Jennifer G., Derin B. Keskin, Siranush Sarkizova, Christina R. Hartigan, Wandi Zhang, John Sidney, Jonathan Stevens, et al. 2017. “Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-Allelic Cells Enables More Accurate Epitope Prediction.” *Immunity* 46 (2): 315–26.

Andreatta, Massimo, Bruno Alvarez, and Morten Nielsen. 2017. “GibbsCluster: Unsupervised Clustering and Alignment of Peptide Sequences.” *Nucleic Acids Research* 45 (W1): W458–63.

Bassani-Sternberg, Michal, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann. 2015. “Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation.” *Molecular & Cellular Proteomics: MCP* 14 (3): 658–73.

Bergstra, James, Brent Komer, Chris Eliasmith, Dan Yamins, and David D. Cox. 2015. “Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization.” *Computational Science & Discovery* 8 (1): 014008.

Bulik-Sullivan, Brendan, Jennifer Busby, Christine D. Palmer, Matthew J. Davis, Tyler Murphy, Andrew Clark, Michele Busby, et al. 2018. “Deep Learning Using Tumor HLA Peptide Mass Spectrometry Datasets Improves Neoantigen Identification.” *Nature Biotechnology* 37 (1): 55–63.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Chowell, Diego, Sri Krishna, Pablo D. Becker, Clément Cocita, Jack Shu, Xuefang Tan, Philip D. Greenberg, Linda S. Klavinskis, Joseph N. Blattman, and Karen S. Anderson. 2015. “TCR Contact Residue Hydrophobicity Is a Hallmark of Immunogenic CD8+ T Cell Epitopes.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (14): E1754–62.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. 2015. “Sequenza: Allele-Specific Copy Number and Mutation Profiles from Tumor Sequencing Data.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 26 (1): 64–70.

Hunt, D. F., R. A. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. L. Cox, E. Appella, and V. H. Engelhard. 1992. “Characterization of Peptides Bound to the Class I MHC Molecule HLA-A2.1 by Mass Spectrometry.” *Science*.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Löffler, Markus W., Daniel J. Kowalewski, Linus Backert, Jörg Bernhardt, Patrick Adam, Heiko Schuster, Florian Dengler, et al. 2018. “Mapping the HLA Ligandome of Colorectal Cancer Reveals an Imprint of Malignant Cell Transformation.” *Cancer Research* 78 (16): 4627–41.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Magadan, Susana, Aleksei Krasnov, Saida Hadi-Saljoqi, Sergey Afanasyev, Stanislas Mondot, Delphine Lallias, Rosario Castro, et al. 2019. “Standardized IMGT® Nomenclature of Salmonidae IGH Genes, the Paradigm of Atlantic Salmon and Rainbow Trout: From Genomics to Repertoires.” *Frontiers in Immunology* 10 (November): 2541.

McGranahan, Nicholas, Rachel Rosenthal, Crispin T. Hiley, Andrew J. Rowan, Thomas B. K. Watkins, Gareth A. Wilson, Nicolai J. Birkbak, et al. 2017. “Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution.” *Cell* 171 (6): 1259–71.e11.

Nielsen, Morten, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Røder, et al. 2007. “NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known.” *PloS One* 2: 796.

Perez-Riverol, Yasset, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J. Kundu, Avinash Inuganti, et al. 2019. “The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data.” *Nucleic Acids Research* 47 (D1): D442–50.

Ramesh, Adithya, and Ian Wheeldon. 2021. “Guide RNA Design for Genome-Wide CRISPR Screens in Yarrowia Lipolytica.” *Methods in Molecular Biology* 2307: 123–37.

Rammensee, H., J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović. 1999. “SYFPEITHI: Database for MHC Ligands and Peptide Motifs.” *Immunogenetics* 50 (3-4): 213–19.

Rammensee, H. G., T. Friede, and S. Stevanovíic. 1995. “MHC Ligands and Peptide Motifs: First Listing.” *Immunogenetics* 41 (4): 178–228.

Sarkizova, Siranush, Susan Klaeger, Phuong M. Le, Letitia W. Li, Giacomo Oliveira, Hasmik Keshishian, Christina R. Hartigan, et al. 2020. “A Large Peptidome Dataset Improves HLA Class I Epitope Prediction across Most of the Human Population.” *Nature Biotechnology* 38 (2): 199–209.

Schuster, Heiko, Janet K. Peper, Hans-Christian Bösmüller, Kevin Röhle, Linus Backert, Tatjana Bilich, Britta Ney, et al. 2017. "The Immunopeptidomic Landscape of Ovarian Carcinomas." *Proceedings of the National Academy of Sciences of the United States of America* 114 (46): E9942–51.

Shukla, Sachet A., Michael S. Rooney, Mohini Rajasagi, Grace Tiao, Philip M. Dixon, Michael S. Lawrence, Jonathan Stevens, et al. 2015. "Comprehensive Analysis of Cancer-Associated Somatic Mutations in Class I HLA Genes." *Nature Biotechnology* 33 (11): 1152–58.

Vita, Randi, James A. Overton, Jason A. Greenbaum, Julia Ponomarenko, Jason D. Clark, Jason R. Cantrell, Daniel K. Wheeler, et al. 2015. "The Immune Epitope Database (IEDB) 3.0." *Nucleic Acids Research* 43 (Database issue): D405–12.

Zhang, Jing, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A. Lajoie, and Bin Ma. 2012. "PEAKS DB:De NovoSequencing Assisted Database Search for Sensitive and Accurate Peptide Identification." *Molecular & Cellular Proteomics* 11 (4): M111.010587.