Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

Priority Brief

Cancer Immunology Research

NetH2pan: A Computational Tool to Guide MHC Peptide Prediction on Murine Tumors

Christa I. DeVette[1], Massimo Andreatta[2], Wilfried Bardet[1], Steven J. Cate[1], Vanessa I. Jurtz[3], Kenneth W. Jackson[1], Alana L. Welm[4], Morten Nielsen[2,3], and William H. Hildebrand[1]

Abstract

With the advancement of personalized cancer immunotherapies, new tools are needed to identify tumor antigens and evaluate T-cell responses in model systems, specifically those that exhibit clinically relevant tumor progression. Key transgenic mouse models of breast cancer are generated and maintained on the FVB genetic background, and one such model is the mouse mammary tumor virus-polyomavirus middle T antigen (MMTV-PyMT) mouse—an immunocompetent transgenic mouse that exhibits spontaneous mammary tumor development and metastasis with high penetrance. Backcrossing the MMTV-PyMT mouse from the FVB strain onto a C57BL/6 genetic background, in order to leverage well-developed C57BL/6 immunologic tools, results in delayed tumor development and variable metastatic phenotypes. Therefore, we initiated characterization of the FVB MHC class I *H-2*^{q} haplotype to establish useful immunologic tools for evaluating antigen specificity in the murine FVB strain. Our study provides the first detailed molecular and immunoproteomic characterization of the FVB *H-2*^{q} MHC class I alleles, including >8,500 unique peptide ligands, a multiallele murine MHC peptide prediction tool, and *in vivo* validation of these data using MMTV-PyMT primary tumors. This work allows researchers to rapidly predict H-2 peptide ligands for immune testing, including, but not limited to, the MMTV-PyMT model for metastatic breast cancer. *Cancer Immunol Res; 6(6); 636–44. ©2018 AACR.*

Introduction

The success of cancer immunotherapies has shed considerable light on the ability of the host immune system to survey and eliminate aberrant tumor cells. Pivotal to such therapies is the ability of the immune system to specifically recognize the tumor, a process that occurs at the immunologic synapse when the T-cell receptor (TCR) of cytotoxic T lymphocytes (CTLs) binds to tumor-specific intracellular peptides presented by major histocompatibility complex (MHC) class I molecules. Such peptides can be derived from pathogen, host, or mutated-self proteins, the latter being classified as tumor neoantigens. Recognition of these peptide–MHC (pMHC) complexes by TCRs is a prerequisite for T-cell activation and tumor cell elimination. This mechanism is so critical to antitumor immunity that successful CTL responses against tumor neoantigens can induce long-term remission in a subset of patients with melanoma (1, 2). Unfortunately, many cancers, including breast cancers, have low mutation burdens and/or "cold" immunologic microenvironments and, so far, have responded poorly to immunotherapy (3). Thus, advanced tools, such as immune-competent mouse models of breast cancer, are needed for studying how antigen-specific CTL responses are regulated in breast and other cancers.

The mouse mammary tumor virus-polyomavirus middle T antigen (MMTV-PyMT) transgenic mouse is widely used for studying breast cancer because tumors arise with high penetrance, spontaneously metastasize to the lungs, and exhibit clinically relevant histology in the FVB/NJ background (4). Other well-established breast cancer models (MMTV-Neu, MMTV-c-Myc, MMTV-Wnt1) are also found in the FVB strain, in large part due to the technical feasibility of introducing transgenes into this strain (5). Given the centrality of the FVB mouse to *in vivo* breast cancer studies, the ability to track antigen-specific T-cell responses would be transformative to the utility of these murine models. Here, we address this issue by combining proteomics and *in silico* MHC immunology technology to develop a tool that allows accurate prediction of tumor peptides presented on MHC class I in the FVB mouse.

Peptides that are recognized by T cells, termed T-cell epitopes, vary depending on which class I MHC molecule presents the peptide. Murine strains possess different class I haplotypes with different peptide binding preferences (6). In mice, the class I MHC α-chain proteins are encoded by *H-2K* and *H-2D*. In FVB and BALB/c mouse strains, a duplication event of the *H-2D* locus spawned the pseudogene *H-2L*, although the physiologic relevance of this duplication remains unknown (7). Because of this, we have focused our efforts on H-2K and H-2D peptide prediction. The FVB strain expresses the *H-2*^{q} haplotype and, despite the preponderance of FVB mice in disease models, the MHC peptide binding motifs of the H-2^q molecules are understudied (8). To

[1]University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma. [2]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina. [3]Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark. [4]Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah.

Note: Supplementary data for this article are available at Cancer Immunology Research Online (http://cancerimmunolres.aacrjournals.org/).

Corresponding Author: William H. Hildebrand, University of Oklahoma Health Sciences Center, 975 NE 10th Street, BRC-317, Oklahoma City, OK 73104. Phone: 405-271-1203; Fax: 405-271-3117; E-mail: william-hildebrand@ouhsc.edu

doi: 10.1158/2326-6066.CIR-17-0298

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

address this paucity of data and the difficulties in studying immune responses in the FVB strain, we DNA sequenced the cDNA forms of *H-2D*^{q} and *H-2K*^{q} gene transcripts and transfected *H-2D*^{q} and *H-2K*^{q} gene constructs to facilitate high-throughput proteomic identification of >8,500 peptides presented by MHC molecules of the *H-2*^{q} haplotype. These data allowed us to develop an online mouse H-2 pan-specific prediction tool—that is, a prediction tool encompassing the *H-2*^{b}, *H-2*^{d}, *H-2*^{k}, and *H-2*^{q} haplotypes.

A pan-specific neural network method that uses pooled data from multiple MHC molecules, as opposed to a single allele, greatly improves predictive power and accuracy (9). We have also shown that a model integrating peptide binding affinity data with eluted ligand data achieves highly improved predictive performance when it comes to the identification of eluted ligands and T-cell epitopes (10). Thus, the H-2D^q and H-2K^q peptide elution data were combined with peptide binding and eluted ligand data from other murine haplotypes (*H-2*^{b}, *H-2*^{d}, *H-2*^{k}). The resulting prediction tool is tailored to the "q" haplotype but allows cancer immunobiologists to accurately predict peptides of other murine class I haplotypes. We named this tool "NetH2pan," as an extension of our other published predictive tools (NetMHCpan). Prior to this, NetMHCpan was the only murine prediction tool available for the *H-2*^{q} haplotype. NetMHCpan, however, could only predict H-2^q ligands based on MHC sequence homology to other haplotypes. NetH2pan improves upon this tool by incorporating eluted peptide data from the *H-2*^{q} haplotype. To validate this pan-specific algorithm, we eluted >2,000 H-2^q peptides from class I MHC on MMTV-PyMT primary tumor cells, identifying peptides derived from 27 cancer-associated source proteins. We confirmed that the NetH2pan prediction tool successfully predicts cancer-associated tumor peptide ligands with high fidelity (top 1%), providing a significant improvement to current murine peptide prediction tools. The combined molecular, proteomic, and *in silico* MHC strategies described here empower a new generation of cancer immunotherapy research in murine models of breast cancer.

Materials and Methods

H-2K and *H-2D* sequencing of the FVB strain

High-resolution MHC typing was performed by Sequence Based Typing (SBT) at the University of Oklahoma Health Science Center CLIA/ASHI-accredited HLA typing laboratory on FVB mouse spleens. Extracted RNA (converted to cDNA) was amplified with two pairs of primers specific for the 5' and 3' UTRs of *H-2K* and *H-2LD* loci. *H-2L* and *H-2D* coamplification was dissected using locus-specific sequencing primers. Primers and nucleotide/protein sequences have been deposited to GenBank (accession numbers MF352192 and MF352193).

Cell lines, transfection, and production of MHC complexes for elution studies

HeLa cells were purchased from ATCC, immediately expanded, and frozen per manufacturer's instructions. 721.221 cells were a kind gift from Dr. Ted Hansen (Washington University, St. Louis). HeLa and 721.221 cells were authenticated by in-house HLA-typing and confirmed with known HLA types of the original cells. Cells were reauthenticated after transfection with soluble MHC (sMHC) constructs and prior to seeding in roller bottles. Soluble MHC constructs were generated as previously described with the addition of a very low-density lipoprotein receptor (VLDLr) purification tag (11). 721.221 and HeLa cells were stably transfected with the sMHC by nucleofection per the manufacturer's cell-line optimized protocols (nucleofection kits R and V, Lonza), G418 drug selection for 10 days, and subcloned by single-cell sorting. sMHC-producing clones were identified using a capture enzyme-linked immunosorbent assay (ELISA) with anti-VLDLr (CRL-2197 hybridoma, ATCC) as the capture antibody and anti-β_2-microglobulin (DAKO) as the detection antibody, and developed with o-phenylenediamine dihydrochloride (Sigma). Transfected cells were seeded into 48 roller bottles and sMHC-containing supernatant was collected. sMHC was purified from the supernatant using affinity chromatography with an antibody to VLDLr as published (12). Complexes were eluted from the column in 0.2 N acetic acid and immediately processed for isolation of the peptide ligands

Liquid chromatography–mass spectrometry (LC-MS) of H-2D^q and H-2K^q peptides

Peptide ligands were eluted as described previously (1, 11). Briefly, peptides were separated from purified MHC with acid boil followed by 3-kDa ultrafiltration (Merck Millipore). Peptides were fractionated with reverse phase-high performance liquid chromatography (RP-HPLC) and then analyzed by LC-MS. Nano-scale LC was performed with an Eksigent nano-LC-400 with an Eksigent autosampler (AB SCIEX). Fractions were combined with internal retention time (iRT, Biognosys) peptides before injection. Eluate was ionized with a NanoSpray III ion source (AB Sciex), and MS1 and MS2 fragment spectra were obtained in data-dependent acquisition (DDA) mode using a SCIEX TripleTOF 5600. Peptide sequences were obtained from spectra using PEAKS 8.0 (Bioinformatics Solutions) at a 5% FDR. Oxidation (M, H, W), deamidation (N, Q), sodium adducts (D, E, C-term), acetylation (N-term), pyro-glu from Q, and cysteinylation (C) were used as variable modifications. The UniProt database with *Homo Sapiens* or *Mus Musculus* taxonomy and iRT peptides were used as a reference library for fragments. Identified sequences have been deposited to the IEDB (submission IDs 1000724, 1000726, and 1000727, www.IEDB.org).

Training a pan-specific murine MHC ligand prediction method

Assembling a panel of peptide–H-2 binding and elution data. The amino-acid sequences of the H-2D^q and H-2K^q molecules were aligned to a reference database of MHC sequences to determine the pseudosequence of MHC residues in direct contact with the peptide, as described in detail previously (13). Both for H-2D^q and H-2K^q elution data sets, potential contaminants, and false positives (1%–5% of the ligands) were filtered out using GibbsCluster (14), applying the default parameters suggested for MHC I ligands of variable length. Binding affinity and ligand elution data for seven additional murine MHC molecules (H-2D^b, D^d, K^b, K^d, K^k, L^d, and L^q) were obtained from the IEDB (15), and consisted of 9,625 pMHC binding affinity measurements and 3,310 eluted ligands of length 8 to 11 residues. Binding affinity values *aff* in IC_{50} nmol/L were rescaled using the relationship $t = 1 - \log(aff)/\log(50{,}000)$, ensuring that target values fell between 0 and 1 (16). Because elution experiments only report positive peptides, negative instances must be generated artificially: for each H-2 molecule, the peptide length with the highest number N of observed ligands was determined, introducing then a flat distribution of 10 × N random natural peptides for each of the

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

lengths 8, 9, 10, and 11 as artificial negatives. Each training point is therefore represented by a triplet consisting of (i) the peptide sequence; (ii) the MHC pseudosequence associated with the peptide; and (iii) the target value, either as a rescaled binding affinity or as a binary value (one for observed ligand, zero for artificial negative).

Neural network training. A neural network ensemble was trained in 5-fold cross-validation as described (9), extending the architecture of the neural network with a second output neuron (10). This addition allows combining heterogeneous training data by utilizing the first output neuron (and its connections from the hidden layer) for binding affinity examples, and the second output neuron to predict ligands. Because the weights between input and hidden layers are shared between the two data types, motifs can be learned and reinforced within and between data types. All other parameters were consistent with the architecture of NetMHCpan-3.0: networks were initialized with 10 alternative random configurations of weights, using a single hidden layer composed of either 56 or 66 neurons and representing peptides and pseudosequences using BLOSUM encoding (9). Up to one insertion and two deletions were allowed to accommodate peptides of lengths 8 to 11 to a common alignment core of nine amino acids. Additional features encoded as input to the networks included the length of the insertion/deletion (if any) and the length of the peptide. The training set for H-2D^q included only the ligands derived from HeLa cells; ligands eluted from 721.221 cells were reserved as an independent set to evaluate the performance of the predictor.

Predictive performance was calculated in terms of area under the ROC curve (AUC) and in positive predictive value (PPV). In line with previous work, the predictions on the ligand data sets were compared against a decoy set of 999 natural random peptides for each positive instance, equally distributed across the four peptide lengths 8, 9, 10, and 11 (17). For cross-validation experiments, the partitions of positive examples were maintained unaltered, whereas negative examples were replaced with the 999 random decoys per positive example. The PPV was then calculated as the fraction of true positives among the top 0.1% ligands predicted by the method.

Length and motif preferences of H-2 molecules. For each MHC molecule included in the model, 400,000 random natural peptides (100,000 for each of the lengths 8, 9, 10, 11) were submitted to the neural networks and ranked by ligand prediction score. The relative frequency of each peptide length among the top 1% scoring peptides was then used to draw the ligand length profile characterizing each H-2 molecule. Similarly, the same top 1% scoring peptides were used to generate sequence motifs of the MHC molecules included in the model using the software Seq2Logo (18). For the analysis of the peptide cleavage preferences, ligands were mapped back to the nonredundant UniProtKB/Swiss-Prot database (19) to retrieve their source protein sequences. Enrichment scores M in the peptide flanking regions were then calculated as $M = \log_2(F_{i,A}/E_A)$, where $F_{i,A}$ is the frequency of amino acid A at position i, and E_A is the expected frequency for amino acid A calculated on all source proteins containing at least one ligand.

Mice

Female, 5-to-8-week-old FVB/NJ (Jackson Laboratories stock number 001800) and MMTV-PyMT (Jackson Laboratories stock number 002374) mice were purchased from The Jackson Laboratories. All animal work was in accordance with policies at the University of Utah, and all studies were approved by the University of Utah IACUC committee. MMTV-PyMT tumor cells were harvested from MMTV-PyMT transgenic mice, collagenase digested, and implanted into cleared mammary fat pads according to standard procedures (20). Tumors were measured weekly and size was calculated using length and width calliper measurements in the ellipsoid formula (Tumor volume = 1/2(length × width^2)). Mice were euthanized when tumors reached a maximum of 2 cm^3 and tumors were flash frozen in liquid nitrogen for peptide extraction.

Peptide identification from MMTV-PyMT tumors

Anti–H-2D^q (28-14-8S hybridoma, ATCC) and anti–H-2K^q (34-1-2S hybridoma, ATCC) were used to generate H-2^q immunoaffinity columns by coupling to CNBR-activated Sepharose 4 Fast Flow (GE Healthcare). Peptides were extracted based on a previously published protocol (Supplementary Fig. S1; ref. 21). Whole tumors were flash frozen in liquid nitrogen, cryogenically milled (MM400, RETSCH), suspended in lysis buffer containing octylphenoxy poly(ethyleneoxy)ethanol (IGEPAL; Sigma) and cOmplete EDTA-free protease inhibitor cocktails (Roche), and clarified by ultracentrifugation. Filtered lysate was passed over sequential anti-H-2D^q and H-2K^q columns. Peptides were eluted in acid and boiled, isolated by RP-HPLC, and analyzed by LC/MS using DDA mode for *de novo* peptide identification as described above. The UniProt database with *Mus Musculus* taxonomy was used as a reference library. Peptide results can be accessed through the IEDB (submission ID 1000724, http://www.iedb.org).

Results

Characterization of *H-2^q* haplotype

We began by sequencing the class I MHC *H-2^q* loci with high-resolution MHC typing. The sequencing strategy used here amplified the regions flanked by the 5' and 3' UTRs for *H-2K* and *H-2L/D* genes of the FVB/NJ strain. Locus-specific oligonucleotide primers were used to separate the *H-2D* sequence from *H-2L*. The amplification primers and nucleotide sequences for *H-2D^q* and *H-2K^q* have been deposited into GenBank (accession numbers MF352192 and MF352193). The *H-2L* locus, although expressed in the "q" haplotype, is present only in a few MHC haplotypes (6), and we therefore chose to characterize only *H-2D^q* and *H-2K^q*. The resulting sequences for *H-2^q* loci differed by 2 and 1 amino acids from previous reports for *H-2K^q* and *H-2D^q*, respectively, a finding that is expected given strain-to-strain variation and improved sequencing techniques. Having determined the sequence of *H-2D^q* and *H-2K^q*, we proceeded to identify peptides presented by the "q" haplotype.

Using the MHC class I *H-2K^q* and *H-2D^q* sequences, we designed sMHC constructs that lacked the transmembrane domain and incorporated a VLDLr purification tag. Engineered purification tags (e.g., VLDLr) eliminated the need for allele-specific antibodies (28-14-8S and 34-1-2S) that may cross-react with other alleles within the same haplotype, allowing isolation of only the peptide:MHC complexes of interest. For alleles in which the motif is not yet characterized, such an approach allowed the characterization of the correct motifs with high confidence. sMHC-bound peptidomes accurately represent the

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

full-length, membrane-bound peptidomes with comparable binding motifs, peptide lengths, predicted binding antigens, and putative source antigens (22). Based on these findings, sMHC is a valuable and physiologically relevant tool for studying the immune peptidome. Tagged, soluble forms of *H-2K*^{q} and *H-2D*^{q} were separately transfected into the MHC class I–negative 721.221 cell line, and high producing single-cell clones were obtained. Soluble *H-2D*^{q} was also transfected into HeLa cells to provide additional unique *H-2D*^{q} peptides. Multiple attempts for a high producing 721.221-H-2K^q were made, but protein expression dwindled with each expansion impeding soluble H-2K^q production in these cells. Peptide–MHC complexes from the supernatant of high-producing clones were purified by anti-VLDLr immunoaffinity columns, and peptides were liberated from the MHC by boiling in acidic conditions. Two-dimensional nano-LC-MS/MS yielded 8500 H-2D^q and 481 H-2K^q different peptides, permitting the visualization of H-2D^q and H-2K^q binding motifs, ligand length distributions, and source protein distribution (Fig. 1). These peptide sequences and their source proteins are available through the Immune Epitope Database (IEDB; submission IDs 1000726 and 1000727, http://www.iedb.org). The high number of unique H-2D^q peptides was due to higher H-2D^q expression in transfected 721.221 and HeLa cells than in *H-2K*^{q} transfected 721.221 cells alone. Physiologic differences in allele expression, peptide diversity, and binding motifs can all contribute to fewer H-2K^q ligands when compared with H-2D^q.

Analysis using Seq2Logo revealed an MHC binding motif for each allele (Fig. 1A and B; ref. 18). H-2D^q favored proline at P2 and hydrophobic residues at PΩ, whereas H-2K^q favored acidic P2 residues and a hydrophobic PΩ (Fig. 1A and B). The ligand length preference differed substantially between the two loci whereby H-2D preferred 9-mers and H-2K bound predominately 8-mers (Fig. 1C). These ligands were derived from 4,000 source proteins, and most source proteins provided a single peptide (Fig. 1D).

Generation of a prediction model for murine MHC class I

Pooling this large set of H-2D^q and H-2K^q ligands with publicly available H-2 ligand elution and binding affinity data, we generated a pan-specific prediction model applicable to nine different mouse MHC molecules. This combination of multiple molecules and data types produced a consistently high performance in terms of AUC for the prediction of binding affinity (BA) and ligand likelihood (EL), even for alleles for which only one data type was available (Table 1). Of particular interest for this study, the

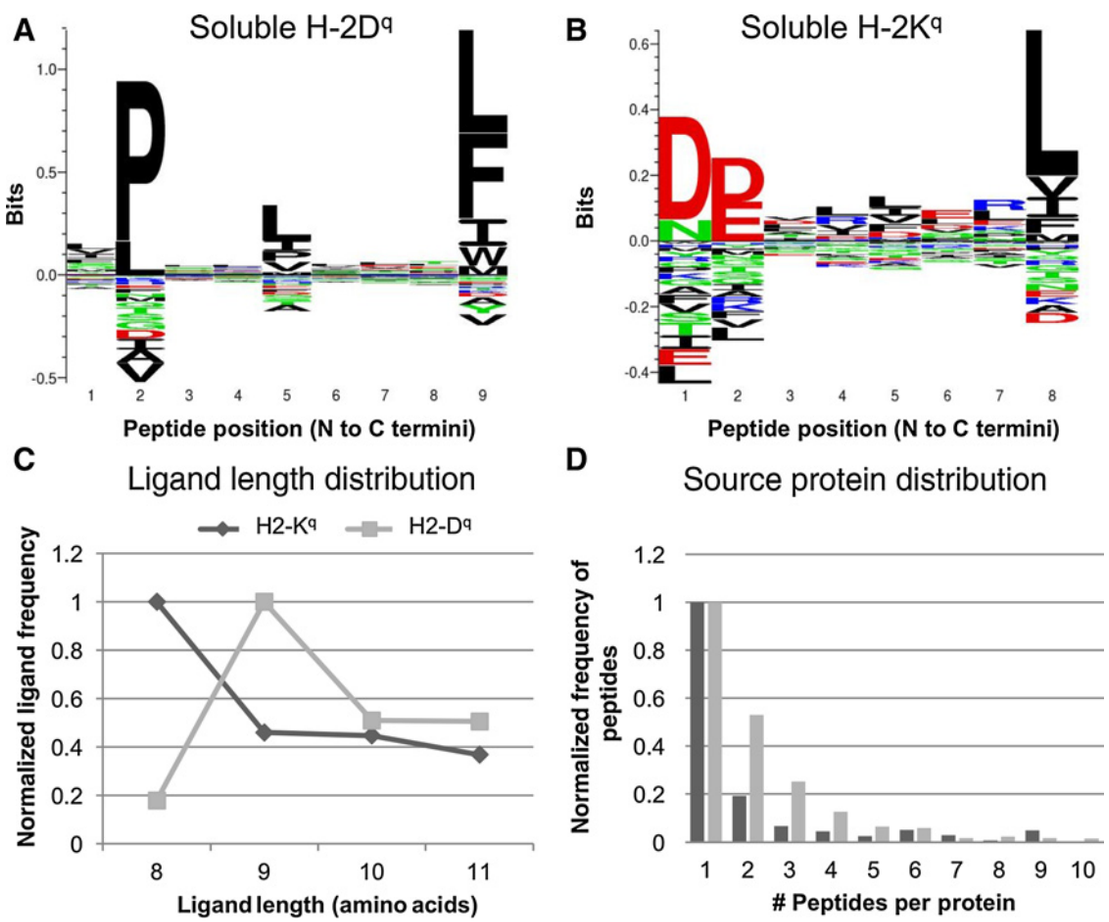

Figure 1.

Soluble MHC elution identifies H-2D^q and H-2K^q ligands. **A** and **B,** Peptide binding motif for H-2D^q and H-2K^q (derived from the subset of 9-mer and 8-mer ligands, respectively). **C,** Ligand length distribution (normalized to the most prevalent length: 8-mer for H-2K^q, 9-mer for H-2D^q. **D,** Source protein analysis of H-2^q peptides (normalized to the most frequent number of peptides per protein). Graphs **C–D** were generated with 3,500 721-H-2D^q ligands and 500 721-H-2K^q ligands.

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

Table 1. Cross-validated performance in AUC for prediction methods trained on binding affinity data only (BA), eluted ligands only (EL), and a combination of both data types (NetH2pan)

	Binding affinity data				Eluted ligands			
Allele	***N***	**Positive**	**BA**	**NetH2pan**	***N***	**Positive**	**EL**	**NetH2pan**
H-2D^b	3,775	730	0.943	0.945	22,616	806	0.981	0.981
H-2D^d	413	47	0.902	0.902	0	0	NA	NA
H-2D^q	0	0	NA	NA	83,107	4,107	0.995	0.993
H-2K^b	3,977	1,338	0.920	0.922	21,509	1,774	0.936	0.933
H-2K^d	843	295	0.859	0.858	24,148	730	0.968	0.968
H-2K^k	371	176	0.847	0.855	0	0	NA	NA
H-2K^q	0	0	NA	NA	8,121	361	0.930	0.855
H-2L^d	243	50	0.939	0.951	0	0	NA	NA
H-2L^q	3	2	NA	NA	0	0	NA	NA

NOTE: For binding affinity measurements, positive instances are defined as having an IC_{50} affinity < 500 nmol/L. For elution assays, the total number of data points (*N*) includes the positives (i.e., the observed ligands) and a uniform distribution of artificial negatives generated as described in Materials and Methods. This tool predicts two properties of each peptide: binding affinity or eluted ligand likelihood. For binding affinity evaluation, the binding affinity predictions were used, and for eluted ligands, the eluted ligand likelihood predictions.

predictor obtains a remarkably high cross-validated AUC = 0.993 for H-2D^q. An alternative useful performance metric is the PPV, which measures the fraction of true positives in the top P_p % predicted ligands. Because each MHC molecule is expected to present approximately 1 of 1,000 peptides in the proteome, the PPV was calculated on the top $P_p = 0.1\%$ predictions. The PPV for the five H-2 alleles characterized by ligand data varied between 35.7% and 61.1%, with an average of 49.8% (Supplementary Table S1; refs. 15, 17). The model maintains a high predictive performance on H-2D^q ligands eluted from another cell line (721.221) that were not included in the original training set (AUC = 0.944, PPV = 59.7%). Ligands detected in both the HeLa and 721.221 cell lines (see Supplementary Fig. S2) were excluded from this validation. When both affinity and ligand elution data were available for a given allele, the combined model did not benefit from the added binding affinity data to predict eluted ligands, compared with a model only trained on eluted ligands (Table 1). However, in cases in which only binding affinity or eluted ligand data were available for a given allele, the combined approach had the advantage of complementing one data type with the other, using a single, unified model. The prediction tool can be accessed at http://www.cbs.dtu.dk/services/NetMHCpan-4.0/NetH2pan/.

The eluted and predicted binding motifs for H-2K^q and H-2D^q both show a preference for hydrophobic residues in the PΩ F pocket (Fig. 1A and B; Fig. 2A and B). This hydrophobic amino acid preference is highly conserved amongst murine and human MHC and is essential for stability within the MHC class I–binding cleft (23). The second anchor for H-2D^q is a proline at the P2 B pocket. H-2K^q, on the other hand, favors peptides with acidic residues at the P2 anchor. The preference for acidic residues at P2 could not be captured by networks trained without H-2K^q eluted ligands. Despite the ability of pan-specific methods to infer binding motifs by similarity to other alleles, dissimilar molecules can be hard to predict. As a rule of thumb, the pan-specific method described here performed accurately for alleles with a distance D (measured in terms of the pseudosequence similarity) to the training data lower than 0.1 (24). In the absence of H-2K^q training data, the minimum distance for H-2K^q was D = 0.33, with H-2K^k being the closest allele in the training data. This observation underscores the importance of collecting experimental elution data across the whole MHC sequence space. Peptides with acidic anchors have limited diversity, which might also explain the low yield of H-2K^q peptides in our experiments.

An important feature of neural networks trained on peptides of variable length is their ability to assimilate the ligand length preferences of MHC molecules (25). Accounting for MHC-dependent length preferences has important implications when scanning for potential ligands in epitope discovery. In this respect, the model shows that H-2K^q has a marked preference for 8-mers (Fig. 2C), reflecting the distribution directly observed in the elution experiments (Fig. 1C). This tendency is similar to the length preference for H-2K^b, both as predicted by our model and as described in the literature (26). H-2D^q has a more canonical peptide length distribution, centered on 9-mers and to a lesser extent on 10-mers and 11-mers (Fig. 1C, 2C). These data confirm that the prediction tools accurately represent the eluted peptide data and that this pan-specific approach can identify subtle differences in length and anchor preferences of the murine MHC I.

Cytoplasmic proteins targeted for degradation undergo antigen processing (defined by proteasomal cleavage, trimming by cytosolic peptidases, and N-terminal processing by endoplasmic reticulum aminopeptidases). Given this, we hypothesized that certain residues would be favored in the source protein regions directly surrounding the MHC-presented ligands. Analysis of flanking source protein sequences revealed a preference for methionine (M) at N-1, whereas proline (P) was disfavored at both termini (N − 1, C + 1; Fig. 2D). Positively charged residues (R, K) appeared to be enriched at C + 1, C + 2, whereas negatively charged residues (D, E) were depleted in these positions. The source proteins of H-2 ligands extracted from the IEDB (Fig. 2E, element-wise correlation PCC = 0.559) and in previously reported peptide cleavage signatures (17) had similar patterns. To test whether these cleavability preferences could improve the identification of H-2 ligands, we retrained NetH2pan, including the 5-residue regions flanking the ligands as additional input features. We did not observe a significant impact of cleavability preferences in cross-validated performance, with average AUC increasing slightly from 0.945 (model without flanking sequences) to 0.947 (with flanks), and PPV decreasing from 49.8% to 49.1% (Supplementary Table S2). On the H-2D^q evaluation set from 721.221 cells, AUC increased from 0.944 to 0.945, and PPV from 59.7% to 60.5%. Although signatures of peptide processing were found consistently across data sets and MHC alleles, their contribution thus did not appear to be beneficial to improving ligand prediction.

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

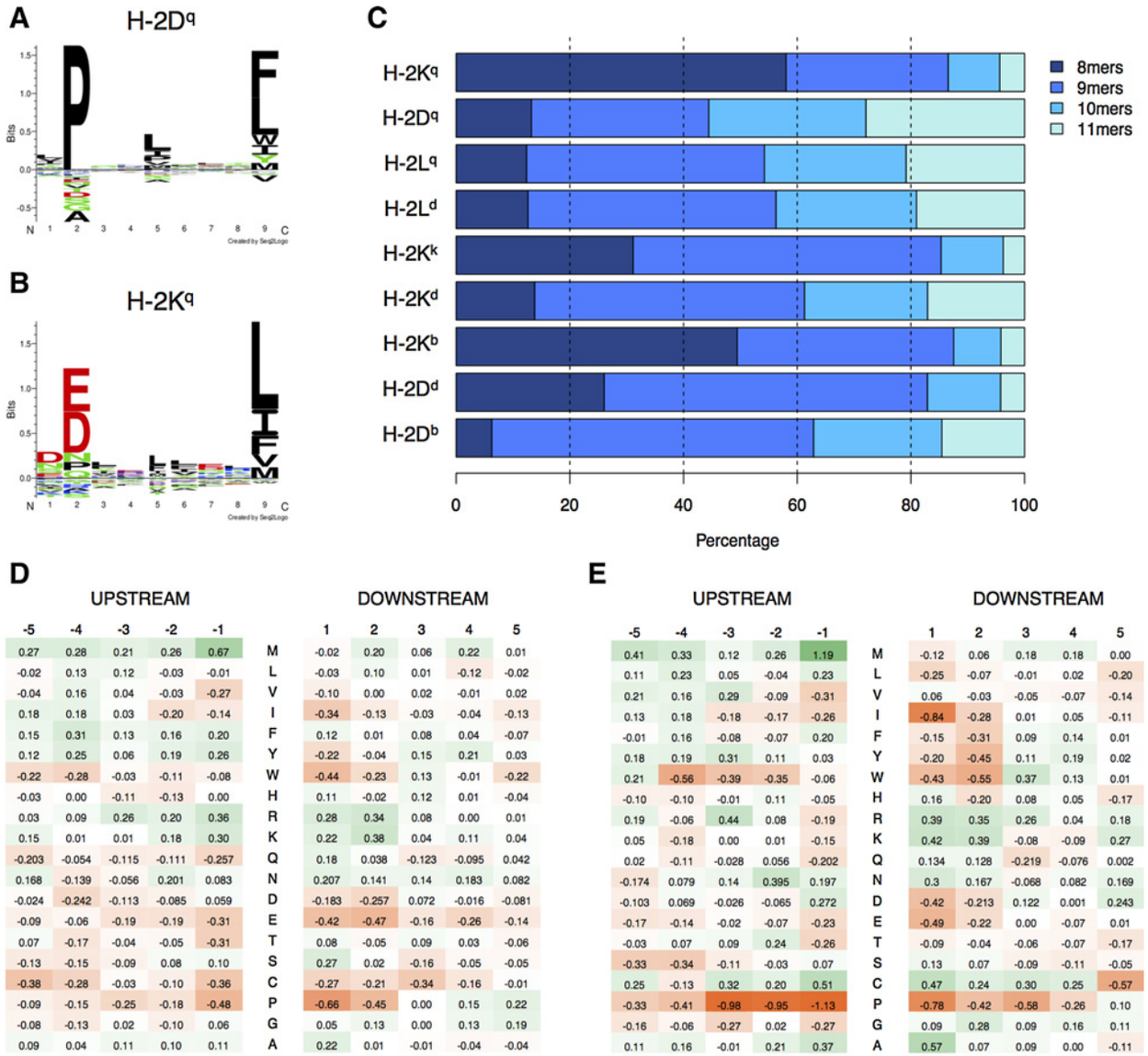

Figure 2.

Binding motifs predicted by the neural network model. **A,** H-2D^q shows a strong preference for proline at P2 and enrichment of hydrophobic/aromatic amino acids at PΩ. **B,** H-2K^q requires acidic residues at P2 and hydrophobic amino acids at PΩ. **C,** Predicted ligand length preferences for the nine H-2 alleles included in the model, calculated from a large set of random natural peptides. **D–E,** Amino-acid enrichment in the source protein regions flanking the ligands, derived from soluble MHC ligands generated in this study (**D**) and from H-2 ligands extracted from the IEDB (**E**). Values in the heat maps are $M = \log_2(F/E)$, where F is the observed frequency and E is the expected frequency in the source proteins containing ligands. The element-wise correlation between the heat maps **D** and **E** is PCC = 0.559.

Tumor antigen discovery validation of the NetH2pan algorithm

Finally, we sought to validate the NetH2pan prediction tool on ligands directly eluted from primary murine tumors as a complementary method to eliminate any bias introduced by the sMHC approach. The ability of NetH2pan to predict peptides presented on full-length H-2D^q and H-2K^q via physiologic antigen processing would provide important validation of this tool. Because aberrancies to MHC class I presentation vary among tumor models, and functions as a mechanism of tumor immune evasion, we confirmed class I MHC expression on >60% of EpCAM$^+$ primary tumor cells using flow cytometry (Supplementary Fig. S3). These percentages correlate with global MHC I transcript expression in healthy mammary tissue and secondary lymphoid organs in mice (27). Compared with splenocytes, tumor cells exhibited heterogeneous MHC class I protein expression (Supplementary Fig. S3E). We next purified class I MHC from the 5 g of murine tumors; we were able to purify >2,000 H-2^q ligands from MMTV-PyMT tumors. These findings suggest that the ability for MMTV-PyMT tumors to thrive in the FVB mouse does not result from a lack of peptide presentation, but likely other T-cell–dampening mechanisms that require further

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

exploration. Therefore, neoantigen discovery paired with peptide prediction remains highly relevant.

Peptide:MHC complexes were extracted from MMTV-PyMT tumors pooled from seven mice, based on a previously published protocol (Supplementary Fig. S1; refs. 21). Whole tumors were flash frozen in liquid nitrogen, cryogenically milled, and suspended in lysis buffer. Lysates were clarified by ultracentrifugation and passed over sequential H-2D^q and H-2K^q columns. Peptides were eluted by acid boil, and the yield of MHC was quantified by the β-2-microglobulin peak in first dimension HPLC (pH 2.85). Peptides were fractionated and analyzed by LC/MS using DDA (H-2D^q and H-2K^q) for *de novo* peptide identification. Full-length purification relied on two H-2^q immunoaffinity columns (28-14-8S and 34-1-2S). Previous literature suggests that the 28-14-8S antibody can cross-react with both H-2D^q and H-2L^q. To prevent the coelution of H-2D^q and H-2L^q peptides from undermining our analysis, we used the sequence available in the public domain for *H-2L^q* (GenBank AAA39573.1) and compared it with the sequence obtained for *H-2D^q*. We indeed find differences in the residues contained in the H-2 peptide-binding domain. These differences suggest subtle variation in the predicted ligands of H-2D^q and H-2L^q, allowing eluted peptides to be sorted using the Gibbs Cluster analysis. Only the peptides matching the H-2D^q motif generated from the sMHC approach were used to validate NetH2pan. Additionally, we found that 34-1-2S also recognized H-2D^q, because LC/MS analysis solved H-2D^q peptides and H-2D^q heavy chains in the eluate. Peptides eluted from 34-1-2S columns were subjected to Gibbs Cluster analysis as well. Because of the cross-reactivity and the strict H-2K^q motif, 40-fold fewer H-2K^q peptides were identified when compared with H-2D^q (Supplementary Table S3).

The list of tumor-presented peptides is accessible through IEDB, and specific MS/MS spectra are available upon request (submission ID 1000724, http://www.iedb.org). Using the H-2^q peptides extracted from PyMT tumors, two dominant motifs emerged (Fig. 3A and B) that were nearly identical to the sMHC eluted and predicted motifs (Figs. 1 and 2). Of the 2,000 tumor-eluted peptides, 27 cancer-associated source proteins were identified and used to test the prediction performance of NetH2pan versus NetMHCpan-3.0 (Fig. 3C; Supplementary Table S4). These 27 source proteins include oncogenes, tumor suppressors, and common biomarkers, but are not canonical

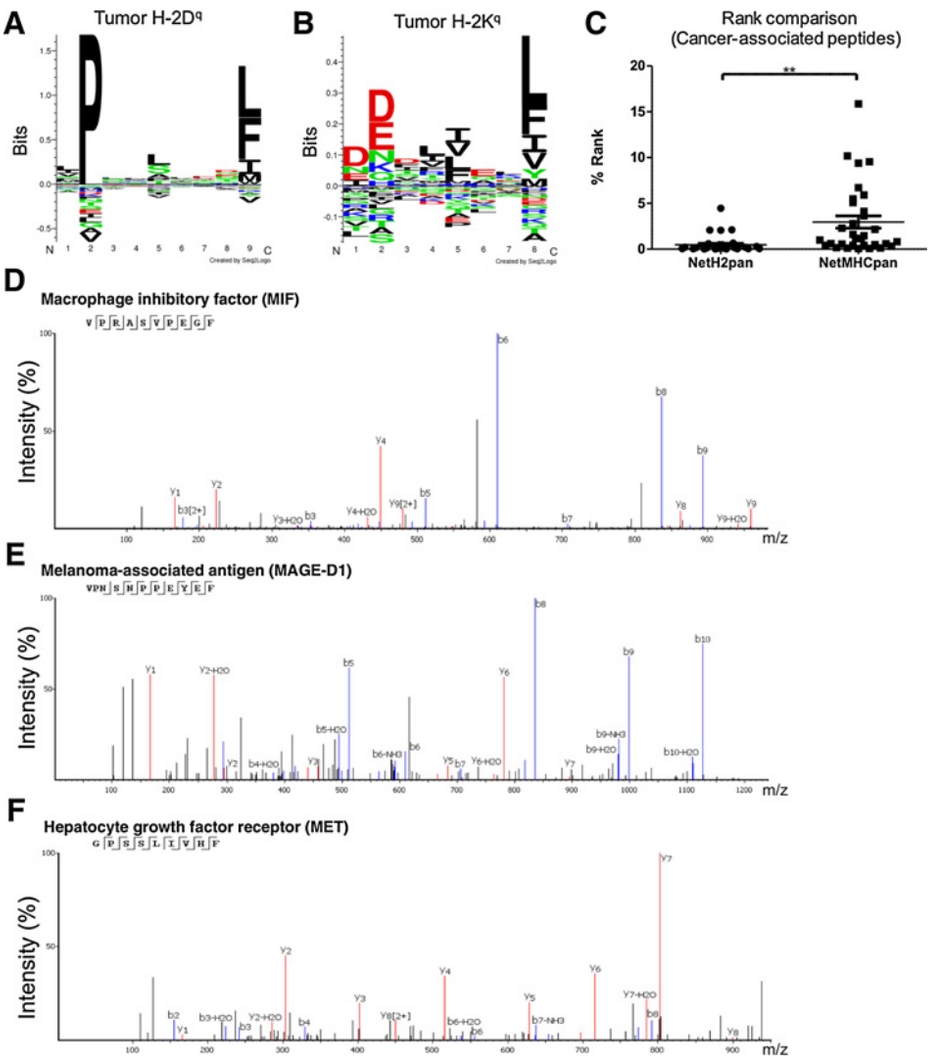

Figure 3.
H-2^q peptides from MMTV-PyMT tumors validate prediction methods. **A–B,** H-2^q binding motifs for peptides directly eluted from tumors, **C,** Comparison of predictions for presented cancer-associated peptides with NetMHCpan; **, $P = 0.0013$ (binomial test). **D–F,** Select H-2D^q peptide MS/MS spectra of cancer-associated source proteins eluted from tumors: macrophage inhibitory factor (MIF), epithelial cell adhesion molecule (EpCAM), and hepatocyte growth factor receptor (MET) oncogene.

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

"tumor antigens." The spectra of 3 (of 27) cancer-associated peptides, macrophage inhibitory factor (MIF), epithelial cell adhesion molecule (EpCAM), and hepatocyte growth factor receptor (MET), are shown to illustrate the unambiguous proteomic data obtained directly from tumors (Fig. 3D–F). The source protein sequences were subjected to NetH2pan and NetMHCpan-3.0 peptide predictions, then the identified eluted peptide was compared with the resulting predicted peptides. Pooling all sequences of the cancer-associated proteins containing at least one H-2D^q ligand and digesting them in all possible 8- to 11-mers resulted in a total of 86,340 peptides. The 32 observed H-2D^q ligands constituted 0.04% of the total pool of 86,340 potential ligands, a number not too distant from the 0.1% estimate assumed in earlier studies (17, 28). Among the top 32 ligands predicted by NetH2pan, 11 were true positives (PPV = 34.4%). In contrast, none of the top 32 ligands predicted by NetMHCpan-3.0 were true positives. In sum, NetH2pan outperformed NetMHCpan on the peptide predictions, providing improved predicted binding affinities (rank scores; Fig. 3C). Of the 33 peptides, 32 were predicted within the top 3%, with the median absolute rank being 3. In comparison, NetMHCpan predictions had a median absolute rank of 38. This means that when using NetH2pan, researchers are likely to find the true ligands by synthesizing only 3 predicted peptides, in contrast to the 38 required using the earlier NetMHCpan tool.

Discussion

In summary, we sequenced the MHC class I loci of FVB mice, produced and isolated their MHC class I proteins, and generated an MHC class I binding motif for the FVB "q" haplotype. Over 8,500 ligands were eluted, allowing us to design a prediction tool for antigens of interest. This tool, "NetH2pan," improves on previous "NetMHCpan" methods by utilizing both binding affinity and elution mouse H2 data to generate its predictions, representing a sophisticated peptide modeling tool for mice. To confirm this, peptides directly identified on MMTV-PyMT primary tumors served as a validation of the predictive power of NetH2pan. This immunopeptidomics study represents the first MHC peptide characterization and prediction for the "q" haplotype and for MMTV-PyMT tumors. Our data also facilitate improved predictions of peptide presentation in other commonly used murine strains, including C57/BL6 and BALB/c.

Antigen-specific T-cell responses have been difficult to study in breast cancer because the antigens are poorly characterized in mouse models with high penetrance of breast cancer. Our goal was to enable studies of antigen-specific responses in models of spontaneous tumor development and metastasis, which are often on the FVB background. Here, the identification of ligands directly eluted from tumors identifies MMTV-PyMT tumor-presented peptides, although immunogenicity studies must be undertaken to validate ligands as T-cell targets. This approach is not confined to a single antigen or murine strain, as we have provided a method improved from the NetMHCpan tool to predict peptides across several MHC class I haplotypes. In this way, NetH2pan will be particularly useful for cancer neoepitope prediction from tumor-specific mutations identified via whole-exome-sequencing (WES) data in C57/BL6 tumor models, as the H-2^b predictions are now more likely to generate truly presented peptides (29). This WES-tumor neoantigen-peptide prediction pipeline has been adopted and successfully used to treat human melanoma patients (1). When used appropriately and in combination with WES data from tumors, NetH2pan can predict CD8 T-cell immune epitopes with fewer false positives. More work must be done to characterize which of these peptides are unique to the tumor and capable of activating CTLs. These studies are now feasible in more murine cancer models. The prediction model developed in this study is publicly available online as a webserver at http://www.cbs.dtu.dk/services/NetMHCpan-4.0/NetH2pan.

Disclosure of Potential Conflicts of Interest

M. Nielsen is a consultant/advisory board member for Sir Sciences. W.H. Hildebrand is chief scientist at, has ownership interest in, and is a consultant/advisory board member for Pure MHC. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: C.I. DeVette, V.I. Jurtz, A.L. Welm, M. Nielsen, W.H. Hildebrand

Development of methodology: C.I. DeVette, M. Andreatta, S.J. Cate, M. Nielsen

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C.I. DeVette, W. Bardet, K.W. Jackson, A.L. Welm

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): C.I. DeVette, M. Andreatta, W. Bardet, M. Nielsen

Writing, review, and/or revision of the manuscript: C.I. DeVette, M. Andreatta, V.I. Jurtz, A.L. Welm, M. Nielsen

Acknowledgments

This work was supported by funding from a Department of Defense Breast Cancer Research Program Innovator and Scholar Concept Award (W81XWH-12-1-0499 to A.L. Welm), a Susan G. Komen Leadership Award (SAC160078 to A.L. Welm), and an NIH NRSA NIAID Training Grant (T32AI007633 to C.I. DeVette).

We would like to thank Ismail Can and Atakan Ekiz for providing MMTV-PyMT tumors, Curtis McMurtrey for excellent technical advice, Saghar Kaabinejadian for critical review, Sean Osborn for HLA-typing, and the Flow Cytometry Laboratory at the University of Oklahoma Health Sciences Center.

Received June 21, 2017; revised January 12, 2018; accepted March 27, 2018; published first April 3, 2018.

References

1. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science 2015;348:803–8.
2. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med 2012;366:2443–54.
3. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
4. Guy CT, Cardiff RD, Muller WJ. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. Mol Cell Biol 1992;12:954–61.
5. Taneja P, Frazier DP, Kendig RD, Maglic D, Sugiyama T, Kai F, et al. MMTV mouse models and the diagnostic values of MMTV-like sequences in human breast cancer. Expert Rev Mol Diagn 2009;9:423–40.
6. Lee DR, Rubocki RJ, Lie WR, Hansen TH. The murine MHC class I genes, H-2Dq and H-2Lq, are strikingly homologous to each other, H-2Ld, and two genes reported to encode tumor-specific antigens. J Exp Med 1988;168:1719–39.

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

7. Rubocki RJ, Lee DR, Lie WR, Myers NB, Hansen TH. Molecular evidence that the H-2D and H-2L genes arose by duplication. Differences between the evolution of the class I genes in mice and humans. J Exp Med 1990;171: 2043–61.
8. Taketo M, Schroeder AC, Mobraaten LE, Gunning KB, Hanten G, Fox RR, et al. FVB/N: an inbred mouse strain preferable for transgenic analyses. Proc Natl Acad Sci USA 1991;88:2065–9.
9. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med 2016;8:33.
10. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immunol 2017;199: 3360–8.
11. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, et al. The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. J Immunol 2016;196:1480–7.
12. Yaciuk JC, Skaley M, Bardet W, Schafer F, Mojsilovic D, Cate S, et al. Direct interrogation of viral peptides presented by the class I HLA of HIV-Infected T cells. J Virol 2014;88:12992–3004.
13. Nielsen M, Dimitrov I, Flower DR, Doytchinova I. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS ONE 2007;2.
14. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res 2017;45: W458–63.
15. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res 2015;43: D405–12.
16. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 2003;12:1007–17.
17. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass spectrometry profiling of HLA-associated peptidomes in monoallelic cells enables more accurate epitope prediction. Immunity 2017;46: 315–26.
18. Thomsen MC, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res 2012;40:W281–7.
19. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-prot, the manually annotated section of the uniprot knowledgeBase: how to use the entry view. Methods Mol Biol 2016;1374: 23–54.
20. Eyob H, Ekiz HA, Derose YS, Waltz SE, Williams MA, Welm AL. Inhibition of Ron kinase blocks conversion of micrometastases to overt metastases by boosting antitumor immunity. Cancer Discov 2013;3:751–60.
21. Purcell AW. Isolation and characterization of naturally processed MHC-bound peptides from the surface of antigen-presenting cells, in HPLC of peptides and proteins: methods and protocols, Aguilar M.-I., Editor. 2004, Springer, New York; Totowa, NJ. p. 291–306.
22. Scull KE, Dudek NL, Corbett AJ, Ramarathinam SH, Gorasia DG, Williamson NA, et al. Secreted HLA recapitulates the immunopeptidome and allows in-depth coverage of HLA A*02:01 ligands. Mol Immunol 2012; 51:136–42.
23. Hunt D, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, et al. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science 1992;255:1261–3.
24. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics 2012;64.
25. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics 2016; 32:511–7.
26. Deres K, Schumacher TN, Wiesmüller KH, Stevanović S, Greiner G, Jung G, et al. Preferred size of peptides that bind to H-2 Kb is sequence dependent. Eur J Immunol 1992;22.
27. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 2014;515: 355–64.
28. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics 2015;14:658–73.
29. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature 2014;515:577–81.

Published OnlineFirst April 3, 2018; DOI: 10.1158/2326-6066.CIR-17-0298

Cancer Immunology Research

NetH2pan: A Computational Tool to Guide MHC Peptide Prediction on Murine Tumors

Christa I. DeVette, Massimo Andreatta, Wilfried Bardet, et al.

Cancer Immunol Res 2018;6:636-644. Published OnlineFirst April 3, 2018.