Supporting Appendix

The C-terminal extension landscape of naturally presented HLA-I ligands

Running title: C-terminal extensions in HLA-I ligands.

Philippe Guillaume[a], Sarah Picaud[b,c], Petra Baumgaertner[a], Nicole Montandon[a], Julien Schmidt[a], Daniel E Speiser[a], George Coukos[a], Michal Bassani-Sternberg[a], Panagis Filippakopoulos[b,c], David Gfeller[a,d,1,2]

[a]Ludwig Institute for Cancer Research, Department of Fundamental Oncology, University of Lausanne, 1066 Epalinges, Switzerland.

[b]Structural Genomics Consortium, Old Road Campus Research Building, Roosevelt Drive, Nuffield Department of Clinical Medicine, University of Oxford, Roosevelt Drive, Oxford OX3 7DQ, UK.

[c]Ludwig Institute for Cancer Research, Nuffield Department of Clinical Medicine, University of Oxford, Roosevelt Drive, Oxford OX3 7DQ, UK.

[d]Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland.

[1] To whom correspondence should be sent: David.Gfeller@unil.ch.

[2] D.G. should be considered both as first and last author.

www.pnas.org/cgi/doi/10.1073/pnas.1717277115

SI Methods

Collection of HLA peptidomics data

HLA peptidomics data used in this study came from seven different studies (1-7) (Dataset S1). In all these studies, peptides were eluted from class I specific antibody purified HLA molecules. None of these data had been filtered with existing predictors, and therefore all had the potential to reveal non-canonical ligands. We only included MS samples generated with < 1% false discovery rate, with available HLA-I typing information and in which HLA-I motifs could be clearly annotated by our motif deconvolution approach for pooled HLA peptidomics data (7, 8). In total, our dataset comprises 43 samples covering 54 different HLA-I alleles (Dataset S1) for a total of 109,953 unique peptides (9- to 12-mers).

Analysis of non-canonical binding modes

Analysis of 9-mer HLA peptidomes

To analyze non-canonical binding modes in pooled HLA peptidomics studies that include ligands from up to six different HLA-I molecules, we developed the pipeline illustrated in Fig. 1. We first characterized the different motifs of the 9-mer HLA peptidomes using our recent motif deconvolution algorithm (8) (Fig. 1B). This algorithm has the major advantage of not depending on *a priori* knowledge of HLA-I binding specificity, and therefore can be applied even in the presence of poorly characterized HLA-I alleles. Each 9-mer peptide was then assigned to its corresponding motif and PWMs were built for each group of peptides, including random counts based on BLOSUM62 (7).

Prediction of N- and C-terminal extensions

To distinguish between the three different models describing longer peptides (i.e. bulge, N-terminal extensions, C-terminal extensions), we developed the algorithm illustrated in Fig. 1C. We first reasoned that longer peptides following the bulge model should have conserved binding specificity around the anchor residues at the second and last positions. We therefore modeled these peptides with the binding specificity of 9-mers at the first three and last two amino acids, and unspecific position in the middle. In the absence of a priori information about the amino acid preferences at terminal extensions, we modeled them as with one unspecific position at the N- or C-terminus. To enable meaningful comparison between the scores of the bulge and the terminal extension models, we did not impose any constrain at middle positions in the N- or C-terminal extension models. Each longer peptide (i.e., *L*-mer, with *L* equal to 10, 11 or 12) was then scored with all models (i.e. bulge, N- or C-terminal extensions) derived from each motif identified in the 9-mer HLA peptidome. Scores for the different models were computed as: $\sum_{i=1}^{3}\log\left(\frac{M_{X_i,i}}{P(X_i)}\right)+\sum_{i=8}^{9}\log\left(\frac{M_{X_{i+S},i}}{P(X_{i+S})}\right)$ for the bulge model; $\sum_{i=1}^{3}\log\left(\frac{M_{X_i,i}}{P(X_i)}\right)+\sum_{i=8}^{9}\log\left(\frac{M_{X_i,i}}{P(X_i)}\right)$ for the C-terminal extension model and $\sum_{i=1}^{3}\log\left(\frac{M_{X_{i+S},i}}{P(X_{i+S})}\right)+\sum_{i=8}^{9}\log\left(\frac{M_{X_{i+S},i}}{P(X_{i+S})}\right)$ for the N-terminal extension model, where M stands for the 20x9 PWM derived from 9-mer ligands for a given allele, X stands for the sequence of a *L*-mer peptide, *S* equals to *L*-9, and *P(X$_i$)* stands for the frequency of amino acid *X$_i$* in the human proteome. Peptides were then assigned to one model and one allele if their score with this allele and this model was higher than a threshold T_1 and no other score for all the other possible models of any allele was larger than the largest score minus T_2. Here we chose values of T_1 and T_2 equal to 2.0, since over 95% of the longer peptides found in our HLA peptidomics datasets had a score larger than this value for at least one allele and one model. We further developed a null model to assign Z-scores to the number of predicted C-terminal extensions with respect the expected number considering only bulges. Alleles with anchor residues at positions 4 to 7 (i.e., HLA-B08:01, HLA-B14:01 and HLA-B14:02) were excluded, as they display non-conserved binding motifs between 9- and 10-mers and much less 10-mer ligands (8).

The fractions of C-terminal extensions shown in Fig. 2B and 2D were computed as the number of peptides unambiguously assigned to the C-terminal extension mode of a given allele, divided by the number of peptides assigned to any of the three models for the same allele. Sequence logos were generated with the LOLA software (http://baderlab.org/Software/LOLA).

Null model

To account for possible noise in MS data and focus on predictions that showed statistical significance, we developed a null model representing the expected HLA peptidome from the bulge model, and described here for the case of 10-mer ligands.

Starting with a list 100’000 10-mer peptides randomly selected from the human proteome, we selected those that passed the threshold T_1 for the bulge model of at least one motif. The list of peptides was further randomly filtered so as to have the same number of peptides assigned to each allele as in the actual 10-mers HLA peptidomics data. We then re-predicted these peptides using all three models for each motif (Fig. 1C). This enabled us to assess how many 10-mer peptides generated from the bulge model of each allele could by chance be assigned to the N- or C-terminal extensions of any allele when considering the three models. The simulations were repeated 100 times to derive a Z-score. Only alleles with at least 50 10-mer peptides assigned to them, at least 10 peptides assigned to the C- or N-terminal extension model and with a Z-score larger than 2 were included in our predictions. Similar predictions were obtained using values ranging between 1.5 and 2.5 for the threshold T_1, or using a different threshold T_1 for each motif given by the score corresponding to the top 2% predictions in a large set of 100’000 10-mer peptides randomly selected from the human proteome (Table S3).

In vitro binding assays

All peptides used in Fig. 3 and 5, Fig S3, S4 and S11 and Table 1 were synthesized with free N and C-termini (1mg of each peptide, > 80% purity) at the Protein and Peptide Chemistry Facility of UNIL. Peptides were incubated separately with denaturated HLA alleles refolded by dilution in the presence of biotinylated beta-2 microglobulin proteins at temperature T=4°C for 48 hours. The solution was then incubated at 37°C. Samples were retrieved at time t=0h, 8h, 24h, 48h and t=72h. Stable complexes indicating interactions between HLA-I molecules and the peptides were detected by ELISA. Signals for 9-mer ligands at time t=0h were used for renormalization, while negative controls consisted of absence of peptides. Half-lives (Table 1) were computed as $\ln(2)/k_{off,}$ where k_{off} were determined by fitting exponential curves to the light intensity values obtained by ELISA, after removing the background signal (i.e., no peptides). Two independent replicates were performed for each measurement.

Expression and purification of HLA-A68:01 and β2M

The heavy chain HLA-A68:01 and the light chain β2M were purified from inclusion-bodies (9) with some small modifications. Briefly, recombinant expression plasmids were transformed into BL21 (DE3) (HLA-A68:01) or XA90 strain (β2M) bacteria. The cells were grown overnight in LB (Luria-Bertani medium) twice concentrated supplemented with 50 µg ml^{-1} kanamycin or 100 µg ml^{-1} ampicillin respectively at 37 °C. One litre of pre-warmed LB was inoculated with 10 ml of the overnight culture and was incubated at 37 °C. At an OD$_{600\ nm}$ of 0.6, expression was induced for 8 h at 37 °C with 1.0 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were then harvested by centrifugation (8700x*g*, 15 min, 4 °C, Beckman Coulter Avanti J-20 XP centrifuge), and then re-suspended in lysis buffer (10 mM TRIS-HCl, pH 8.0 at 20 °C complemented with 100 µg ml^{-1} Lysozyme (Sigma Aldrich), 250 units of Benzonase (Novagen), 1 mM EDTA and 1:1000 (v/v) Protease Inhibitor Cocktail III (Calbiochem)). After 20 min of continuous rocking at 22 °C, cells were lysed 3 times at 4 °C using a Basic Z-Model Cell Disrupter (Constant Systems Ltd, UK). Inclusion-bodies were isolated by centrifugation (16000x*g* for 1 h at 4 °C, JA 25.50 rotor, on a Beckman Coulter Avanti J-20 XP centrifuge). The supernatant was decanted and the contaminated material was removed by scrapping the outer rings (viscous and dark coloured) leaving the more compact and lighter coloured inner ring (containing the protein of interest) intact. The pellet was then fully dispersed in 20 ml of washing buffer (10 mM TRIS-HCl, pH 8.0 at 20 °C) complemented with 1:1000 (v/v) Protease Inhibitor Cocktail III (Calbiochem) and centrifuged again for 10 min (16000x*g* at 4 °C, JA 25.50 rotor, on a Beckman Coulter Avanti J-20 XP centrifuge). The washing/scrapping steps were repeated 5 times (until the outer rings vanished). The pellet containing the recombinant protein was then dissolved in 20 ml of solubilisation buffer (100 mM TRIS-HCl, pH 8.0 at 20 °C, 8 M urea). Insoluble material was precipitated by centrifugation (16000x*g* for 1 h at 4 °C, JA 25.50 rotor, on a Beckman Coulter Avanti J-20 XP centrifuge). Solubilized HLA-A68:01 heavy chain was immediately flash frozen in liquid nitrogen and stored at -80 °C. The recombinant β2M protein in urea was refolded by dialysis against 10 mM TRIS-HCl pH7.0 using SnakeSkin® Dialysis tubing (3.500 MWCO, Thermo Scientific) and purified by ion exchange on Hi-Trap Q HP (5 ml, GE Healthcare) column, with a linear gradient from 0 to 100 mM NaCl. Fractions containing pure β2M were dialysed overnight against water, concentrated with a 10 MWCO concentrator (Amicon® Ultra, MILLIPORE) to 2 mg ml^{-1}, flash frozen in liquid nitrogen and stored at -80 °C.

HLA:β2M:peptide complex assembly

The protein complex was reconstituted by dilution of the denatured HLA-A68:01 heavy chain (3 μM) and β2M (6 μM) in presence of the peptide ([H]-E-T-S-P-L-T-A-E-K-L-[OH], 10 μM) into 200 ml of refolding buffer (100 mM TRIS-HCl , pH 8.0 at 20 °C, 400 mM L-Arginine HCl (Sigma Aldrich), 2 mM EDTA, 5 mM reduced L-glutathione (Sigma Aldrich), 0.5 mM oxidized L-glutathione (Sigma Aldrich) and with 1:1000 (v/v) Protease Inhibitor Cocktail III (Calbiochem)). The refolding mixture was incubated at 10 °C during 36 h under constant stirring. Every 12 h another batch of the denatured HLA-A68:01 heavy chain (3 μM) was added to the mix. The 200 ml were then concentrated to 5 ml using a 3 MWCO concentrator (Amicon® Ultra, MILLIPORE). The concentrated protein mixture was further submitted to exclusion size chromatography (HiLoad™ 16/60 Superdex™ 75 prep grade, on Äkta Pure GE Healthcare) and each peak collected was submitted to SDS-PAGE gel and mass spectrometry (Agilent 6530 QTOF (Agilent Technologies Inc. - Palo Alto, CA)) to assess the simultaneous presence of the 3 components of the complex. Fractions of interest were then stored at 4°C until further use.

Crystallization

Prior to crystallization, the buffer of the protein complex was exchanged to 25 mM MES pH6.5 at 20 °C and 150 mM NaCl on a Superdex™ 200 Increase 10/300 GL column using an Äkta Pure system (GE Healthcare). The complex was then concentrated to 10.96 mg ml^{-1} final using a 3 kDa MWCO concentrator (Amicon® Ultra, MILLIPORE). Aliquots of the complex were set up for crystallization using a mosquito® crystallization robot (TTP Labtech, Royston UK). Coarse screens were typically setup onto Greiner 3-well plates using three different drop ratios of precipitant to protein per condition (100+50 nl, 75+75 nl and 50+100 nl). Crystallization was carried out using the sitting drop vapor diffusion method at 4 °C. Crystals were grown by mixing 100 nl of the protein complex (10.96 mg/ml) with 50 nl of reservoir solution containing 0.1 M HEPES pH 7.5, 12 % PEG3350, 0.005 M $CoCl_2$, 0.005 M $NiCl_2$, 0.005 M $CdCl_2$ and 0.005 M $MgCl_2$. Diffraction quality crystals grew within a few days.

Data Collection and Structure Refinement

Crystals were cryo-protected using the well solution supplemented with additional ethylene glycol and were flash frozen in liquid nitrogen. Data were collected at Diamond beamline I24 on a Pilatus3 6M detector at a wavelength of 0.96864 Å. Indexing and integration was carried out using XDS (10) and scaling was performed with SCALA (11). Initial phases were calculated by molecular replacement with PHASER (12) using a model of HLA/β2M peptide complex (PDB:5T6X). Initial models were built by ARP/wARP (13) followed by manual building in COOT (14). Refinement was carried out in REFMAC5 (15). Thermal motions were analyzed using TLSMD (16) and hydrogen atoms were included in late refinement cycles. Data collection and refinement statistics can be found in Table S4. The model and structure factors have been deposited with PDB accession code: 6EI2.

HLA-I sequence and structure analysis

For all 54 alleles considered in this work, the sequence of the peptide binding domains were retrieved from IMGT database (17) and aligned with MUSCLE (18). To investigate the molecular determinants of C-terminal extensions, we selected all amino acids surrounding the F pocket (75-81, 84, 114-118, 142-147, following residue numbering in HLA-I structures). The eight positions displaying the largest changes in amino acid frequencies (Jensen-Shannon divergence larger than 0.1) are shown in Fig. 4C. Using 20-dimensional encoding for each amino acid, we trained a logistic regression model based on *glmnet* package in R (19). Four-fold cross-validation was performed and repeated 10 times for different random seeds to compute AUCs.

Available HLA-I crystal structures were then collected for alleles considered in this work. In total, 20 alleles among those with available HLA peptidomics data had experimental X-ray crystal structures. For each allele, a reference structure was selected by prioritizing complexes with 9-mer ligands and highest resolution (Table S1). In each structure, the distance between the C-alpha of residues 80 and 143 was computed (Fig. 4D and Table S1). The AUC of a predictor using as input feature this distance was 0.92, which is comparable to the AUC of the predictor based on HLA-I sequences when considering only the 20 HLA-I alleles with available crystal structures (AUC=0.93, three-fold cross-validation).

HLA-I ligand predictors explicitly incorporating C-terminal extensions

For each allele displaying C-terminal extensions, we retrained our predictor based on HLA peptidomics data modeling C-terminal extensions as multiple motifs. In practice, all 10-mer ligands predicted to follow the C-terminal extension mode of each allele were treated separately and a PWM (M^C) was built with them (7). All other 10-mer ligands (non-C-terminal extensions) were used to build another PWM describing the bulge model (M^b). For alleles with C-terminal extensions, the score of a 10-mer peptide X=(X_1, … X_{10}) with this predictor is then computed as:

$$S(X) = \frac{1}{10}\log\left\{w^C \prod_{i=1}^{10} \frac{M^C_{X_i,i}}{P(X_i)} + w^b \prod_{i=1}^{10} \frac{M^b_{X_i,i}}{P(X_i)}\right\}$$

where w^C, respectively w^b, stands for the fraction of ligands predicted to follow C-terminal extensions, respectively bulges. For alleles without C-terminal extensions, the predictions did not change compared to the previous version of our predictor (7).

Benchmarking of the algorithm was done using HLA-A03:01 and/or HLA-A68:01 positive samples, since these are the two HLA-I alleles displaying C-terminal extensions in multiple samples in our collection of HLA peptidomics studies. Careful cross-validation was performed, where for each sample used as testing set, the predictor was trained only on the data from the other samples. To this end, we further excluded from our analysis samples where some other HLA-I alleles did not appear in other samples (i.e., Fibroblast and RA957). For each sample, 4-fold excess of random peptides from the human proteome were added as negatives and all peptides were ranked with our new predictor, using for each peptide the highest score across the different HLA-I alleles present in the sample. The fraction of positives found in the top 20% of the predictions (which in this case is equivalent to the recall since the number of true positives is equal to the number of predictions) as well as the AUC were then computed (Fig. 5A and Fig. S10). The performance of the predictor explicitly modeling C-terminal extensions (MixMHCpred1.1) was compared with the former version of MixMHCpred (v1.0, based on a single PWM for each allele) (7), NetMHC4.0 (20) and NetMHCpan3.0 (21).

Human PBMC, IFNγ-ELISpot assays and Multimer analysis

Healthy volunteers donated circulating leukocytes according to the standards of the Blood Transfusion Center in Epalinges, Switzerland (Service Vaudois de Transfusion Sanguine). Samples from patients were obtained under written informed consent following study protocol approval by the Human Research Ethics Committee of the Canton de Vaud (Switzerland).

PBMCs were peptide stimulated in vitro for 12 days in vitro in presence of 100 U/ml IL-2. Subsequently, the Elispot was performed using the ELISpot^{PRO} kit for Human IFNγ from MABTECH (3420-2APT-10), following the standard supplier instructions. 100'000 cells per well were re-challenged with the peptide for 16h. The spots were analysed by the iSpot Robot ELISpot reader (AutoImmun Diagnostika GMBH).

Monomeric HLA-A03:01/`TVRSHCVSKI` complexes were prepared by refolding procedures using heavy chain HLA-A03:01 and the light chain β2M, as described before. The heavy chain contained added BSP (BirA enzyme Substrate Peptide). The enzymatic biotinylation of HLA-A3:01-BSP was performed over night at 25°C with ATP, biotin and the biotin ligase Bir A. Multimer complexes were prepared by mixing biotinylated HLA-A03:01/`TVRSHCVSKI` monomers with PE-labeled streptavidin (Invitrogen).

CD8 T cells of the CMV 10-mer peptide stimulated PBMC after 12 days in vitro stimulation were stained with a PE labeled HLA-A03:01/`TVRSHCVSKI` multimer and co-stained with anti-CD8 antibody (BC A94683) and DAPI for dead cell exclusion. Multimer+ CD8+ T cells were analysed at the BD ARIA III instrument equipped with the FACS Diva software. The analysis was performed with the FlowJo software (FLOWJO.LLC). The negative control consists of an EBV-derived peptides (`RVRAYFYSKV`)/HLA-A03:01 tetramer and a PBMC sample from another HLA-A03:01 and EBV seropositive healthy donor, for which we did not observe any recognition by CD8 T cells.

IFNγ-ELISpot, tetramer and refolding assays for the I(10)L, K(9)L and truncated 9-mer peptides derived from the CMV epitope were later carried out in the same way (Fig. S11A-C), including also the 10-mer CMV epitope as a positive control for the tetramer analysis. T cells recognizing the wild-type peptide (upper box in Fig. 5D) were further sorted, expanded and tested for cross-reactivity with the 9-mer epitope (Fig S11D).

SI Figures

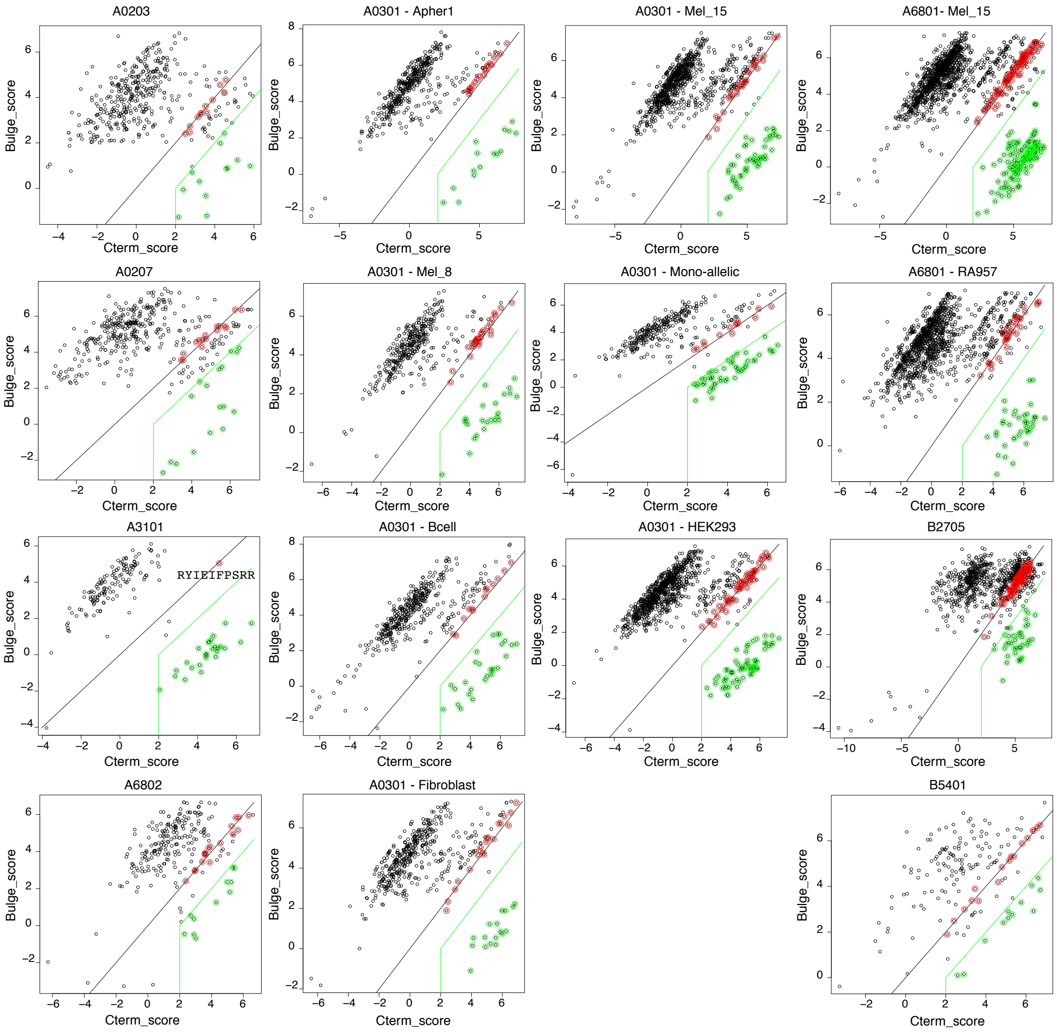


Fig. S1: Plots of the scores of peptides with the C-terminal extension model (x-axis) and the bulge model (y-axis) for all cases of predicted C-terminal extensions. Black lines show the y=x line, green circles show predicted C-terminal extensions (Dataset S3), green lines represent the thresholds used (SI methods) and red circles show peptides with very similar scores with both models (Dataset S4). Only peptides unambiguously assigned to the corresponding allele in pooled samples are shown (SI Methods).

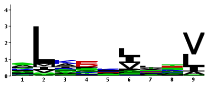
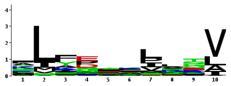
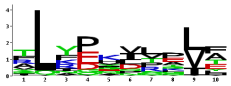
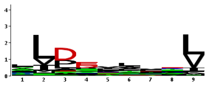
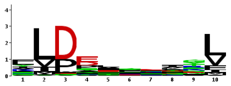
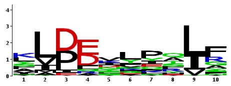
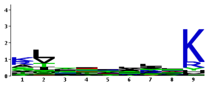
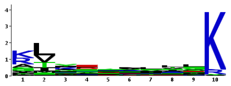
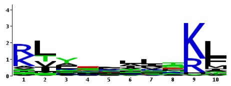
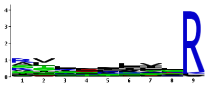
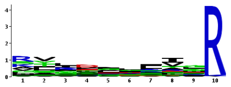
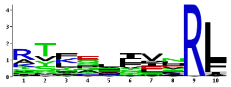
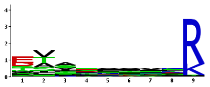
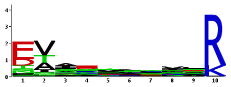
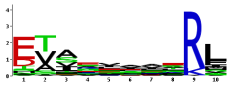
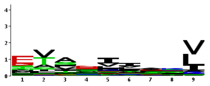
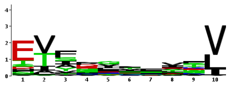
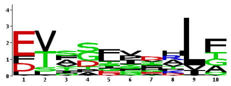
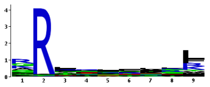
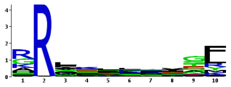
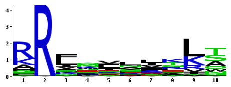
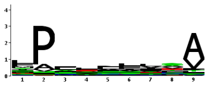
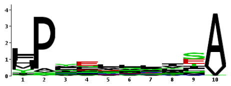
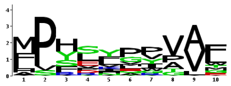
	9-mer motifs	10-mer bulges	10-mer C-terminal extensions
A02:03			
A02:07			
A03:01			
A31:01			
A68:01			
A68:02			
B27:05			
B54:01			

Fig. S2: Comparison between 9-mer motifs, 10-mer motifs from peptides predicted to follow the bulge model and 10-mer motifs from peptides predicted to display C-terminal extensions for the eight alleles with such extensions. For HLA-A03:01 and HLA-A68:01, peptides have been pooled from the different studies where C-terminal extensions were identified to generate the logos.

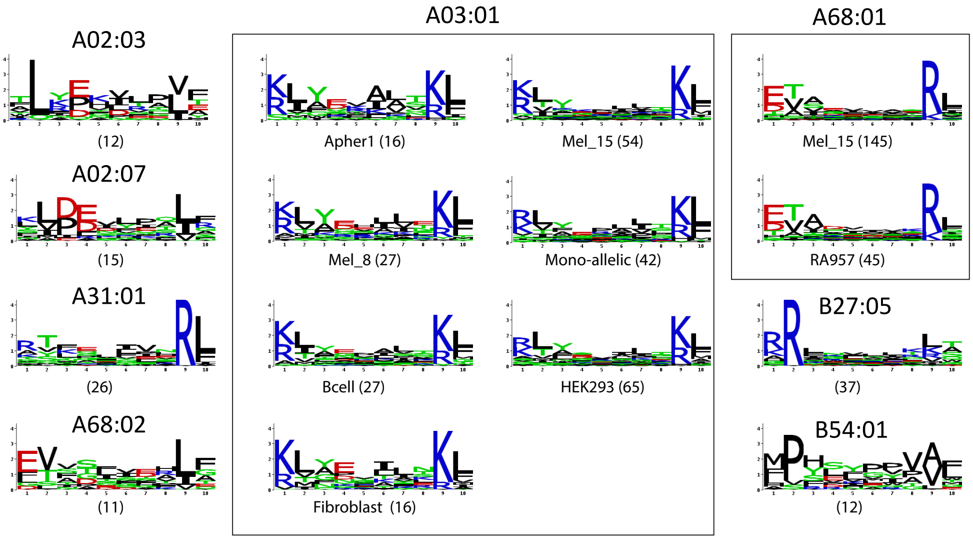


Fig. S3: Predictions of C-terminal extensions among 10-mer peptides in the presence of 5% of noise in all HLA peptidomics datasets.

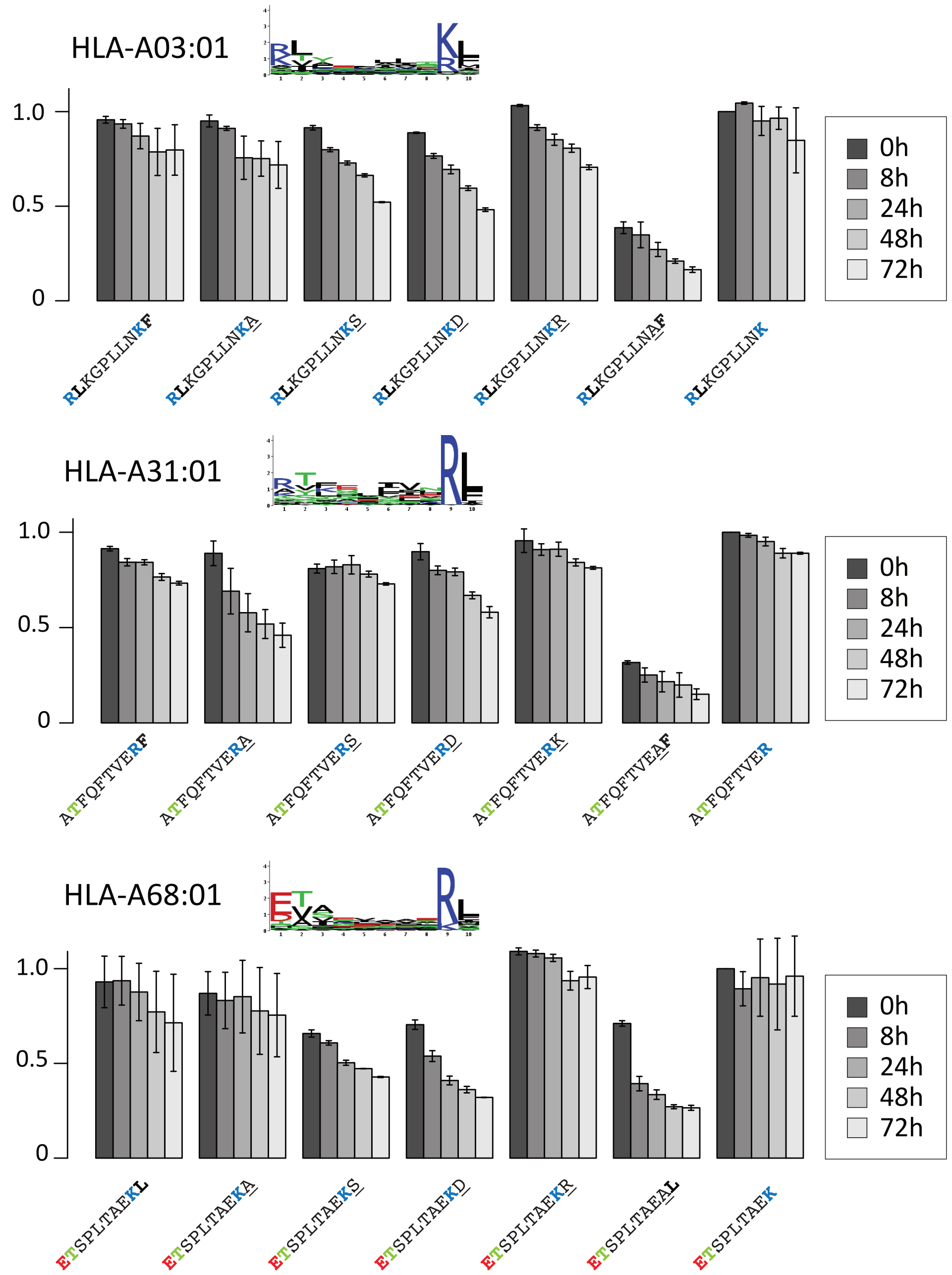


Fig. S4: Binding stability results for other P10 mutants (S,D,R/K) that did not match the specificity at P10 predicted in our analysis of MS data.

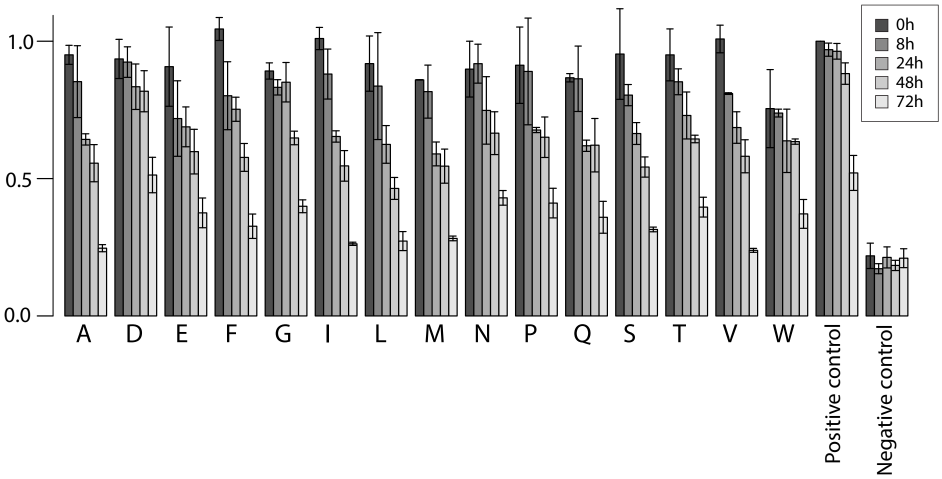


Fig. S5: Stability analysis of 11-mers built from the C-terminally extended 10-mer KLAYTLLNKL HLA-A03:01 ligand (Fig. 3) with all possible C-terminal amino acids not compatible with the specificity of the second anchor for HLA-A03:01 (i.e., KLAYTLLNKL[A/D/E/F/G/I/L/M/N/P/Q/S/T/V/W]). ILRGSVAHK was used as positive control. Negative control consists of absence of peptide.

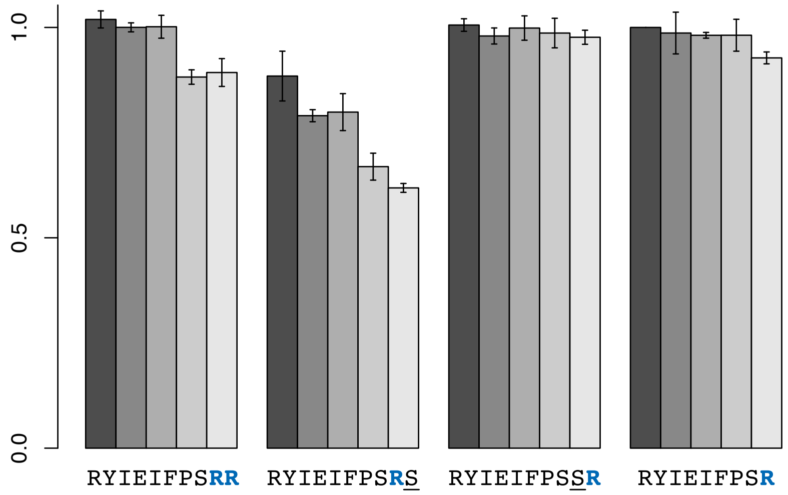


Fig. S6: Binding stability of the 10-mer RYIEIFPSRR, R(9)S and R(10)S mutants, and the 9-mer RYIEIFPSR in complex with HLA-A31:01.

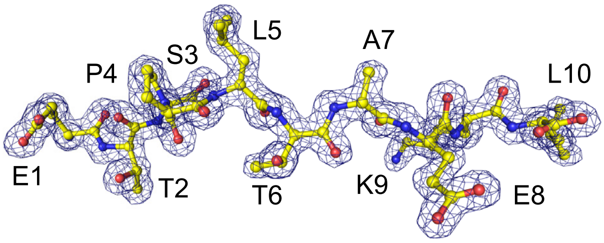


Fig. S7: Map of the peptide experimental electron density in the new X-ray structure shown in Fig. 4A (Resolution 1.6Å).

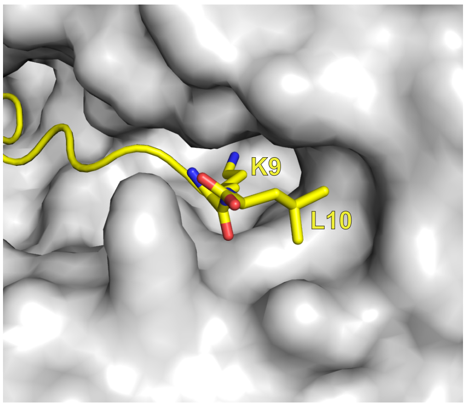


Fig. S8: Surface view of the new structure of HLA-A68:01 in complex with the C-terminally extended peptide (ETSPLTAEKL).

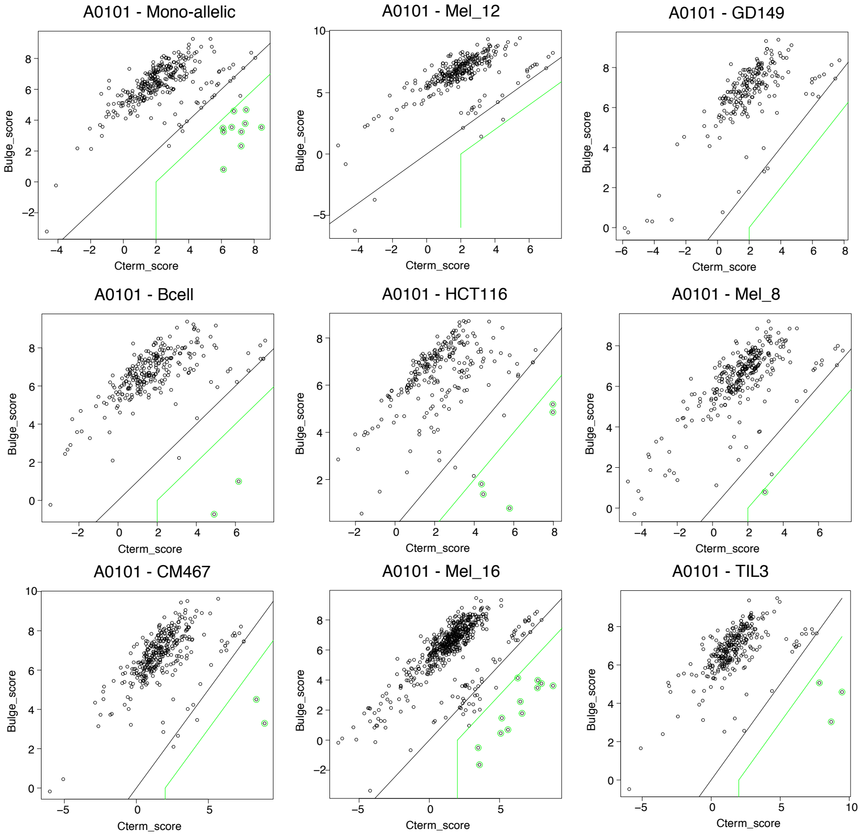


Fig. S9: Bulge versus C-terminal extension scores of peptides assigned to HLA-A0101 in the different samples where this allele is present.

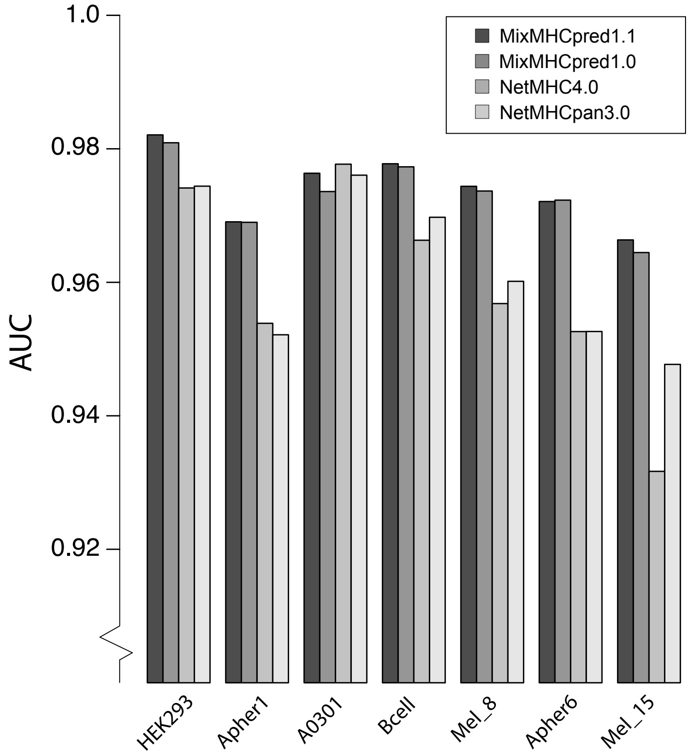


Fig. S10: AUC values for different predictors when re-predicting HLA peptidomics data from samples containing HLA-A03:01 or HLA-A68:01 alleles (same cross-validation study as in Fig. 5A).

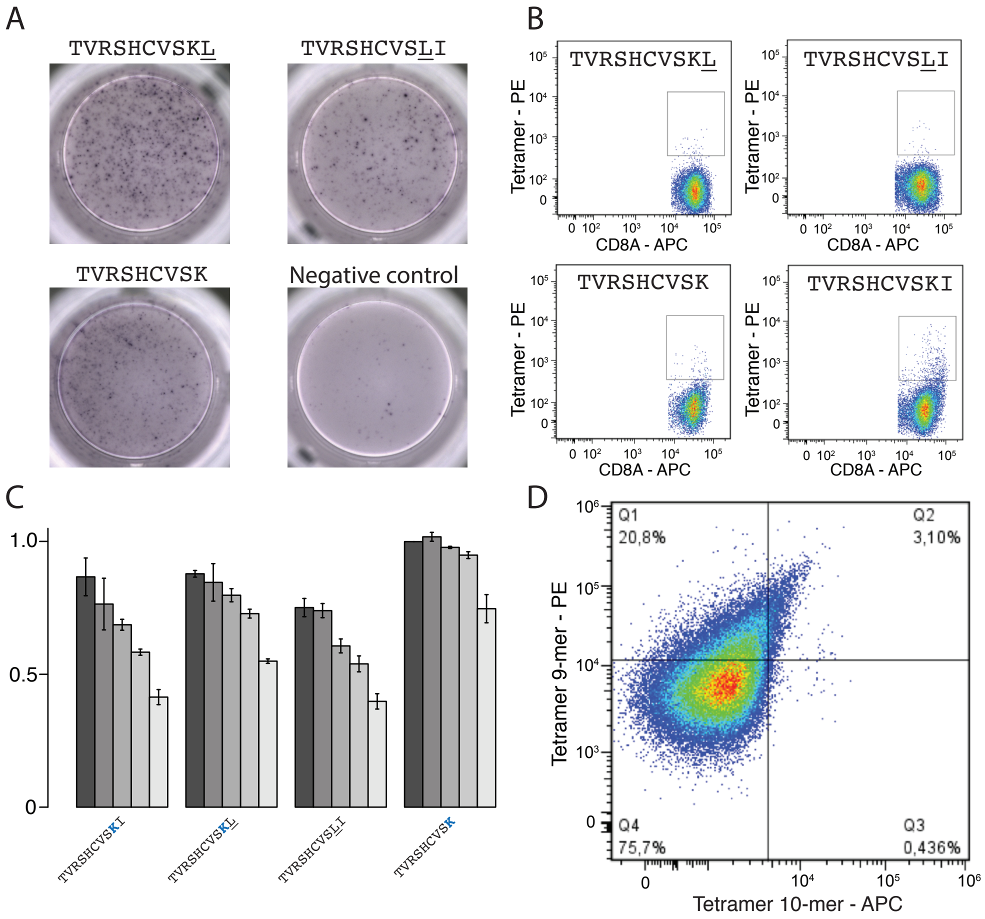


Fig. S11: **A:** Results of the IFNγ-ELISpot with the P10 mutant (TVRSHCVSKL), the P9 mutant (TVRSHCVSLI) and the 9-mer (TVRSHCVSK). Negative control consists of absence of peptide. **B:** Tetramer analysis with the P10 mutant (TVRSHCVSKL), the P9 mutant (TVRSHCVSLI), the 9-mer (TVRSHCVSK) and the initial 10-mer ((TVRSHCVSKI, repeats of Fig. 5D), gated on the CD8+. **C:** Stability analysis of the C-terminally extended CMV epitope (TVRSHCVSKI), P10 mutant, P9 mutant and the 9-mer in complex with HLA-A03:01. **D:** Cross-reactivity of TCRs binding to 10-mer (TVRSHCVSKI) with the 9-mer (TVRSHCVSK) epitope.

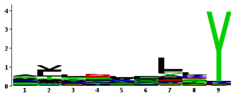
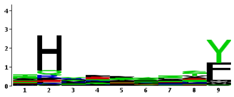
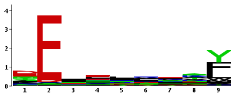
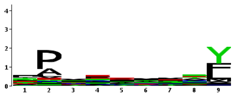
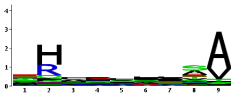
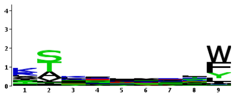
HLA-A29:02	
HLA-B15:18	
HLA-B18:01	
HLA-B35:01	
HLA-B39:06	
HLA-B57:01	

Fig. S12: Examples of alleles, together with the 9-mer motifs, for which no C-terminal extensions were observed and that do not only show preference for hydrophobic amino acids at the C-terminal anchor residue.

SI Tables

Table S1: List of X-ray structures available for alleles studied in this work. The last column shows the distance between C-alpha atoms of residues 80 and 143 (residue numbering following X-ray structures).

Allele	PDB	Resolution	Ligand	Residue at 80	Residue at 143	Distance [Å]
A01:01	3bo8	1.8	EADPTGHSY	T	T	10.1714
A02:01	2bnq	1.7	SLLMWITGV	T	T	9.99682
A02:03	3ox8	2.1	FLPSDFFPSV	T	T	10.21521
A02:07	3oxs	1.75	FLPSDFFPSV	T	T	10.20377
A03:01	3rl1	2	AIFQSSMTK	T	T	10.06951
A24:02	3vxn	1.95	RYPLTFGWCF	I	T	9.98535
A68:01	4hwz	2.4	AIFQSSMTK	T	T	10.12923
A68:02	4hx1	1.8	SVYDFFVWL	T	T	10.26965
B07:02	5eo0	1.7	RPMTFKGAL	N	T	9.33824
B08:01	1m05	1.9	FLRGRAYGL	N	T	9.28667
B18:01	4xxc	1.43	DELEIKAY	N	T	9.28902
B27:04	5def	1.6	RRKWRRWHL	T	T	9.86348
B27:05	1ogt	1.47	RRKWRRWHL	T	T	10.19104
B35:01	2cik	1.75	KPIVVLHGY	N	T	9.38096
B39:01	4o2e	1.98	SHVAVENAL	N	T	9.40402
B44:02	1m6o	1.6	EEFGRAFSF	T	T	9.74925
B44:03	1n2r	1.7	EEFGRAFSF	T	T	9.8785
B51:01	1e27	2.2	LPPVVAKEI	I	T	10.26281
B57:01	5t6w	1.9	SSTRGISQLW	I	T	9.7761
C03:04	1efx	3	GAVDPLLAL	N	T	9.58623

Table S2: C-terminal extension predictions in *T. gondii* HLA-A02:01 ligands from (22).

	Number of peptides	C-terminal extension	Expected	SD	Z-score
10-mers	22	2	0.13	0.3	6.231
11-mers	25	5	0.09	0.23	21.308
12-mers	23	5	0.05	0.14	34.668

Table S3: Sensitivity of the C-terminal extension predictions with respect to different choices for the thresholds T_1. Only alleles that passed the thresholds are shown.

T_1=1.5	**Sample**	**Allele**	**10-mers**	**C-terminal Extension**	**Expected**	**SD**	**Z-score**
Abelin	A0203	A0203	446	13	4.54	2.09	4.056
Abelin	A0207	A0207	424	17	5.06	2.12	5.629
Ritz	HEK293	A0301	1221	63	4.7	2.08	28.094
Melanoma	Mel_8	A0301	1283	25	13.21	5.04	2.34
Abelin	A0301	A0301	381	51	8.05	2.41	17.811
Melanoma	Mel_15	A0301	4865	48	13.27	4.68	7.414
Abelin	A3101	A3101	193	26	0.4	0.55	46.695
Melanoma	Mel_15	A6801	1944	38	14.18	4.74	5.026
Lausanne	RA957	A6801	4865	137	24.19	7.2	15.676
Abelin	A6802	A6802	357	11	4.36	1.68	3.957
Abelin	B5401	B5401	194	12	6.58	2.39	2.27

T_1=2.5	**Sample**	**Allele**	**10-mers**	**C-terminal Extension**	**Expected**	**SD**	**Z-score**
Abelin	A0203	A0203	446	10	2.31	1.42	5.416
Abelin	A0207	A0207	424	17	2.83	1.66	8.529
Ritz	HEK293	A0301	1221	62	0.09	0.29	212.804
Lausanne	Apher1	A0301	1273	14	2.5	1.65	6.952
Melanoma	Mel_8	A0301	1283	24	2.83	2.23	9.483
Mommen	Bcell	A0301	2521	26	5.49	2.95	6.965
Abelin	A0301	A0301	381	45	1.93	1.39	31.095
MCP	Fibroblast	A0301	859	17	2.25	1.89	7.797
Abelin	A3101	A3101	193	25	0.08	0.23	106.889
Melanoma	Mel_15	A6801	4865	136	7.65	3.25	39.49
Lausanne	RA957	A6801	1944	38	1.66	1.83	19.849
Abelin	A6802	A6802	357	10	1.57	0.96	8.736
Mommen	Bcell	B0702	2521	18	10.17	2.93	2.676
Mommen	Bcell	B2705	2521	17	6.22	2.12	5.093
Melanoma	Mel_15	B2705	4865	37	12.5	3.77	6.501
Abelin	B5401	B5401	194	12	4.56	1.94	3.842

T_1=top 2%	**Sample**	**Allele**	**10-mers**	**C-terminal Extension**	**Expected**	**SD**	**Z-score**
Abelin	A0203	A0203	446	11	2.8	1.46	5.597
Abelin	A0207	A0207	424	17	3.64	1.9	7.037

Ritz	HEK293	A0301	1221	63	0.33	0.59	107.027
Lausanne	Apher1	A0301	1273	15	5.19	2.4	4.082
Melanoma	Mel_8	A0301	1283	25	6.73	3.74	4.891
Mommen	Bcell	A0301	2521	27	11.91	4.23	3.563
Abelin	A0301	A0301	381	48	3.28	1.85	24.142
Melanoma	Mel_15	A0301	4865	48	7.8	3.82	10.535
MCP	Fibroblast	A0301	859	17	4.69	2.76	4.464
Abelin	A3101	A3101	193	25	0.15	0.35	70.09
Lausanne	RA957	A6801	1944	38	7.06	3.12	9.932
Melanoma	Mel_15	A6801	4865	137	16.55	5.7	21.137
Abelin	A6802	A6802	357	11	1.98	1.15	7.837
Abelin	B5401	B5401	194	12	5.55	1.96	3.284

Table S4: Crystallographic data collection and refinement statistics (* Values in parentheses correspond to the highest resolution shell).

Data Collection	
PDB ID	**6EI2**
Protein/Ligand	HLA-A*68:01/β2M/peptide
Space group	$P2_12_12_1$
Cell dimensions: a, b, c (Å)	59.06 80.25 110.88
α, β, γ (deg)	90.00 90.00 90.00
Resolution* (Å)	1.61 (1.70-1.61)
Unique observations*	68846 (9873)
Completeness* (%)	99.9 (99.5)
Redundancy*	6.6 (6.6)
Rmerge*	0.094 (0.602)
I/ σI*	11.3 (2.8)
Refinement	
Resolution (Å)	1.61
R_{work} / R_{free} (%)	15.9 / 18.1
Number of atoms (protein/other/water)	3165 / 30 / 442
B-factors ($Å^2$) (protein/other/water)21.85	19.95 / 33.69 / 30.45
r.m.s.d bonds (Å)	0.016
r.m.s.d angles (°)	1.714
Ramachadran Favoured (%)	97.33
Allowed (%)	2.67
Disallowed (%)	0.00

SI Datasets

Dataset S1: List of HLA peptidomics samples considered in this work with HLA typing information. Melanoma (2); MCP (1); Mommen (3); Ritz (4); Lausanne_2016 (7); Abelin (5); Hilton (6).

Dataset S2: Results of predictions of C-terminal extensions for 10-mer and 11-mer peptides. The first two columns indicate the sample of origin. Column 3 indicates the allele to which one 9-mer motif was annotated. Column 4 shows the total number of 10-mers with one-residue long C-terminal extension, respectively 11-mers with two-residue long C-terminal extensions. Column 5 shows the number of C-terminal extensions predicted for the allele in Column 3. Column 6 and 7 show the expected number and standard deviation over 100 realizations of the predicted 10-mer peptidome assuming only bulges (SI Methods). Column 8 shows the Z-score. NA stands for either motifs that could not be annotated, or cases where the number of 10-/11-mer ligands associated with the allele in column 3 did not pass our threshold (mainly B08:01 and HLA-C alleles which poorly bind longer peptides).

Dataset S3: List of 10-/11-mer peptides predicted to display C-terminal extensions for samples and alleles for which the number of C-terminal extensions passed our thresholds. These peptides were used in Fig. 2 and pooled together to define the final motifs for each allele (Fig. S2, last column).

Dataset S4: List of 10-mer peptides that had similar scores for the bulge and the C-terminal extension models in all cases predicted to display C-terminal extensions (red circles in Fig. S1).

References

1. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M (2015) Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics* 14(3):658–673.

2. Bassani-Sternberg M, et al. (2016) Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* 7:13404.

3. Mommen GPM, et al. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc Natl Acad Sci USA* 111(12):4507–4512.

4. Ritz D, et al. (2016) High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics* 16:1570–1580.

5. Abelin JG, et al. (2017) Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 46(2):315–326.

6. Hilton HG, et al. (2017) The Intergenic Recombinant HLA-B∗46:01 Has a Distinctive Peptidome that Includes KIR2DL3 Ligands. *Cell Rep* 19(7):1394–1405.

7. Bassani-Sternberg M, et al. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol* 13(8):e1005725.

8. Bassani-Sternberg M, Gfeller D (2016) Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions. *J Immunol* 197(6):2492–2499.

9. Garboczi DN, Hung DT, Wiley DC (1992) HLA-A2-peptide complexes: refolding and crystallization of molecules expressed in Escherichia coli and complexed with single antigenic peptides. *Proc Natl Acad Sci USA* 89(8):3429–3433.

10. Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):125–132.

11. Evans P (2017) *SCALA - scale together multiple observations of reflections.* (MRC Laboratory of Molecular Biology).

12. McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ (2005) Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr* 61(Pt 4):458–464.

13. Perrakis A, Morris R, Lamzin VS (1999) Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* 6(5):458–463.

14. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126–2132.

15. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53(Pt 3):240–255.

16. Painter J, Merritt EA (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr* 62(Pt 4):439–450.

17. Robinson J, et al. (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 43(Database issue):D423–31.

18. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.

19. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33(1):1–22.

20. Andreatta M, Nielsen M (2016) Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32(4):511–517.

21. Nielsen M, Andreatta M (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* 8(1):33.

22. McMurtrey C, et al. (2016) Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. *Elife* 5:246.