

## **Supplemental information:**

Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction

## **Supplemental Materials**

### **Supplemental Figure Legends**

**Fig. S1, related to Figure 1:** Associated material for main text Fig 1.

**Fig. S2, related to Figure 2:** Associated material for main text Fig 2.

**Fig. S3, related to Figure 3:** Associated material for main text Fig 3.

**Fig. S4, related to Figure 4:** Associated material for main text Fig 4.

**Fig. S5, related to Figure 5:** Associated material for main text Fig 5.

### **Supplemental Tables**

**Table S1, related to Figure 1:** *Tab A.* Negative Control Peptides; *Tab B.* Master List Controls Removed;

*Tab C.* Cysteinylated Peptides

**Table S2, related to Figure 2:** *Tab A.* Sample Summary; *Tab B.* Raw Files

**Table S3, related to Figure 4:** Nested sets of peptides

**Table S4, related to Figure 5:** *Tab A.* Positive predictive value (PPV) calculations from ASROCL Model, per allele; *Tab B.* Internal Evaluation on the LC-MS/MS dataset; *Tab C.* PPV and AUC evaluation results on DFRMLI competition data; *Tab D.* External HIV epitope dataset. *Tab E.* PPV and AUC evaluation results on T cell response data; *Tab F.* PPV and AUC evaluation results on Mann and Trolle data sets.

### **Supplemental Experimental Procedures**

**Figure S1, related to Figure 1:** **a)** The number of peptide spectrum matches (PSMs) identified from both the no enzyme and *HLA*-specific rounds of database searches are shown for each *HLA* allele dataset. These PSMs represent the unique peptide identifications reported in Table 2. **b)** *HLA* cell surface presentation of single-*HLA* cell lines were compared to primary lymphocytes using FACS analysis. Cell lines that resulted in high (top; *HLA*-A\*02:01, -A\*02:07) and low (bottom; *HLA*-A\*31:01, -B\*35:01) numbers of HLA-associated peptides identifications by LC-MS/MS are shown. The number of total LC-MS/MS peptide identifications correlates with total cell surface HLA presentation. **c)** Comparison of the number of overlapping and unique peptides between our LC-MS/MS dataset and IEDB (all lengths). **d)** The overlap of unique peptides identified from biological replicates of our LC-MS/MS data (orange) and published data (purple)(Bassani-Sternberg et al., 2015) generated from immunopurifications of *HLA*-A\*02:01 expressing cells. Unique peptide overlap between our *HLA*-A\*02:01 dataset and this published dataset is also shown. **e)** The distribution of peptide modifications represented by the “Modified peptides” category in Figure 1 is shown as a pie chart. Peptide modifications included oxidized Met (m), deamidation (n), N-term Pyroglutamate (q), phosphorylation (sty), and cysteinylolation (c). **f)** Comparison of amino acid frequency between peptides in IEDB and peptides and peptides in the(Bassani-Sternberg et al., 2015; Trolle et al., 2016) respectively. To avoid biases due to anchor residues, for each comparison, 300 peptides per allele were selected at random for the alleles in the corresponding data set (Trolle: A\*01:01, A\*02:01, A\*24:02, B\*51:01; Mann: A\*01:01, A\*02:01, A\*03:01, A\*24:02, A\*31:01, B\*51:01) and pooled together before amino acid frequency was calculated. Amino acid frequencies are similar to those observed for our data (**Fig. 1f**). See also **Figure 1**.

**Figure S2, related to Figure 2:** **a.** Sequence logos generated using 9mer data for the 16 *HLA* alleles characterized by LC-MS/MS and all 9mers with quantitative measurement in IEDB for the same alleles. **b.** Individual allele entropy calculations for each amino acid positions within 9mer peptides sequenced by LC-MS/MS (entropy is normalized by  $\log(20)$  and shown on to [0,1] scale). **c.** Summary plot of entropy per 9mer position across all *HLA* alleles in LC-MS/MS (orange) and IEDB (blue) datasets (entropy is normalized by  $\log(20)$  and shown on to [0,1] scale). **d.** NMDS plots showing HLA-associated 9mer peptide clustering for individual *HLA* alleles. **e.** Average distance comparisons between pairs of 9mer peptides (orange bars-LC-

MS/MS data; blue bars-IEDB data) presented by an allele. The average distance between IEDB and LC-MS/MS peptides, purple bars. **f, g.** Specific examples of non-metric multidimensional scaling (NMDS) visualization of clusters of peptides for *HLA-A\*02:01* (**f**, well represented cluster in both MS and IEDB) and *HLA-A\*29:02* (**g**, enriched MS cluster). Each circle represents a unique 9mer peptide from either the LC-MS/MS (orange) or IEDB (blue) datasets, with the size of each circle proportional to a peptide's NetMHCpan-2.8 predicted binding affinity. Sequence logos representing these LC-LC-MS/MS and IEDB data are also shown for the highlighted peptide. **i.** Average peptide pairwise distances for a subset of peptides favorable for mass spectrometry detection. **j.** Experimental validation of peptides for allele *HLA-B\*35:01* as in **Figure 2d**. See also **Figure 2**.

**Figure S3, related to Figure 3: a.** Enrichment/depletion of protein sequence features among LC-MS/MS peptides. Each MS peptide was matched to 10 random decoy 9mers from the same source transcript. The relative rates at which hits and decoys mapped to Uniprot-defined sequence features (alpha helices, beta strands, signal peptides, and so on) were calculated as ratios and assessed by chi-square test. **b.** Expression of proteasome genes in B721.221 cells and in high-purity (>95%) samples from TCGA. Purity was determined according to the "percent tumor cell" field in the clinical slide review; if more than five samples were of sufficient purity for a given tumor type, only the top five were used. See also **Figure 3**.

**Figure S4, related to Figure 4:** Top: Example ROC curves for two hypothetical predictors (red and blue) with identical AUC scores. Bottom: PPVs for the same to two predictors at different rank thresholds (x-axis) for defining a "positive". At threshold 0.001, at which only the top 0.1% of evaluations are called as positive, the two predictors differ dramatically in performance. See also **Figure 4**.

**Figure S5, related to Figure 5: a.** Saturation analysis. For each allele, neural network models with peptide-intrinsic features and dummy sequence encoding only were built with increasing number of positive training examples, from 15 to the total number peptides identified by LC-MS/MS per allele. The PPV for each model was evaluated and plotted as a function of the number of binders in the training set. Allele complexity scores, defined as a weighted average of the entropy at each peptide position, are shown in the figure legend (below plot). **b.** Internal evaluation average PPV (top) and AUC (bottom) across the 16 alleles achieved by

NetMHCpan-2.8, NetMHC-4.0, and the two MS-based ensembles trained on LC-MS/MS dataset. **c.** Standard AUC plots are shown per allele based on NetMHC-4.0, NetMHCpan-2.8, 'MSIntrinsic' and 'MSIntrinsicEC', internal evaluations (5-fold cross-validation) of  $n$  hits merged with  $999n$  decoys, where  $n$  is the number of binders for the allele in the LC-MS/MS data (left) and AUC zooming into the [0,0.1]% false positive rate (right). **d.** Bars show AUC (top) and PPV (bottom) results per allele for the 'MSIntrinsic' model trained and evaluated only on IEDB data (5-fold cross-validation) as compared to NetMHC-4.0 and NetMHCpan-2.8. Note that only true non-binders from the IEDB data set were used in both training and evaluation and no random decoys were introduced. See also **Figure 5**.

**Table S1, related to Figure 1: Tab A. Negative Control Peptides** List of peptides identified in negative control immunopurification experiments. **Tab B. Master List Controls Removed** A complete list of HLA-associated peptides identified across 16 *HLA* alleles with peptides identified from the negative control immunopurifications removed. Individual *HLA* allele lists including peptides from the negative control immunopurifications are also reported. **Tab C. CysteinyLATED Peptides** A list of cysteinyLATED peptides identified from all mono-allelic cell lines.

**Table S2, related to Figure 2: Tab A. Sample Summary.** Summary of the samples used for HLA-peptide identification. A description of the global allele frequency, the amount of cell equivalents from each immunopurification used for MS analysis, the number of MS raw files, and the total validation yield from each *HLA* allele are reported. **Tab B. Raw Files** A list of the raw files and the MS instrument used for data collection are also included.

**Table S3, related to Figure 4:** Nested sets of peptides identified from all the *HLA* alleles characterized.

**Table S4, related to Figure 5: Tab A.** Positive predictive value (PPV) calculations from SLECA Model used to quantify the relative contribution of variables to HLA-peptide presentation for individual *HLA* alleles. **Tab B.** Machine Learning model performance for individual *HLA* alleles with available stability predictions. **Tab C.** Internal Evaluation. AUC and PPV machine learning model performance for individual *HLA* alleles as evaluated on the LC-MS/MS data set. **Tab D.** PPV and AUC evaluation results on DFRMLI competition data set along with the number of binders and non-binders per allele. **Tab E.** Due to the small size of the data set, the rank of each evaluated HIV epitope is shown for ‘MSIntrinsic’, NetMHC-4.0 and NetMHCpan-2.8 predictors, instead of PPV and AUC evaluations. **Tab F.** PPV and AUC evaluation results on Mann and Trolle data sets (Bassani-Sternberg et al., 2015; Trolle et al., 2016); NetMHC and NetMHCpan performance values shown with and without correction for MS bias.

## **Supplemental Experimental Procedures**

### **Cell Culture and HLA-peptide immunopurification**

Single *HLA* class I allele-expressing B cells were generated by transduction of the *HLA* class I negative 721.221 cells with a retroviral vector to express a single *HLA* class I allele as described previously (Reche et al., 2006) (cells expressing *HLA*-A\*02:01, -A\*24:02 and -B\*44:03 purchased from the Fred Hutchinson Research Cell Bank, University of Washington; cell expressing *HLA*-A\*03:01 were a gift from Dr. Marcus Altfeld, Ragon Institute; others were a gift from Dr. E.L. Reinherz, DFCI). The class I *HLA*-types of cell lines were confirmed by standard molecular typing (Brigham and Women's Hospital Tissue Typing Laboratory, Boston MA). Cells were cultured and HLA-peptide immuno-purification was performed.  $5-10 \times 10^7$  single HLA-allele expressing 721.221 cells were dissociated using 2 ml of protein lysis buffer (20 mM Tris [pH 8.0], 1 mM EDTA, 100 mM NaCl, 1% Triton X-100, 60 mM *n*-octylglucoside, phenylmethylsulfonyl fluoride (Sigma-Aldrich, St. Louis, MO) and protease inhibitors (Complete Protease Inhibitor Cocktail tablets, Roche Life Science, Indianapolis, IN) 200 units of DNase (Roche Life Science, Indianapolis, IN). Cell membranes were further disrupted using 500 watts, 20kHz, QSonica500 sonicator (QSonica, Newtown, CT) at 35% amplitude using 10 sec pulses until all the visible precipitates were solubilized. Lysates were pre-cleared using microfuge centrifugation for 20 minutes at 12,000 rpm at 4°C. Soluble lysates were co-incubated with 20  $\mu$ l of GammaBind Plus Sepharose beads (GE Lifesciences, Piscataway, NJ) non-covalently linked to W6/32 antibody (Santa Cruz Biotechnology, Dallas, Texas) for 3 hours. Beads were washed four times with lysis buffer without protease inhibitors and Triton X-100, four times with 10 mM Tris (pH 8.0) and once with distilled water.

### **HLA-peptide elution and desalting**

Peptides were eluted from HLA complexes and desalted on in-house built Empore C18 StageTips (3M, 2315) (Rappsilber et al., 2007). Sample loading, washes, and elution were performed on a tabletop centrifuge at a maximum speed of 1,500-3,000  $\times g$ . StageTips were equilibrated with  $2 \times 100 \mu$ L washes of methanol,  $2 \times 50 \mu$ L washes of 50% acetonitrile/0.1% formic acid, and  $2 \times 100 \mu$ L washes of 1% formic acid. In a tube, the dried beads from HLA-associated peptide IPs were thawed at 4°C, reconstituted in 50  $\mu$ L 3% ACN/5% formic acid, and loaded onto StageTips. The beads were washed with 50  $\mu$ L 1% formic acid, and peptides were further eluted using two rounds of 5 minute incubations in 10% acetic acid. The combined wash and elution volumes were combined and

loaded onto StageTips. The tubes containing the IP beads were washed again with 50  $\mu$ L 1% formic acid, and this volume was also loaded onto StageTips. Peptides were washed twice on the StageTip with 100  $\mu$ L 1% formic acid. Peptides were eluted using a step gradient of 20  $\mu$ L 20%ACN/0.1% formic acid, 20  $\mu$ L 40%ACN/0.1% formic acid, and 20  $\mu$ L 60%ACN/0.1% formic acid. Step elutions were combined and dried to completion.

### **Whole proteome analysis of single-HLA allele expressing cell lines**

25  $\mu$ g of trypsin-digested cell lysate (Mertins et al., 2013) from HLA-A\*29:02 and HLA-B\*51:01 expressing cell lines were fractionated using a previously described high-pH reverse phase StageTip protocol (Dimayacyac-Esleta et al., 2015). 5 fractions were collected from each cell line using the following increasing acetonitrile concentrations (10%, 15%, 35%, 55%, and 80%), dried to completion, and reconstituted in 9  $\mu$ L 3% acetonitrile/5% formic acid solution. Approximately half of each sample (4  $\mu$ L) was analyzed in a single-shot MS run as described below. Greater than 70% overlap (>4,300 proteins) was observed between the unique protein identification (>2 unique peptides per protein) from HLA-A\*29:02 (>5,200 proteins) and HLA-B\*51:01 (>5,100 proteins) expressing cell lines.

### **HLA-Peptide sequencing by tandem mass spectrometry**

All nanoLC-ESI-MS/MS analyses employed the same LC separation conditions described below. Samples were chromatographically separated using a Proxeon Easy NanoLC 1000 (Thermo Scientific, San Jose, CA) fitted with a PicoFrit (New Objective, Woburn, MA) 75  $\mu$ m inner diameter capillary with a 10  $\mu$ m emitter was packed under pressure to ~20 cm with of C18 Reprosil beads (1.9  $\mu$ m particle size, 200  $\text{Å}$  pore size, Dr. Maisch GmbH) and heated at 50  $^{\circ}$ C during separation. Samples were loaded in 3  $\mu$ L 3% ACN/ 5 % formic acid and peptides were eluted with a linear gradient from 7–30% of Buffer B (either 0.1% FA or 0.5% AcOH and 80% or 90% ACN) over 82 min, 30–90% Buffer B over 6 min and then held at 90% Buffer B for 15 min at 200 nL/min (Buffer A, either 0.1% FA or 0.5% AcOH and 3% ACN) to yield ~13 (FA)-18 (AcOH) sec peak widths. During data-dependent acquisition, eluted peptides were introduced into either a Q-Exactive plus (QE+) or Q-Exactive HF (QE-HF) mass spectrometer (Thermo Scientific) equipped with a nanoelectrospray source (James A. Hill Instrument Services, Arlington, MA) at 2.15kV. A full-scan MS was acquired at a resolution of 70,000 (QE+) or 60,000 (QE-HF) from 300 to 1,800  $m/z$  (AGC target  $1e6$ , 5ms Max IT). Each full scan was followed by top 12 (QE+) or 15 (QE-HF) data-dependent MS2 scans at resolution 17,500 (QE+) or 15,000 (QE-HF), using an

isolation width of 1.7 m/z with a 0.3 m/z offset, a collision energy of 25 (QE+) or 27 (QE-HF), an ACG Target of  $5e4$ , and a max fill time of 120 ms (QE+) or 100 ms (QE-HF) Max ion time. An isolation offset of 0.3 m/z was used so that doubly charged precursor isotope distributions would be centered in the isolation window. HLA peptides tend to be short, <15 amino acids, so the monoisotopic peak is nearly always the tallest peak in the isotope cluster and the mass spectrometer acquisition software places the tallest isotopic peak in the center of the isolation window in the absence of a specified offset. Dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 15 secs (QE+) or 10 secs (QE-HF). Charge state screening was enabled along with monoisotopic precursor selection using Peptide Match Preferred to prevent triggering of MS/MS on precursor ions with charge state 1 (only for alleles with basic anchor residues), >6, or unassigned.

### **Interpretation of LC-MS/MS Data, related to Figure 1**

Mass spectra were interpreted using the Spectrum Mill software package v5.1 pre-Release (Agilent Technologies, Santa Clara, CA). MS/MS spectra were excluded from searching if they did not have a precursor MH<sup>+</sup> in the range of 600-2000, had a precursor charge > 5, or had a minimum of <5 detected peaks. Merging of similar spectra with the same precursor m/z acquired in the same chromatographic peak was disabled. MS/MS spectra were searched against a database that contained all UCSC Genome Browser genes with hg19 annotation of the genome and its protein coding transcripts (63,691 entries; 10,917,867 unique 9mer peptides). A two-round search strategy was used (**Fig. 1c**). Prior to both search rounds, all MS/MS had to pass the spectral quality filter with a sequence tag length >2, i.e. minimum of 3 masses separated by the in-chain mass of an amino acid. In the first-round search, all spectra were searched using a no-enzyme specificity, fixed modification of cysteine as unmodified, no variable modifications, a precursor mass tolerance of  $\pm 10$  ppm, product mass tolerance of  $\pm 20$  ppm, and a minimum matched peak intensity of 50%. Peptide spectrum matches (PSMs) for individual spectra were automatically designated as confidently assigned using the Spectrum Mill autovalidation module to apply target-decoy based FDR estimation at the PSM rank to set scoring threshold criteria. Peptide auto-validation was done separately for each HLA allele with an auto thresholds strategy using a minimum sequence length of 7, automatic variable range precursor mass filtering, and score and delta Rank1 – Rank2 score thresholds optimized across all LC-MS/MS runs for an HLA allele. This yielded a PSM FDR estimate for precursor charges 1 thru 3, except for alleles A\*02:01 and A\*01:01 (charges 1 thru 4), of <1.0% for each precursor charge state. All confidently identified



peptides for each allele were used to define HLA-specific cleavage specificity for the position 2 and terminal anchors. In the second round search, all remaining spectra that were not confidently identified in the first round were searched using the HLA-specific cleavage specificity with the following allowed variable modifications added: oxidized methionine, pyroglutamic acid (N-term q), deamidation (n), and phosphorylation (s,t,y). An additional round of FDR thresholding as described above was applied to PSM's from the second-round search. The combined PSM's from each round had a peptide FDR <2.0% for each HLA allele. Cysteinylation (c) was searched in a third round using the same HLA-specific cleavage specificity because the under-representation of HLA-bound cysteine-containing peptides was revealed in our bioinformatic analyses.

The creation of decoy sequences during the Spectrum Mill search was adapted so that the target decoy thresholding above better mimicked HLA-peptide populations. Decoy sequence generation typically involves reversing an entire protein sequence (preserves enzyme cleavage frequency), scrambling peptide sequences randomly, or reversing the internal sequence while keeping the ends fixed to enable FDR estimation within a specified confidence interval based on the amounts of decoy and target matches. When generating decoys in Spectrum Mill for every sequence passing the precursor mass filter the peptide C-terminus was held fixed during the no enzyme search round. The second position was additionally held fixed during the HLA allele-specific cleavage round since HLA-associated peptides contain anchor residues at position 2 and last position.

#### **Database Search Evaluations, related to Figure 1**

The validation yield (number of valid PSMs /filtered PSMs) across our HLA datasets was calculated to be approximately 9% (range of 2%-26%). This median validation yield was similar to the identification rate reported for high-energy collisional dissociation (HCD) only HLA-associated peptide sequencing (Mommen et al., 2014). We then compared our HLA-A\*02:01 allele dataset to a high resolution dataset recently published for the HLA-A\*02:01 positive B cell line, JY (Bassani-Sternberg et al., 2015) (**Supplemental Fig. 1d**). Both datasets were searched using our strict filtering criteria and no enzyme specificity, as this was the specificity used by Bassani-Sternberg et al. A large degree of unique peptide overlap between our biologic replicates (71%) was observed, while a lower overlap (42%) was observed between two biological replicates of JY reported. We also calculated

the number of PSMs that passed our strict quality filters and 1% FDR estimation cutoff from the no enzyme and HLA-specific rounds of database searching (**Supplemental Fig. 1a**).

### **Assessment for MS bias, related to Figure 1**

To assess whether data gathered via mass spectrometry may exhibit technical biases, we first utilized the Enhance Signature Peptide (ESP) algorithm (Fusaro et al., 2009) to predict high-intensity peptides (“MS Observability Index”) within peptides in our MS dataset as well as within peptides recorded in the IEDB. Fourteen out of the 16 alleles in our study were included in this analysis due to the very low number of peptides in IEDB for two of the alleles: HLA-A\*02:04 and HLA-A\*02. Since anchor positions have allele-specific residue preferences and more data is available for some alleles than others, we considered 300 9mer binders chosen at random for each of the 14 alleles from each dataset (MS and IEDB), where for alleles with less than 300 identified binders the random sampling was performed with replacement. With the data thus formed, the ESPPredictor (available on GenePattern <http://genepattern.broadinstitute.org/>) was run for each peptide and the distributions of observability scores of peptides in the two data sets were compared (**Fig 1e**). To further probe for technical biases, we used the same data to evaluate the frequency of occurrence of each of the 20 amino acids within peptides in our MS and peptides in IEDB (**Fig. 1f**). Similarly, amino acid frequency was also compared between peptides in IEDB and two additional external mass spectrometry data sets (Bassani-Sternberg et al., 2015; Trolle et al., 2016) (**Supplemental Fig. 1f**).

### **Sequence properties of MS-identified peptides compared to IEDB, related to Figure 2**

#### *IEDB dataset*

A curated set of previously identified HLA-I bound peptides was downloaded from the Immune Epitope Database (IEDB) at <http://www.iedb.org/> (accessed on 10/26/2015) (Vita et al., 2015). The ‘MHC Assay Details’ option under ‘Specialized Searches’ was used and all ‘Linear Epitopes’ (under ‘Epitope’ menu box) associated with ‘MHC Class I’ (under ‘Assay’ menu box) were selected for each of the 16 alleles in our study. Furthermore, any epitope, which did not have a quantitative measure, was excluded.

#### *Affinity and length*

For each allele, MS-observed 9mer peptides were scored by NetMHCpan-v2.8 and compared to 1 million random 9mers drawn from the proteome (**Fig. 2a**). MS peptides (all lengths) were assessed in terms of their length distributions (**Fig. 2b**).

#### *Heatmap of positional amino acid differences*

We tabulated the amino acid counts for each allele at each position (1-9) within 9mer peptides, first for the MS dataset and separately for the IEDB dataset (for IEDB data, peptides with measured binding affinity of less than 500nM were considered). Alleles HLA-A\*02:04 and HLA-A\*02:07 have less than 10 binders peptides in IEDB and were excluded from the analysis, leaving 14 out of the 16 alleles in our study. At each (allele, position, amino acid) tuple, the number of peptides which contain the amino acid and the number of peptides which do not are counted and a chi-squared test is used to assess for differences between the MS and IEDB data sets.

#### *Sequence logo plots*

To capture and compare binding motifs between groups of peptides, sequence logo plots were generated using the motifStack R package (**Supplemental Fig. 2a**).

#### *Entropy*

The entropy at each 9mer position (1 through 9) was calculated for each allele based on all LC-MS/MS 9mer peptides identified for that allele (MS entropy) and then similarly for all IEDB 9mers binders (nM<500) (IEDB entropy). The computation was performed with MolecularEntropy() function from HDMD R package, where entropy values are normalized by log(20) such that entropy of 0 indicates a position with no variation while entropy of 1 indicates that all amino acids are equally likely to be observed at that position (**Supplemental Fig. 2b,c**).

#### *Peptide distance*

The following peptide distance metric was defined and computed between every pair of 9mer peptides in the MS and IEDB sets:

$$d(s_1, s_2) = \frac{1}{9} \sum_{i=1}^9 \text{distPMBEC}(s_{1i}, s_{2i}) * (1 - \text{entropy}_i),$$

where  $s_1$  and  $s_2$  are peptide sequences (e.g. KVLPIIQRW and HSRPIVTVW);  $s_{1i}$  is the amino acid at position  $i$  of the first peptide sequence;  $PMBEC$  is a pre-calculated matrix of residue similarities biased by their HLA binding properties (Kim et al., 2009) and  $\text{distPMBEC}$ , defined as  $\max(PMBEC) - PMBEC$ , is a 20x20 matrix capturing residue dissimilarities;  $\text{entropy}_i$  is the [0,1]-scaled entropy at position  $i$  for the allele associated with  $s_1$  and  $s_2$ . The average of MS and IEDB entropy was used in the distance metric computation.

#### *Peptide distance visualization and clustering*

A pairwise peptide distance matrix was computed between every pair of peptides 9mer peptides in the MS and IEDB sets as described above. Since the matrix contains relative peptide distances rather than absolute Cartesian coordinates, we used non-metric multidimensional scaling (NMDS) to visualize the peptides in two dimensions (nmds() function from ecodist R package). Density based clustering was then performed to assign peptides to clusters with dbscan() function from package dbscan (**Supplemental Fig. 2d**).

#### *Further assessing for mass spectrometry bias*

To assess the possibility that MS data clusters closely together due to mass spectrometry-related technological biases, we considered only the subset of peptides from MS and IEDB datasets with physicochemical properties that are favorable for MS detection. Namely, we selected peptides with one charged residue (by counting the R, H, and K residues per peptide) and peptides with moderate hydrophobicity by removing peptides which had hydrophobicity scores in the lowest and highest decile (a hydrophobicity score for each peptide was assigned with the hydrophobicity() function in Peptides R package). Analysis of the average peptide distances between MS and IEDB datasets and NMDS visualizations per allele were then carried out for this subset of favorable for MS peptides (36% of IEDB and 54% of MS peptides remained; alleles *HLA-A\*02:04* and *A\*02:07* were excluded due to low number of IEDB peptides), where the number of peptides was samples to be equal in the two data sets (**Supplemental Fig. 2h,i**).

#### *Direct affinity measurement*

To determine whether the MS dataset can be used to predict novel HLA-bound peptides, we built a binary (bound/not bound) generalized linear model for each of the 16 only using the MS data in addition to a random set of decoys from the proteome. We used these models to score each MS peptide. MS peptides were also evaluated with NetMHCpan-2.8 and those that scored in the top 10 percentile by MS-based models but bottom 10th percentile by NetMHCpan-2.8 were selected for experimental validation. Thirty three peptides across five alleles (*HLA-A\*01:01*, *-A\*29:02*, *-B\*35:02*, *-B\*51:01*, *-B\*54:01*) were synthesized (RS Synthesis, Louisville KY) and tested for binding to HLA molecules ( $IC_{50} < 500$  nM) by competitive HLA class I allele-binding per gel filtration protocol (Sidney et al., 2001) (**Fig. 2d and Supplemental Fig. 2j**).

### **Peptide Processing Analyses, related to Figure 3**

For each MS hit, the upstream 10 amino acids and downstream 10 amino acids were determined. To account for peptides near the beginning or end of their source protein, a 21<sup>st</sup> “amino acid”, denoted as “-”, was introduced to represent blank positions. For the minority of hits mapping to multiple genes, a selection was made at random. Each MS peptide was matched to 100 random 9mer peptides (drawn from the human proteome) but matched according to the first two and last two amino acids (to control for confounding signals from non-random sequence patterns in the proteome). In comparing the sequence context of the MS hits to the sequence context of the decoys (**Fig. 3a**), the relative enrichment for each amino acid at each position was calculated as a percent change, and the significance was calculated by chi-squared 2x2 contingency table test. Additional previously published MS datasets representing other cell types were analyzed using this same approach (**Fig. 3c-h**). The amino acids frequency analysis in **Fig. 3b**, which considers amino acids frequencies *within* the peptide, uses a separate set of decoys comprising 1,000,000 9mers drawn at random from the proteome (*i.e.* no matching on first two and last two amino acids of the peptide).

To understand the motifs favored by the cleavage prediction algorithm NetChop (Keşmir et al., 2002; Nielsen et al., 2005) (**Fig. 3i**), 1,000,000 random proteome 9mers and their corresponding sequence contexts were scored by the algorithm. The top-scoring 25% and bottom-scoring 25% were identified and analyzed in the manner of **Figure 3a** (top 25% treated as hits; bottom 25% treated as decoys).

To assess whether peptides might be enriched or depleted with respect to source protein sequence features, every MS peptide was matched to ten random 9mers from the same source gene. Then each hit or decoy was marked according to whether it intersected one of the Uniprot ([ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/uniprot\\_sprot.dat.gz](ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/uniprot_sprot.dat.gz)) sequence features: “STRAND”, “HELIX”, “TURN”, “SIGNAL”, or “COILED”. The relative frequency of these features was calculated for hits and decoys, and p-values were calculated by chi-square test (**Supplemental Fig. 3a**).

To determine how the relative expression of proteasome and immunoproteasome components in B721.221 cells compared to other tissues (**Supplemental Fig. 3b**), expression values (represented in transcripts per million) were compared against high-purity TCGA tumors (>95% according to the "percent tumor cell" field in the clinical slide review). If more than five samples were of sufficient purity for a given tumor type, only the top five were used.

To determine whether peptides were likely to be binding in non-canonical overhang conformations, 9mer and 10mer pairs were identified where the sequences were identical aside from 1 additional amino acid at the 10mer's n- or c-terminus (*i.e.* an “extension” of the 9mer, which one might presume binds with overhang). For each pair, another 100 10mers were *simulated* by extending the 9mer with a random amino acid (sampling at proteome frequencies). This procedure was repeated with 9mer+11mer pairs, and three peptide groups – the “core” 9mers, the “extended” 10mers and 11mers, and the simulated 10mers and 11mers – were compared in terms of their predicted binding affinities. Binding predictions were made by concatenating the first 5 and last 4 amino acids of each peptide and processing it with NetMHCpan-v2.8 as a 9mer. This prediction approach assumed that anchors remain at a fixed distance from the peptide termini (regardless of peptide length), which should be true if peptides always bind in a “tucked” conformation rather than an “overhang” conformation. If overhang conformation was common among the 10mers and 11mers of these nested sets, then the true 10mers and 11mers would not be expected to have better binding scores than the simulated 10mers and 11mers. On the other hand, if the true 10mers/11mers have similar scores as the 9mers, it suggests that nested sets only occur when short and long isoforms can both achieve a tucked conformation (strongly suggesting the overhang occurs rarely or never; **Supplemental Fig. 3j**).

#### **Relationship between expression and affinity, related to Figure 4**

RNA from 721.221 cells expressing HLA-A\*29:02-, B\*51:01-, B\*54:01-, and B\*57:01 was isolated using RNeasy mini kit (QIAGEN). The Nextera XT kit from Illumina and the Smart-seq2 protocol were employed to generate full length cDNA and sequencing libraries that were sequenced (50bp paired-end) on a HiSeq2500 Rapid Run. RNA-Seq data were deposited in the NCBI Gene Expression Omnibus (GEO accession number GSE93315). Expression of peptides were determined using RNA-Seq data from four libraries that were aligned to the UCSC transcriptome annotation (hg19, downloaded June 2015) using Bowtie2 (bowtie2-2.2.1, default parameters (Langmead and Salzberg, 2012)). Gene expression was quantified according to RSEM (rsem-1.2.19, default parameters (Li and Dewey, 2011)). Records for non-coding transcripts (per the UCSC annotation) were dropped and transcript per million (TPM) values were re-scaled and averaged across the four cell lines to yield a single expression value for each protein-coding transcript. The expression of a peptide was determined as the sum of the expression of the transcripts containing that peptide.

To assess the relationship between expression and affinity, the 9mer MS peptides for each of the 16 profiled alleles were binned according to their predicted expression and affinity (NetMHCpan-v2.8 prediction). Meanwhile, 1,000,000 random proteome decoy 9mers were binned in the same manner (for each allele). Finally, for each expression-affinity bin, the ratio of MS hits to decoys was calculated (**Fig. 4a**).

To understand the potential differences between observed MS peptides and HLA ligands that fail to be sampled in the MS, we identified peptides that were readily detected (top 10% of precursor ion intensity) to those that were just barely detected (bottom 10% of precursor ion intensity). Expression and affinity values (per NetMHCpan-v2.8 prediction) were compared for the two peptide sets (**Fig. 4b**).

To identify the potential impact of translational efficiency, the count of ATG 3mers upstream of the canonical ATG start codon was determined for each protein coding gene (per UCSC annotation). Each MS hit was matched to 10 9mer decoy peptides, which were chosen based on having similar RNA-Seq expression (minimum absolute log fold change) but different source gene. To avoid having all 10 decoys come from the same gene (which would add noise to the analysis), they were required to come from 10 different genes. In this manner, hits could be

compared to decoys in terms of the relative count of upstream ATGs in a manner controlled for relative gene expression (**Fig. 3c**). The significance of the association was determined by t-test (comparing the upstream ATG counts of hits vs. decoys).

#### **Impact of processing pathways, related to Figure 4**

##### *Localization*

Localization information was obtained from "SUBCELLULAR LOCATION" records in Uniprot's curated protein annotation ([ftp://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/complete/uniprot\\_sprot.dat.gz](ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz)). Uniprot's ID mapping table ([ftp://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/idmapping/by\\_organism/HUMAN\\_9606\\_idmapping.dat.gz](ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz)) as well as the UCSC-to-Uniprot ID mapping available from UCSC table browser (<https://genome.ucsc.edu/cgi-bin/hgTables>) were used to sync these data with UCSC annotations. Proteins were tagged as "Cell Membrane" if the localization field contained the text "cell membrane"; "Mitochondria" if "mitochondr"; "Nucleus" if "nucle"; "Cytoplasm" if "cytoplasm"; "ER" if "Endoplasmic reticulum"; "Secreted" if "secret"; "Late Endosome" if "late endo". It was possible for a protein to be associated with more than one localization. To assess enrichments, 10 random 9mer peptides were drawn from the genome per MS hit, and the hits and decoys were compared in terms of their relative localization frequencies (**Fig. 4d**, left). In addition, an "expression-corrected" analysis was conducted (**Fig. 4d**, right) by using the decoy generation approach employed in **Fig 4c**.

##### *Aborted translation*

Two vectors (length 30000) representing protein positions originating at the N-terminus were created (initialized to zeros) and designated *O* ("observed") and *E* ("expected"). For each hit, 1 was added to the position determined for the peptide within the host transcript, and  $1/n$  was added to positions 1 through  $n$  in *E*, where  $n$  is the total number of positions that the peptide possibly could have come from (the total length of the protein minus the length of the peptide). The resulting *O/E* ratios, representing the ratio of observed to expected hits per position, were averaged at binning intervals of 100 (**Fig. 4e**).



### *Unfolded protein response*

All proteins in the proteome were scored according to a sequence-based estimate of protein instability (Guruprasad et al., 1990). MS hits and expression-matched decoys (using the expression-matching approach employed in **Fig. 4c**) were binned according to the instability scores of their source proteins. The relative ratio of hits in each bin was compared to that observed for the decoys (**Fig. 4f**). The significance of the association was determined by t-test (comparing the instability scores of hits vs. decoys).

In a second analysis, all protein-coding genes were assessed in terms of their content of intrinsically disordered sequence. Disordered sequence predictions from 6 tools (iupred-l.disrange, iupred-s.disrange, espritz-d.disrange, espritz-n.disrange, espritz-x.disrange, and anchor.disrange; <http://d2p2.pro> (Oates et al., 2013) were available for the Gencode V19 human gene annotation; 12mers that were disordered according to three or more of the predictors were identified and counted for each gene in the UCSC gene annotation. MS hits and expression-matched decoys were compared according to the percent disorder (disordered 12mers divided by total 12mers) in their source proteins (**Fig. 4g**). The significance of the association was determined by t-test (comparing the percent disorder of hits vs. decoys).

### *Ubiquitination*

Previously published ubiquitin-targeting IP-MS/MS experiments in KG-1, Jurkat, or MM1S cells (Kronke et al., 2015; Krönke et al., 2014; Udeshi et al., 2012) were pooled to define a set of putative ubiquitination sites, and these sites were counted per gene in the UCSC annotation. Hits and expression-matched decoys were compared in terms of their counts of ubiquitination sites, and significance was determined by t-test (comparing the site counts in hits vs. decoys). The p-value is presented as “0” since it was less than the machine precision of our operating system (approximately  $1 \times 10^{-300}$ ) (**Fig. 4h**).

### *Protein turnover pathways*

Results from nearly 200 IP-MS/MS experiments targeting various protein turnover pathways genes (<http://besra.hms.harvard.edu/ipmsmsdbs/cgi-bin/downloads.cgi>; <http://www.nature.com/nature/journal/v466/n7302/full/nature09204.html>) were downloaded, and the protein identifications in each

experiment were sorted according to their “Weighted D-Score”, a measure of confidence that the given protein physically interacts with the bait. Each set was trimmed to include the only top 100 identifications to deplete it of non-specific binders. Then, for each set, we counted the number of MS hit peptides (vs. the number of expression-matched decoy peptides) that could be assigned to a protein in the set. Enrichment was assessed as the rate of hits in the set divided by the rate of decoys in the set, and the p-value was determined using a chi-squared 2x2 contingency table (**Fig. 4i**).

### **Development of new epitope selection algorithms, related to Figure 5**

#### *Positive predictive value (PPV)*

To assess the prediction performance of new models, we needed a metric well-suited to the epitope selection problem, where which only a small fraction of candidate peptides are expected to be presented. Indeed, each HLA allele is expected to present a repertoire of approximately ~10,000 peptides (Bassani-Sternberg et al., 2015; Hunt et al., 1992; Rammensee et al., 1995, 1999; Vita et al., 2015) among the  $1.1 \times 10^7$  9mer peptides in the proteome, meaning that approximately only 1 out of thousand peptides gets presented. Therefore, we evaluated models in terms of their ability to distinguish MS-observed 9mers among a 999-fold excess of decoy peptides (9mers randomly drawn from the proteome and not observed to bind). The 0.1% top-scoring peptides were considered as positives, and positive predictive value ( $TP/(TP+FP)=PPV$ ) was assigned according to this threshold. This PPV metric was preferred for this analysis (and all subsequent analyses) because it better represents predictor performance among the most highly ranked peptides. On the other hand, AUC scores integrate performance over all possible thresholds for defining a positive, even though many of these thresholds are not well justified (*i.e.* one would not reasonably nominate 10% of a random peptide set as binders). Demonstrating the importance of this distinction, two hypothetical predictors with identical AUC can have drastically different performance when only the most highly ranked hits are considered (*e.g.* top 0.1%; **Supplemental Fig. 4**). We also note that the focus on most highly ranked hits is important in the context of biological follow-ups in which only high-likelihood candidates can realistically be pursued.

#### *Multivariate prediction*

To determine the synergism that might be achieved with models that incorporate multiple variables (predicted affinity, expression, cleavability, *etc.*), we built various logistic regression models (for each allele) to discriminate  $n$  MS-observed peptides from  $999n$  decoy peptides. Since some predictor variables had highly non-normal distributions, they were transformed in the following ways:

1. **NetMHCpan-v2.8** (Hoof et al., 2009) **affinity**: the log of the hit:decoy ratio was calculated for logarithmically spaced affinity bins and the overall curve was smoothed monotonically using the `isoreg()` function in R (Team, 2014). This log-ratio value was used rather than nM affinity directly.
2. **NetMHCStabPan** (Jørgensen et al., 2014) **stability**: half-lives were used directly since they were normally distributed
3. **RNA-Seq Expression**: the log of the hit:decoy ratio was calculated for logarithmically spaced expression bins and the overall curve was smoothed monotonically using `isoreg()`. This log ratio was used rather than the TPM values directly.
4. **Protein Expression**: “iBAQ” values (calculated by summing the intensities of observed peptides for a given gene by the theoretical count of tryptic peptides in the gene (Ishihama et al., 2005)) were log-transformed (with zeros set to one tenth the minimum observed iBAQ value).
5. **Cleavability scores**. A logistic model (described in next section) was built to distinguish MS peptides from decoys (using external data sets) and applied to the B721.221 data (for more details, see next section). The resulting predicted probabilities were then logit transformed. (Logit-transformed NetChop scores were also used for comparison).
6. **Localization**: Seven dummy (0/1) variables were created to encode the various possible cellular localizations (defined by Uniprot as previously described).

All 63 possible subsets of the 6 variables were evaluated for each allele according the PPV metric (**Supplemental Table 4a**). PPVs were averaged across all alleles (shown for select variable combinations in **Fig. 5a**). In addition, we found the order of variable addition that yielded the most PPV improvement soonest and determined the incremental improvement associated with each variable, considering this as its “explanatory contribution” (**Fig. 5b**).

*De novo prediction of cleavability*

An MS-based cleavability predictor was developed by training on previously published MS data sets that profiled melanomas (Bassani-Sternberg et al., 2016), peripheral blood, and the C1R cell line (Caron et al., 2015). To create a set of negative examples, each MS-observed peptide was first mapped to all possible length-matched peptides in the proteome that *a*) have identical amino acids in the N1, N2, C2, and C1 positions and *b*) are not observed as positive training examples. Among these candidate negative examples (typically hundreds), ten were selected at random (with replacement) using a probability weight proportional to the count of positive training examples mapping to the source transcript. This approach was taken to ensure that targets and decoys would be drawn from a similar set of source genes and resulted in a training set with 10 negative examples per positive example. Training was based on an encoding representing amino acid identities and properties (*i.e.* isA, isC, isD, isE, isF, isG, isH, isI, isK, isL, isM, isN, isP, isQ, isR, isS, isT, isV, isW, isY, and isBlank plus pKA, volume, and polarity ([http://www.proteinsandproteomics.org/content/free/tables\\_1/table08.pdf](http://www.proteinsandproteomics.org/content/free/tables_1/table08.pdf))) and included positions U3, U2, U1, N1, N2, N3, C3, C2, C1, D1, D2, and D3 as well as a weighted average of positions U30...U4 ( $W=1...27$ ), a weighted average of positions D4...D30 ( $W=27...1$ ), and an unweighted average of positions N3...C3. These data were used to train a neural network (2 hidden layers of 50 and 10 nodes; 20% dropout for regularization; keras neural networks library (<https://github.com/fchollet/keras>)). To eliminate MS bias against cysteines, cysteines in cysteine-containing peptides were converted to serines for the purpose of forward prediction.

The analyses in the paper apply this predictor to data sets that are distinct from those on which it was trained (**Figs. 5a** and **5b** apply the predictor to our B721.221 data; **Fig. 5d** applies the predictor to other external data sets (Bassani-Sternberg et al., 2015; Trolle et al., 2016). Thus overfitting is not expected.

#### *Machine learning, overview*

HLA-peptides sequenced by mass spectrometry along with a set of random decoys were used to build binary classifiers (one classifier per HLA allele) to predict whether a given peptide will bind to a specific HLA allele. Generalized linear models were first trained with the glmnet R package in a 5-fold cross-validation scheme. Theano was used to train two types of neural networks: three models which incorporate one of the sequence encoding schemes with the rest of the peptide-intrinsic features (amino acid properties, and peptide

characteristics), and three models which incorporate one of the sequence encoding schemes along with expression and cleavage features. Scores of the first three models were averaged together to form the ‘MSIntrinsic’ ensemble and scores of all six models were averaged together to form the ‘MSIntrinsicEC’ ensemble.

#### *Machine learning, model features*

Five different classes of features were used in various combinations:

1. Each 9mer peptide amino acid sequence was represented as a numerical vector of length 180 in three ways 1.1) dummy (or binary) encoding, 1.2) blosum62-based encoding, 1.3) a fuzzy encoding where the each position in the vector represent the similarity between the true amino acid at the current peptide position with each of the 20 amino acids according to the PMBEC matrix(Kim et al., 2009) ;
2. Each residue in a peptide was represented by the first three principle components of PCA on amino acid properties (27 features)(Bremel and Homan, 2010);
3. Eight different peptide characteristics extracted from the Peptides package in R (“boman”, “hmoment”, “hydrophobicity”, “helixbend”, “sidechain”, “xstr”, “partspec”, “pkc”);
4.  $\text{Log}_2(\text{TPM}+1)$  expression (as measured here);
5. MS-based cleavage score (as defined above).

#### *Neural Network Models*

Artificial neural networks were built following the same cross-validation procedure with an equal number of positive and negative training examples: a random sample of all hit peptides of size 10x the number of hits was taken (with replacement) and supplemented with a random set of decoys of the same size. The network architecture for the ‘MSIntrinsic’ model consisted of an input layer with 215 features (180 coding scheme, 27 amino acid properties, 8 peptide properties) and single hidden layer with 50 hidden units (**Supplemental Fig. 5a**). The final model scores were defined as the average of the outputs of 3 networks trained with different random initialization seeds. To compose the ‘MSIntrinsicEC’ ensemble model, first neural networks with 182 features (180 coding scheme, 1 expression, 1 cleavage) and the same number of hidden layer units were trained with 3 random initializations. The Final ‘MSIntrinsicEC’ scores were then calculated by taking the average of these networks and the ‘MSIntrinsic’ networks. The same 5-fold splits were used to train both types of neural networks

to ensure 'MSIntrinsicEC' improvements were not due to seeing more positive training examples. All neural network training was done using Theano and code development followed the deep learning tutorial at <http://deeplearning.net/software/theano/>.

### *Saturation Analysis*

To determine the number of peptides required to build a strong predictor, we carried out saturation analysis by training models with varying numbers of positive training examples (minimum of 15 and maximum the full set of MS-identified peptides) and by measuring PPV on a test set of fixed size. Performance improvement was seen to plateau at several hundred peptides (**Supplemental Fig. 5b**), with variation across alleles likely due to the varying complexity of the peptide repertoire per allele. Indeed, complexity score, defined as a decay-weighted average of the entropies at each peptide position, ranked the alleles with strongest performance, *HLA-A\*01:01*, -*B\*44:03*, -*B\*44:02*, -*A\*29:02*, as 1, 2, 3, and 5 of 16 respectively, from least to most complex.

### *Predicting external datasets using our MS-trained neural networks*

Performance of the MS-trained models was evaluated on 6 independent external data sources. First, we used a competition dataset of eluted 9mer peptides and non-binders (Zhang et al., 2011). 'MSIntrinsic' performed better for 2 of 4 alleles compared to NetMHC-4.0 and NetMHCpan-2.8, even though most of the competition dataset was included in IEDB and likely in NetMHC training (**Supplemental Table 4c**). Second, we evaluated our methods using a curated and orthogonal dataset of 52 HIV-1 epitopes (that were associated with 12 HLA alleles from our study) for which T cell responses had been detected in patients (Llano A, Williams, A, Overa, A, Silva-Arrieta, S, Brander, 2013). We evaluated on the set of all HIV 9mer epitopes (excluding any that overlap with our data) mixed with a set of all HIV decoys (all tiled 9-mers across HIV proteins, excluding true HIV epitopes, ~3000 peptides). After scoring and ranking all peptides, 'MSIntrinsic' was able to predict the top-ranked true epitope at the same or higher position compared to NetMHC-4.0 or NetMHCpan-2.8 for 9 of the 12 alleles (**Supplemental Table 4d**). Third, we made predictions on 9mer T cell response epitopes retrieved from IEDB (Chowell et al., 2015) by accessing PPV and AUC (**Supplemental Table 4e**). To compute PPV, the top 0.1% of the model's predicted peptides were considered true positives. We ruled out 0.01% because we have directly observed more than 1000 9mers for some alleles, and 1% would imply that 100,000 peptides are presented per

allele, which is inconsistent with previous biochemical estimates (Walz et al., 2015). We thus define PPV as the fraction of LC-MS/MS peptides found within the model's 0.1% top scoring peptides. In this way, we test how effectively a model calls MS peptides from a background of random peptides (e.g. for  $n$  MS-observed 9mer peptides, we mix in  $999n$  random 9mer decoy peptides from the human genome). Fourth, we predicted HLA-bound peptides an independent source of peptides eluted from purified HLA molecules using LC-MS/MS from 7 cell lines that express multiple HLA alleles (Bassani-Sternberg et al., 2015). For each allele that overlapped with our study, we first excluded peptides that were predicted to bind other alleles ( $<150\text{nM}$  by NetMHCpan-2.8) but not the allele of interest ( $>1000\text{nM}$ ), and then added  $999n$  decoys (**Fig. 5d, Supplemental Table 4f**). Finally, we evaluated our models on the soluble HLA single-allele mass spectrometry dataset generated by Trolle and colleagues. Similarly,  $999n$  decoys were introduced to the identified peptides and PPV and AUC were evaluated. Since the data is allele-specific, there was no uncertainty in assigning peptides to alleles (**Fig. 5d, Supplemental Table 4f**). To determine whether NetMHC's weaker performance related to MS bias, a second set of NetMHC-based predictions were made by B721.221-trained logistic regressions based on log NetMHC affinity, ESP observability, and count of cysteines. Expression data from for the cell lines in the two studies was downloaded from CCLE and ENCODE.

### Supplemental References

Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., et al. (2016). Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* *7*, 13404.

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.

Hassan, C., Kester, M.G.D., de Ru, A.H., Hombrink, P., Drijfhout, J.W., Nijveen, H., Leunissen, J.A.M., Heemskerk, M.H.M., Falkenburg, J.H.F., and van Veelen, P.A. (2013). The Human Leukocyte Antigen-presented Ligandome of B Lymphocytes. *Mol. Cell. Proteomics* *12*, 1829–1843.

Mommen, G.P.M., Frese, C.K., Meiring, H.D., van Gaans-van den Brink, J., de Jong, A.P.J.M., van Els, C.A.C.M., and Heck, A.J.R. (2014). Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (ET<sub>h</sub>CD). *Proc. Natl. Acad. Sci.* *111*, 4507–4512.

Rammensee, H.-G., Bachmann, J., Emmerich, N.P.N., Bachor, O.A., and Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* *50*, 213–219.

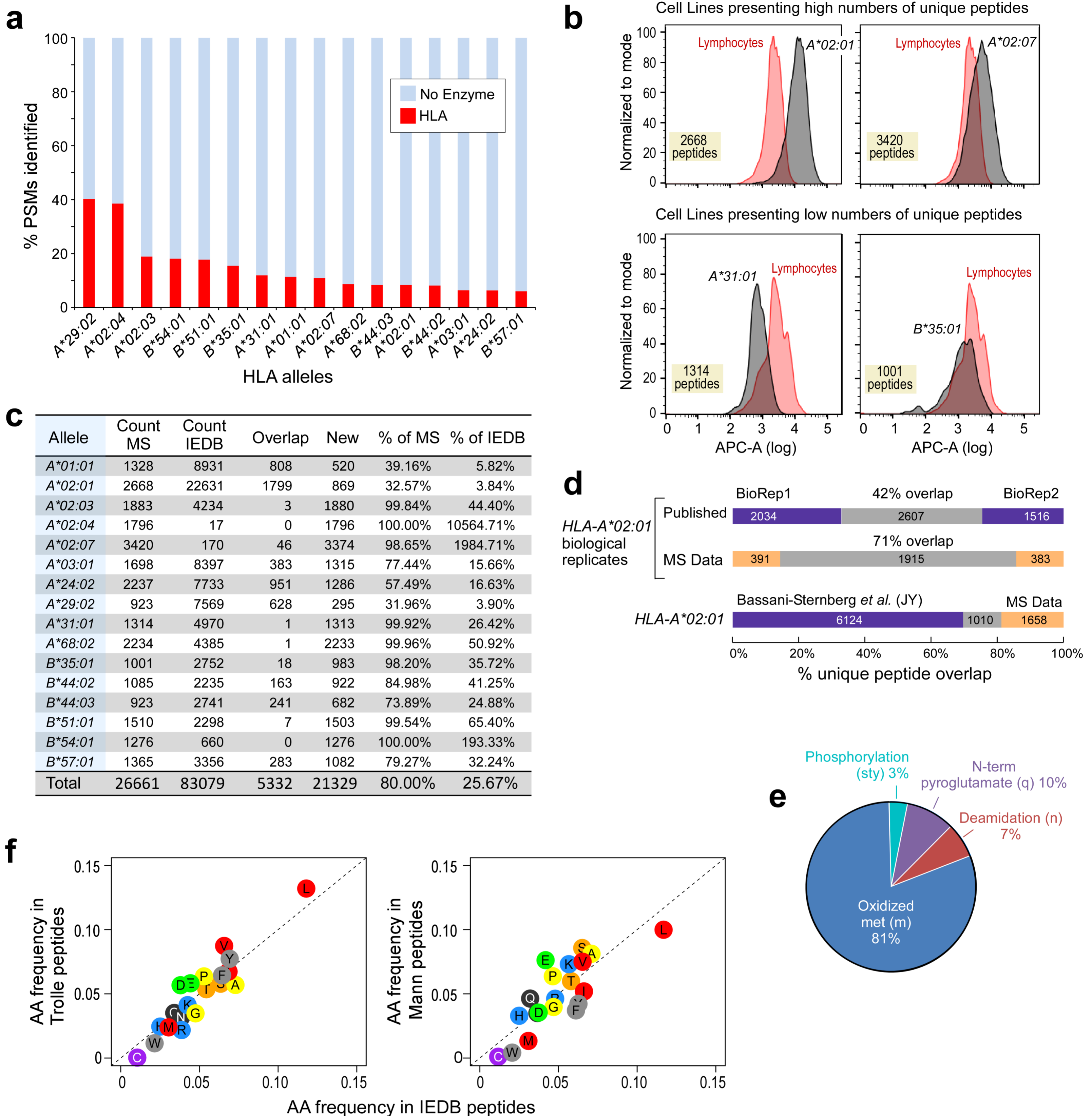
Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-

Team, R.C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

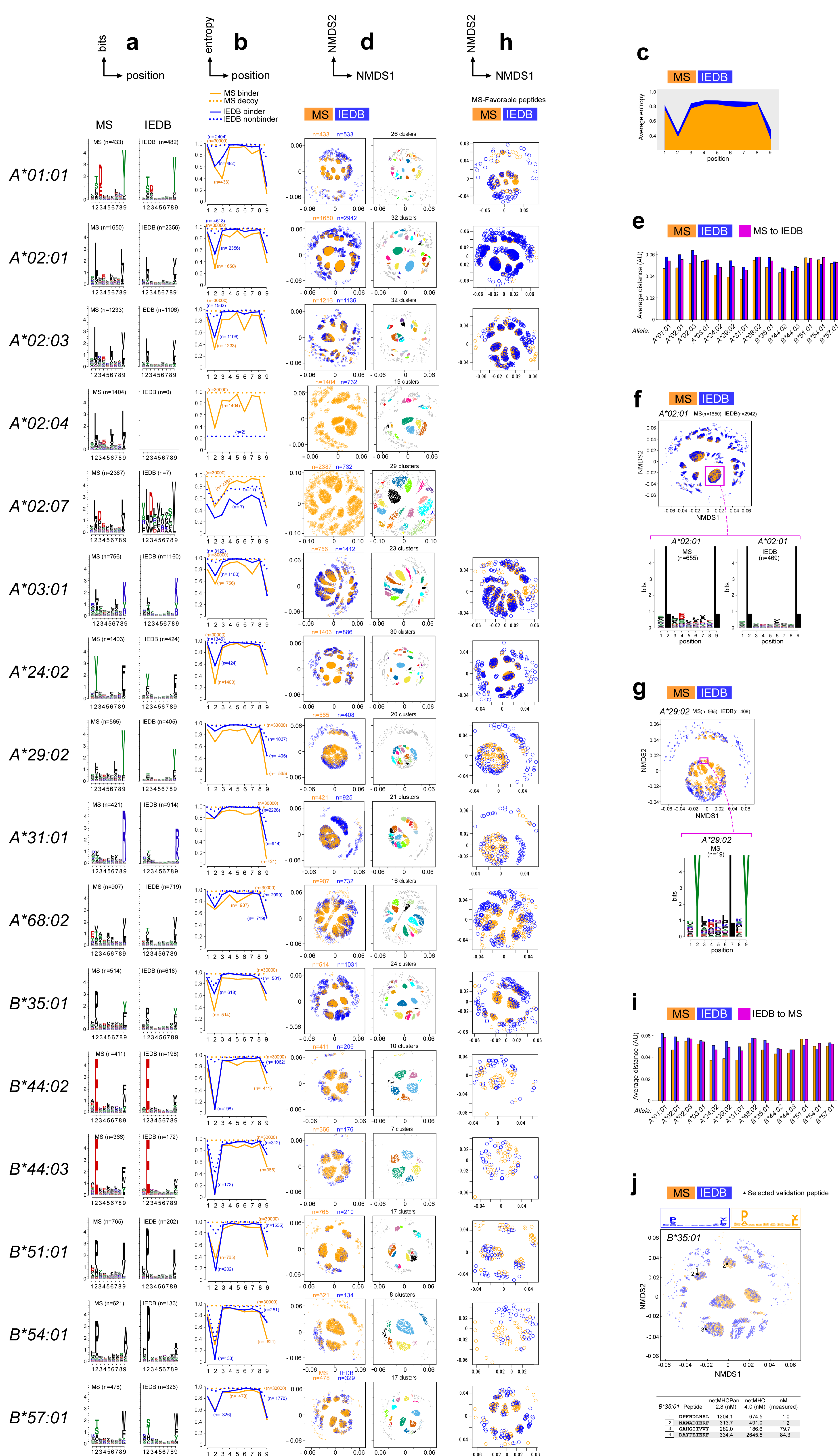
Walz, S., Stickel, J.S., Kowalewski, D.J., Schuster, H., Weisel, K., Backert, L., Kahn, S., Nelde, A., Stroh, T., Handel, M., et al. (2015). The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for T-cell-based immunotherapy. *Blood* *126*, 1203–1213.

Zhang, G.L., Lin, H.H., Keskin, D.B., Reinherz, E.L., and Brusic, V. (2011). Dana-Farber repository for machine learning in immunology. *High-Throughput Methods Immunol. Mach. Learn. Autom.* *374*, 18–25.



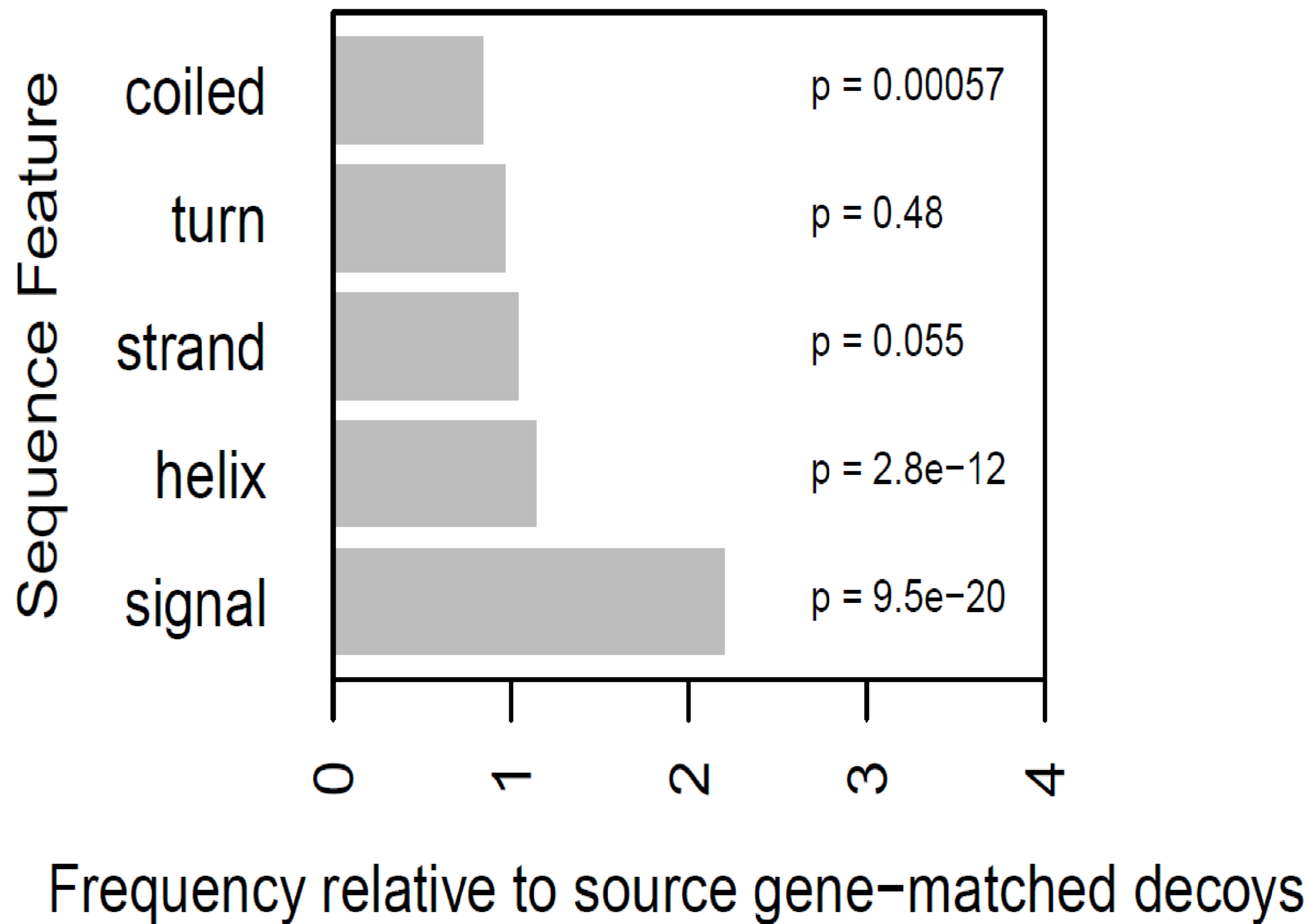
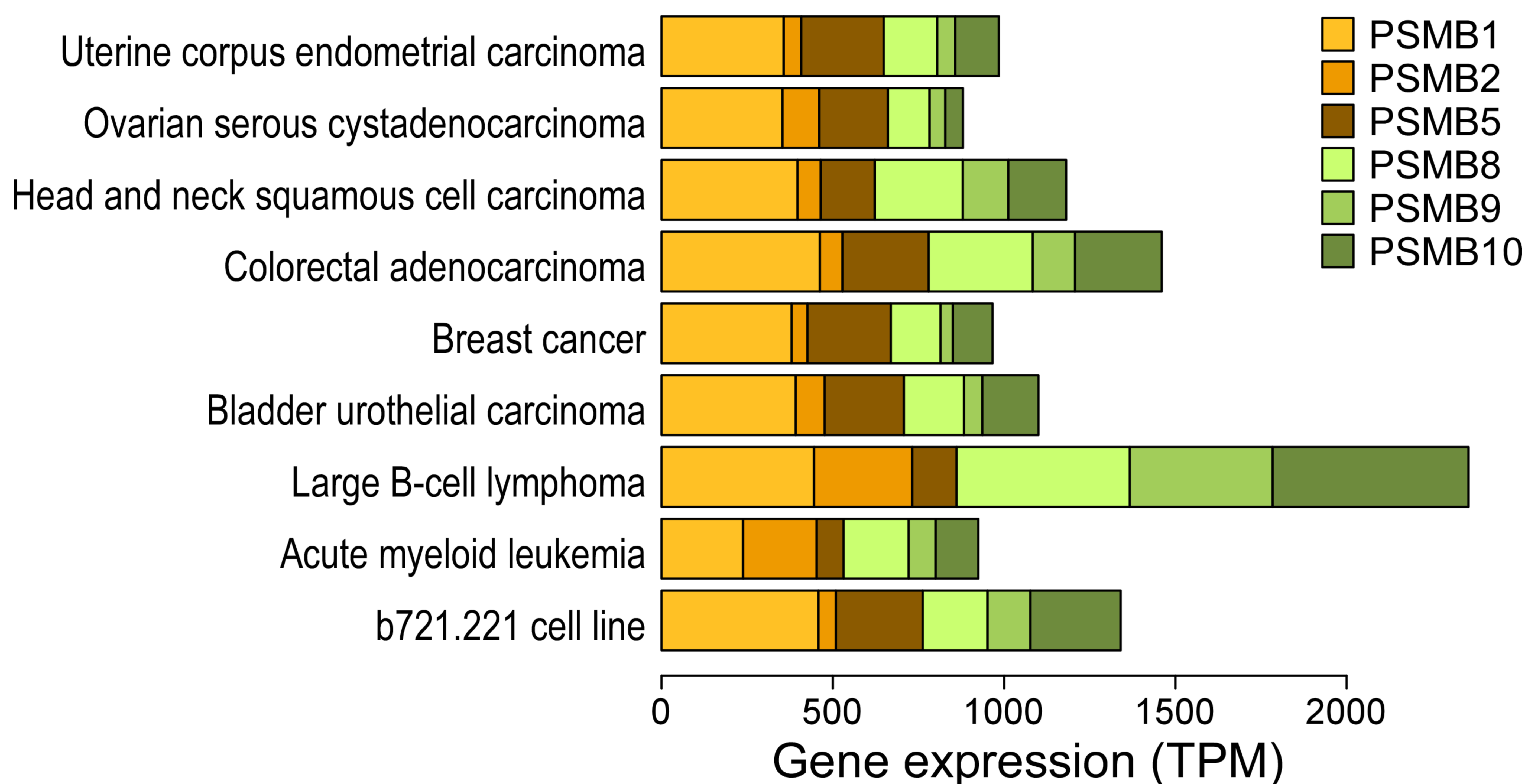


**Figure S1, related to Figure 1:** a) The number of peptide spectrum matches (PSMs) identified from both the no enzyme and HLA-specific rounds of database searches are shown for each HLA allele dataset. These PSMs represent the unique peptide identifications reported in Table 2. b) HLA cell surface presentation of single-HLA cell lines were compared to primary lymphocytes using FACS analysis. Cell lines that resulted in high (top; HLA-A\*02:01, -A\*02:07) and low (bottom; HLA-A\*31:01, -B\*35:01) numbers of HLA-associated peptides identifications by LC-MS/MS are shown. The number of total LC-MS/MS peptide identifications correlates with total cell surface HLA presentation. c) Comparison of the number of overlapping and unique peptides between our LC-MS/MS dataset and IEDB (all lengths). d) The overlap of unique peptides identified from biological replicates of our LC-MS/MS data (orange) and published data (purple) (Bassani-Sternberg et al., 2015) generated from immunopurifications of HLA-A\*02:01 expressing cells. Unique peptide overlap between our HLA-A\*02:01 dataset and this published dataset is also shown. e) The distribution of peptide modifications represented by the “Modified peptides” category in Figure 1 is shown as a pie chart. Peptide modifications included oxidized Met (m), deamidation (n), N-term Pyroglutamate (q), phosphorylation (sty), and cysteinylolation (c). f) Comparison of amino acid frequency between peptides in IEDB and peptides in the (Bassani-Sternberg et al., 2015; Trolle et al., 2016) respectively. To avoid biases due to anchor residues, for each comparison, 300 peptides per allele were selected at random for the alleles in the corresponding data set (Trolle: A\*01:01, A\*02:01, A\*24:02, B\*51:01; Mann: A\*01:01, A\*02:01, A\*03:01, A\*24:02, A\*31:01, B\*51:01) and pooled together before amino acid frequency was calculated. Amino acid frequencies are similar to those observed for our data (Fig. 1f). See also **Figure 1**.

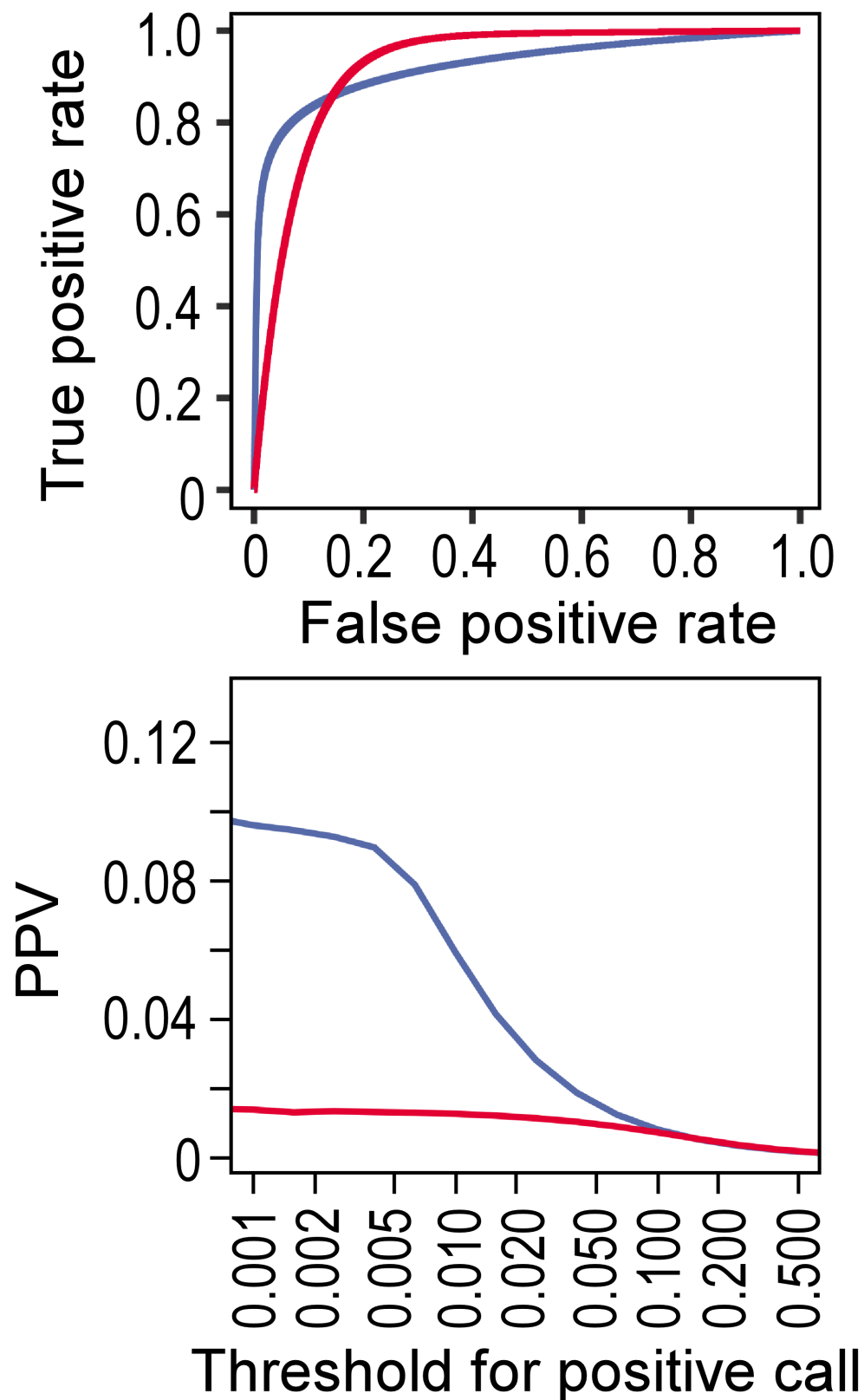


**Figure S2, related to Figure 2:** a. Sequence logos generated using 9mer data for the 16 HLA alleles characterized by LC-MS/MS and all 9mers with quantitative measurement in IEDB for the same alleles. b. Individual allele entropy calculations for each amino acid positions within 9mer peptides sequenced by LC-MS/MS (entropy is normalized by  $\log(20)$  and shown on to  $[0, 1]$  scale). c. Summary plot of entropy per 9mer position across all HLA alleles in LC-MS/MS (orange) and IEDB (blue) datasets (entropy is normalized by  $\log(20)$  and shown on to  $[0, 1]$  scale). d. NMDS plots showing HLA-associated 9mer peptide clustering for individual HLA alleles. e. Average distance comparisons between pairs of 9mer peptides (orange bars-LC-MS/MS data; blue bars-IEDB data) presented by an allele. The average distance between IEDB and LC-MS/MS peptides, purple bars. f, g. Specific examples of non-metric multidimensional scaling (NMDS) visualization of clusters of peptides for HLA-A\*02:01 (f, well represented cluster in both MS and IEDB) and HLA-A\*29:02 (g, enriched MS cluster). Each circle represents a unique 9mer peptide from either the LC-MS/MS (orange) or IEDB (blue) datasets, with the size of each circle proportional to a peptide's NetMHCpan-2.8 predicted binding affinity. Sequence logos representing these LC-LC-MS/MS and IEDB data are also shown for the highlighted peptide. i. Average peptide pairwise distances for a subset of peptides favorable for mass spectrometry detection. j. Experimental validation of peptides for allele HLA-B\*35:01 as in Figure 2d. See also **Figure 2**.

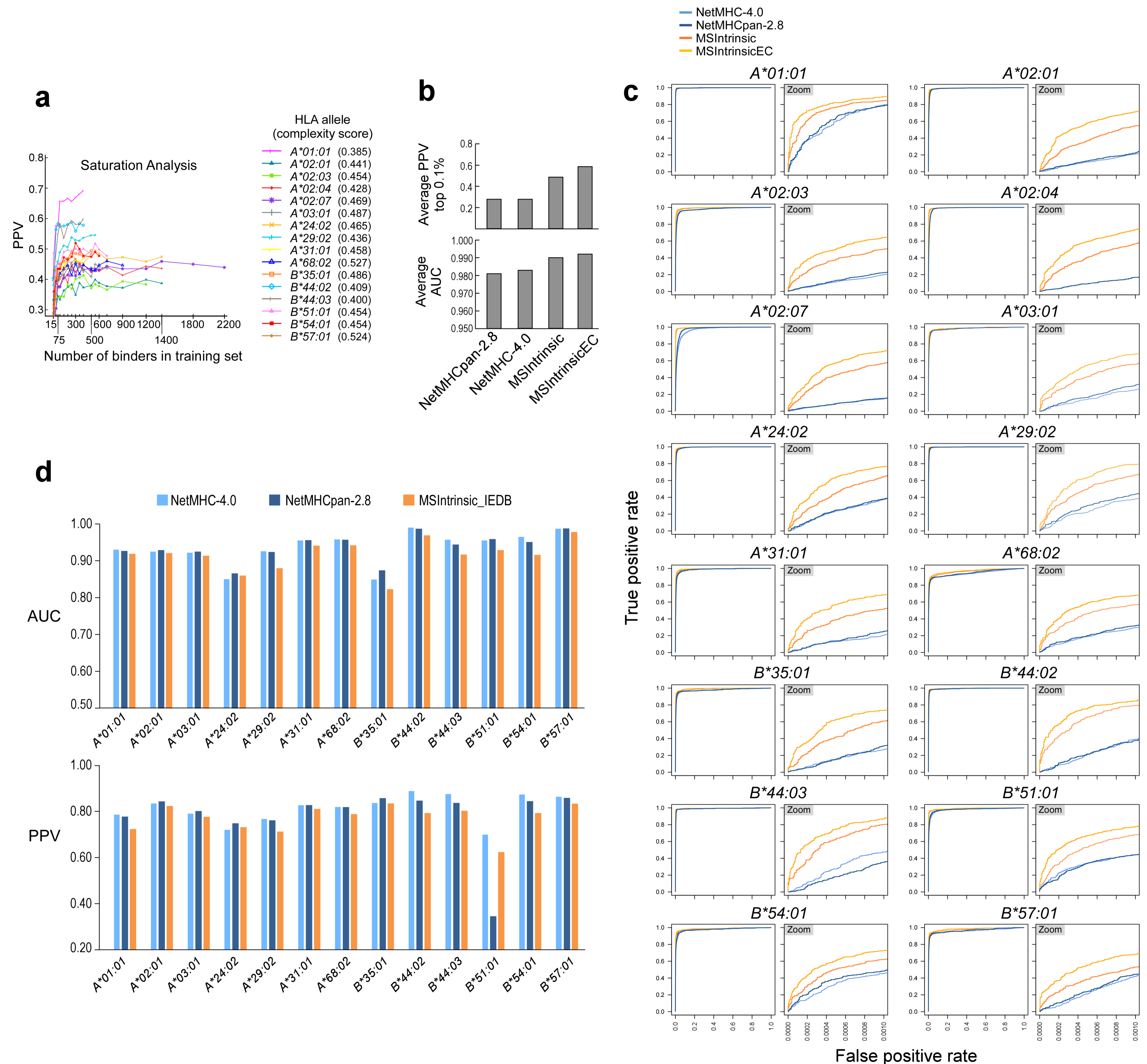
B*35:01	Peptide	netMHCpan 2.8 (nM)	netMHC 4.0 (nM)	nM (measured)
1	DPFRDLHSL	1204.1	674.5	1.0
2	NAWADIERF	313.7	491.0	1.2
3	GAGGILVVF	289.0	186.6	79.7
4	DAVPEIERF	334.4	2645.5	84.3

**a****b**

**Figure S3, related to Figure 3: a)** Enrichment/depletion of protein sequence features among LC-MS/MS peptides. Each MS peptide was matched to 10 random decoy 9mers from the same source transcript. The relative rates at which hits and decoys mapped to Uniprot-defined sequence features (alpha helices, beta strands, signal peptides, and so on) were calculated as ratios and assessed by chi-square test. **b)** Expression of proteo-some genes in B721.221 cells and in high-purity (>95%) samples from TCGA. Purity was determined according to the "percent tumor cell" field in the clinical slide review; if more than five samples were of sufficient purity for a given tumor type, only the top five were used. See **Figure 3**.



**Figure S4, related to Figure 4: Top:** Example ROC curves for two hypothetical predictors (red and blue) with identical AUC scores. **Bottom:** PPVs for the same to two predictors at different rank thresholds (x-axis) for defining a “positive”. At threshold 0.001, at which only the top 0.1% of evaluations are called as positive, the two predictors differ dramatically in performance.



**Figure S5, related to Figure 5:** **a)** Saturation analysis. For each allele, neural network models with peptide-intrinsic features and dummy sequence en-coding only were built with increasing number of positive training examples, from 15 to the total number of peptides identified by LC-MS/MS per allele. The PPV for each model was evaluated and plotted as a function of the number of binders in the training set. Allele complexity scores, defined as a weighted average of the entropy at each peptide position, are shown in the figure legend (below plot). **b)** Internal evaluation average PPV (top) and AUC (bottom) across the 16 alleles achieved by NetMHCpan-2.8, NetMHC-4.0, and the two MS-based ensembles trained on LC-MS/MS dataset. **c)** Standard AUC plots are shown per allele based on NetMHC-4.0, NetMHCpan-2.8, 'MSIntrinsic' and 'MSIntrinsicEC', internal evaluations (5-fold cross-validation) of  $n$  hits merged with  $999n$  decoys, where  $n$  is the number of binders for the allele in the LC-MS/MS data (left) and AUC zooming into the [0,0.1]% false positive rate (right). **d)** Bars show AUC (top) and PPV (bottom) results per allele for the 'MSIntrinsic' model trained and evaluated only on IEDB data (5-fold cross-validation) as compared to NetMHC-4.0 and NetMHCpan-2.8. Note that only true non-binders from the IEDB data set were used in both training and evaluation and no random decoys were introduced. See Figure 5.