The Journal of Immunology

The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference

Thomas Trolle, Curtis P. McMurtrey, John Sidney, Wilfried Bardet, Sean C. Osborn, Thomas Kaever, Alessandro Sette, William H. Hildebrand, Morten Nielsen and Bjoern Peters

J Immunol 2016; 196:1480-1487; Prepublished online 18 January 2016;
doi: 10.4049/jimmunol.1501721
http://www.jimmunol.org/content/196/4/1480

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
1451 Rockville Pike, Suite 650, Rockville, MD 20852

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference

Thomas Trolle,* Curtis P. McMurtrey,† John Sidney,‡ Wilfried Bardet,† Sean C. Osborn,† Thomas Kaever,‡ Alessandro Sette,‡ William H. Hildebrand,† Morten Nielsen,*,§ and Bjoern Peters‡

HLA class I–binding predictions are widely used to identify candidate peptide targets of human CD8$^+$ T cell responses. Many such approaches focus exclusively on a limited range of peptide lengths, typically 9 aa and sometimes 9–10 aa, despite multiple examples of dominant epitopes of other lengths. In this study, we examined whether epitope predictions can be improved by incorporating the natural length distribution of HLA class I ligands. We found that, although different HLA alleles have diverse length-binding preferences, the length profiles of ligands that are naturally presented by these alleles are much more homogeneous. We hypothesized that this is due to a defined length profile of peptides available for HLA binding in the endoplasmic reticulum. Based on this, we created a model of HLA allele–specific ligand length profiles and demonstrate how this model, in combination with HLA-binding predictions, greatly improves comprehensive identification of CD8$^+$ T cell epitopes. ***The Journal of Immunology*****, 2016, 196: 1480–1487.**

The identification of HLA class I (HLA-I)–restricted epitopes recognized by human T cells has benefited greatly from the development of reliable binding-prediction tools for different HLA molecules. For a given HLA molecule and a given peptide length, several benchmarks showed that binding predictions correlate well with measured binding affinities (1–4) and that peptides with high predicted affinity contain the vast majority of T cell epitopes (5, 6). This has allowed comprehensive mapping of epitopes in entire pathogens by focusing testing on a manageable number of top-predicted binders, saving vast amounts of resources (7–12).

However, it is not clear how peptides of different lengths should be treated in such prediction-guided approaches. Traditionally, there has been a focus on 9mer peptides when mapping HLA-I–restricted T cell epitopes, but peptides of other lengths can bind HLA-I molecules (13) and elicit immune responses, as evidenced by multiple dominant epitopes of length 8, 10, and 11 (14–17), and occasionally much longer peptides, up to length 15 (17–19). MHC-binding predictions for peptides of noncanonical lengths are available, but in many cases their predictions are extrapolated from 9mer data (20) and will predict a roughly similar affinity range for peptides of any given length. Thus, when considering all peptides of length 8–15 that have predicted affinities stronger than a given threshold, the number of peptide candidates would go up drastically compared with when only 9mers are considered.

The length distribution of T cell epitopes should largely reflect the length distribution of peptide ligands that are presented to T cells by MHC molecules. In turn, the MHC ligand length distribution should reflect at least two factors: the MHC allele–specific ability to bind peptides of different lengths and the MHC allele–independent availability of peptides of different lengths for binding to MHCs, which is shaped by the Ag- processing and -presentation machinery preceding MHC binding, such as proteasomal cleavage and TAP transport (21). The goal of this study was to determine the length distribution of MHC class I (MHC-I)–restricted ligands, to what degree this length distribution is allele specific, and how this knowledge can be used to optimize MHC-I–binding predictions for CD8$^+$ T cell epitope mapping.

Materials and Methods

MHC-binding assays

Quantitative in vitro competitive binding assays using purified MHC-I and an [^{125}I]labeled standard probe peptide were performed using a mAb capture assay platform, essentially as described previously (22). Briefly, 0.1–1 nM radiolabeled peptide was coincubated at room temperature with 1 μM to 1 nM purified MHC-I in the presence of a mixture of protease inhibitors and 1 μM human β_2-microglobulin (Scripps Laboratories). Following a 2-d incubation, MHC-I–bound radioactivity was determined by capturing MHC-I/peptide complexes on W6/32 (anti–HLA-I mAb)–coated LUMITRAC 600 plates (Greiner Bio-one, Frickenhausen, Germany) and measuring bound radioactivity using the TopCount (Packard Instrument, Meriden, CT) microscintillation counter. The concentration of peptide

*Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark; †Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104; ‡Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037; and §Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, San Martín, B 1650 HMP Buenos Aires, Argentina

ORCIDs: 0000-0003-0762-2198 (T.T.); 0000-0002-0987-405X (J.S.); 0000-0003-4709-6356 (T.K.).

Received for publication July 31, 2015. Accepted for publication December 13, 2015.

This work was supported in whole or in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract HHSN272201200010C. M.N. is a researcher at the National Scientific and Technical Research Council-Argentina.

The datasets presented in this article have been submitted to the Immune Epitope Database and Analysis Resource (http://www.iedb.org/subID/1000685) under accession number 1000685.

Address correspondence and reprint requests to Dr. Bjoern Peters, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037. E-mail address: bpeters@lji.org

The online version of this article contains supplemental material.

Abbreviations used in this article: AP, peptides available for binding to MHC; ER, endoplasmic reticulum; HLA-I, HLA class I; IEDB, Immune Epitope Database; MHC-I, MHC class I; sHLA, soluble HLA.

yielding 50% inhibition of the binding of the radiolabeled peptide was calculated. Under the conditions used, where [label] < [MHC] and $IC_{50} \geq$ [MHC], the measured IC_{50} values were reasonable approximations of the true K_D values (23, 24). Each competitor peptide was tested at six concentrations covering a 100,000-fold dose range and in three or more independent experiments. As a positive control, the unlabeled version of the radiolabeled probe was also tested in each experiment.

Twenty-seven HLA alleles representative of the most frequent HLA-I specificities in the human population were considered (14). Their peptide length preferences were determined by testing panels of combinatorial peptide libraries. Each library contained peptides of a uniform fixed length. Libraries ranged from 8 to 15 aa residues. Furthermore, each library, for each length, was defined by a fixed C-terminal residue that was either I, K, or F. For each HLA allele, the libraries with the C-terminal residue that gave the highest affinity were chosen to determine the length preference for that HLA allele, to take into account the different C-terminal–binding preferences of different alleles. It should be noted that the length-binding profile for a given HLA molecule is estimated from a series of independent peptide libraries of different lengths. As a result of variations in library synthesis and binding assay variability, substantial noise is to be expected in the measured binding values, leading to some degree of nonmonotonic behavior of the length-profile curves.

Cell lines and production of HLA complexes for elution studies

HeLa cells were cultured and propagated in DMEM with 10% FCS. HeLa cells were stably transfected with a soluble form of HLA-A*01:01, HLA-A*02:01, HLA-A*24:02, HLA-B*07:02, and HLA-B*51:01, as previously described (25, 26). Soluble HLA (sHLA) constructs were generated with a truncation at the *trans*-membrane and cytoplasmic domains with the addition of a VLDLr purification tag and cloned into pcDNA3.1. HeLa cells were stably transfected with the sHLA by electroporation, followed by drug selection and subcloning. sHLA-producing clones were identified using a capture ELISA with the pan-class I Ab W6/32 as the capture Ab and a β_2-microglobulin Ab as a detector. sHLA-producing clones were expanded and seeded into a hollow fiber bioreactor where sHLA-containing supernatant was collected. sHLA was purified from the supernatant using affinity chromatography with an anti-VLDLr Ab. Complexes were eluted from the column in 0.2 M acetic acid and immediately processed for isolation of the peptide ligands.

Elution of naturally presented MHC-I ligands

Eluted MHC-I ligand datasets were generated for five common HLA alleles: HLA-A*01:01, HLA-A*02:01, HLA-A*24:02, HLA-B*07:02, and HLA-B*51:01. The elution of the peptide ligands was done, as described in detail previously (25, 26). Briefly, the peptide ligands were eluted from the complex with an acid boil, and peptide ligands were separated from the α- and β-chains with a 3-kDa cut-off filtration using a Millipore 3-kDa molecular mass limit ultrafiltration membrane (Merck Millipore). A 3-kDa cutoff corresponds to the cutoff of a peptide ∼ 30 aa. Because we use a maximum of 15mers in our predictions (or 1.5 kDa, which is half the molecular mass limit of the filter), we should have little to no bias in the number of 15mer peptides. However, the filtration efficiency may not be identical for all peptides between 8 and 15 residues long, and longer peptides may be underrepresented. Peptide pools were initially separated into ∼40 fractions using pH10 RP HPLC. Peptide-containing fractions (fractions 22–60) were analyzed individually with liquid chromatography mass spectrometry. Nano-scale liquid chromatography was performed with an Eksigent nano-LC-4000 with an Eksigent autosampler (AB Sciex). Fractions were loaded on a C18 trap (350 μm [i.d.] × 0.5 mm long; ChromXP) and desalted before separated with a gradient elution into a ChromXP C18 separation column (75 μm [i.d.] × 15 cm long). Column media consisted of 3-μm particles with 120-Å pores. The elution mobile phase consisted of two linear gradients using solvent A (98% water, 2% acetonitrile, 0.1% formic acid) and solvent B (95% acetonitrile, 5% water, 0.1% formic acid): 10–40% B for 70 min and then 40–80% for 10 min. Eluate was ionized with a NanoSpray III ion source (AB Sciex), and MS1 and MS2 fragments were obtained in IDA mode using an AB Sciex TripleTOF 5600 System, as described previously (26).

Peptide sequences were derived from the resulting fragment spectra using PEAKS 7.0 (Bioinformatics Solutions), with a precursor ion tolerance of 50 ppm and a product ion tolerance of 0.05 Da. The National Center for Biotechnology Information nonredundant database with *Homo sapiens* taxonomy was used. Posttranslational modifications consisting of N-terminal acetylation, deamidation of Asn and Gln, oxidation of Met, His, Trp, sodium adducts of Asp, Glu, C terminus, and the pyroglutamate derivative of glutamic acid were searched as variable modifications. Positive-sequence assignments were determined at a 1% false-discovery rate using the decoy fusion approach (27). Most positive peptide identifications were within 25 ppm of theoretical mass. Any peptides from the sHLA construct (HLA α-chain and β_2-microglobulin) and a contaminating protein TERA were removed from the data, because these are likely not ligands. Peptides resulting from D|P, D|A, and D|T cleavages were also removed because these are peptides likely created from acid hydrolysis of larger ligands.

Corrected elution datasets

In addition to the eluted peptide dataset described above, two corrected datasets were created. The first corrected dataset was obtained by filtering out ligands that did not conform to the canonical MHC-binding motif of the given allele to remove likely contaminants. Binding affinities for all eluted peptides were predicted using *NetMHCpan-2.8* (28, 29). In addition to binding affinities in nM, *NetMHCpan* returns a percentage rank score for each peptide, indicating how strong a peptide's binding affinity is compared with a large pool of naturally occurring peptides. A rank score of 10% means a peptide falls within the top 10% strongest binders among the pool of naturally occurring peptides. The standard *NetMHCpan* rank score is based on predicted binding affinities of 9mer peptides only. In this study we extend this and calculate the rank score compared with pools of peptides matching the length of the query peptide. This was done to remove any artificial bias in the rank scores imposed by the use of the extrapolation model from 9mer data mentioned earlier (20). A rank score of 10% was used as the threshold for defining a binder. Note that this is a very tolerant threshold, because earlier studies demonstrated that the vast majority of known $CD8^+$ epitopes are predicted to bind to the restricting MHC molecules with a rank score ≤ 2% (5, 30).

The second and final corrected dataset took into account that some peptides are degraded before mass spectrometry identification, leading to the recovery of fragments of the original full-length ligand. To identify such peptide-degradation events, we mapped all predicted nonbinders back to their source proteins and extended the peptides in silico with up to five amino acids at either the N or C terminus, up to a maximum of length 15, while searching for potential predicted binders. If a high-affinity binder, defined using a rank score threshold of 2%, was discovered in this process, it was substituted for the nonbinding fragment. We decided to use a more stringent rank score threshold of 2% in this case, because we were only interested in including extended peptides that had a very high likelihood of binding their given HLAs. In contrast, the previous filtering used a 10% rank score threshold, because there our goal was to exclude only peptides that had a very low probability of binding their given HLAs. It should be noted that similar results were obtained using an unfiltered data set, as well as data sets in which the thresholds for identifying peptide-degradation events were 1 and 10% (data not shown).

Reconstruction of the peptide-length profile available for binding to MHC

Assuming that peptides available for binding to MHC (AP) can be approximated by a Boltzmann distribution, the ratio of the number of peptides of a given length L bound to MHC, $P_{MHC}(L)$, compared with the number of peptides bound of length 9, P_{MHC} (9), is determined by the ratio of peptides available for binding of length L, AP(L), and those of length 9, AP (9), and the difference in binding free energy of these peptides. In our assay conditions, $\log(IC_{50})$ approximates binding free energy; thus, we can write:

$$\frac{P_{MHC}(L)}{P_{MHC}(9)} = \frac{AP(L)}{AP(9)} \cdot e^{-\beta \log\left(\frac{IC_{50}(L)}{IC_{50}(9)}\right)} = \frac{AP(L)}{AP(9)} \cdot \left(\frac{IC_{50}(9)}{IC_{50}(L)}\right)^{\beta}$$

where β is a positive unknown parameter, and the IC_{50} values and bound length distributions (P_{MHC}) are known based on the affinity measurements and elution experiments for five HLA alleles. Thus, we can fit β and the unknown available peptide length distribution by minimizing the squared distance between measured and calculated $P_{MHC}(L)/P_{MHC}(9)$ values.

Benchmark data

A T cell epitope evaluation dataset was retrieved from the Immune Epitope Database (IEDB) (31). Because it is of particular importance for our study that the optimal-length peptide epitope was identified, we restricted ourselves to multimer/tetramer assays. Peptides between the lengths of 8 and 5 aa were included, in which the tetramer used was 1 of the 27 IEDB reference HLA alleles. Source proteins for each epitope were downloaded from GenBank using the accession number annotated in the IEDB. A total of 535 T cell epitopes matching our selection criteria was downloaded.

These epitopes were filtered to remove predicted nonbinders using the same approach as for the elution dataset, reducing the dataset by 42 epitopes. Finally, five epitopes were removed from the dataset because they could not be mapped to their annotated source protein. Because a majority of the epitopes (59%) in the dataset were HLA-A*02:01 restricted, we created a balanced dataset: 20 epitopes for each HLA allele were selected at random. If there were <20 epitopes for an allele, all of the epitopes were included in the balanced dataset. Binding-affinity predictions were generated for each T cell epitope, as well as for all overlapping 8–13mers in the source proteins, using *NetMHCpan*. Because no 14–15mers were present in the final datasets, these lengths were excluded from the benchmark.

In addition to the T cell epitope dataset, three recently published MHC-I ligand datasets by Granados et al. (32), Marcilla et al. (33), and Thommen et al. (34) were retrieved from the IEDB (IEDB reference IDs 1027559, 1027269, and 1027076). The datasets were filtered to remove predicted nonbinders, as previously described, removing 48, 81, and 27 peptides, respectively. Binding-affinity predictions were generated for overlapping peptides in the source proteins, as described for the IEDB dataset. Overlapping 8–11mers were included in the Granados et al. (32) benchmark, whereas 8–13mers were included in the Marcilla et al. (33) and Thommen et al. (34) benchmarks, reflecting the ligand lengths found in each dataset.

Adjusting binding-affinity predictions for length preference

NetMHCpan predictions were adjusted to result in a distribution of predicted binders that reflect the length profile of ligands for a given HLA allele. This was achieved by dividing the predicted rank scores for each peptide of length L by the relative frequency at which peptides of this length are found bound to the MHC:

$$\text{Adjusted peptide rank score} = \text{peptide rank score}/(P_{MHC}(L)/\ P_{MHC}(9))$$

where the ratio ($P_{MHC}(L)$ / P_{MHC} (9)) was estimated in an MHC-specific manner using the model described above. This meant that 9mer predictions were left unchanged, whereas predictions for all other lengths were modified depending on the peptide length and MHC restriction. Peptide lengths that were enriched compared with 9mers received enhanced length-corrected rank scores, and vice versa for peptide lengths that were less preferred compared with 9mers for the given MHC. For peptides with MHC restrictions outside of our panel of 27 HLA-I alleles, an averaged MHC binding-length preference was used instead. This averaged length distribution (Supplemental Table I) was determined by calculating the geometric mean of the 27 HLA-I length preferences. Finally, the corrected rank scores were transformed back to binding-affinity values using the underlying percentile affinity distribution for the given allele, which corresponds to an effective IC_{50} value.

Results

MHC binding-length preference

We set out to determine the preferred length of peptides binding to a panel of 27 human MHC-I alleles making up commonly expressed molecules in the human population. Affinities of combinatorial libraries of peptides with different lengths were determined and normalized to the affinity of a library of 9mer peptides, as described in *Materials and Methods*. The resulting length preferences are shown in Supplemental Table I, and Fig. 1 depicts data for five alleles that are representative of the spectrum of observed patterns: HLA-A*02:01, HLA-A*24:02, and HLA-B*07:02 showed the typical preference for 9mer peptides, HLA-A*01:01 had a preference for 10mers, and HLA-B*51:01 had a preference for 8mers. These data confirm that there are MHC allele–specific differences in binding affinity for peptides of different length.

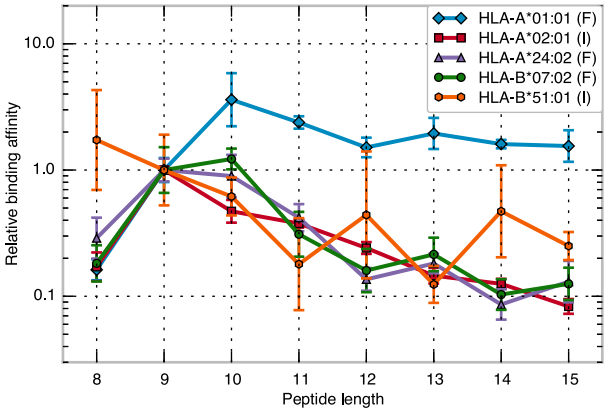

FIGURE 1. Peptide binding-length preference for five common HLA alleles. The length preference for each HLA was determined by measuring the binding affinity of a series of fixed C-terminal combinatorial libraries of different length. Three series were tested, with I, K, or F at the C-terminal. The series with the strongest binding affinity was selected to represent the HLA allele. The selected series is denoted in the parentheses in the inset. IC_{50} binding affinities for each length were calculated as geometric means of three to six experiments. The relative binding affinities plotted were calculated as $IC_{50}(9)/IC_{50}(L)$, where *L* is the peptide length. Error bars indicate SEs of the geometric means.

Length distribution of naturally presented MHC ligands

To examine whether allele-specific differences in binding-length preferences have an impact on the length distribution of naturally processed MHC ligands, we performed peptide-elution studies on the five alleles for which binding data are shown in Fig. 1. Elution of peptide ligands from secreted MHCs and ligand identification by mass spectrometry were performed as described previously (25, 26) and in *Materials and Methods*. To remove contaminants and to control for the effect of peptide degradation, the dataset was further corrected with the help of binding predictions. The final datasets (Supplemental Table II) contain an average of 3197 identified peptides/allele, ranging from 1275 for HLA-B*51:01 to 4456 for HLA-A*02:01 (Table I). These datasets were submitted to the IEDB and provide a large publicly available dataset of naturally presented MHC ligands with well-defined restrictions for different alleles gathered with a consistent methodology. The data can be accessed at http://www.iedb.org/subID/1000685.

Next, we examined the length distribution of the ligands identified in the elution studies. Fig. 2 shows, for each allele, the number of ligands identified for a given length normalized by the number of peptides identified for the HLA at length 9. Raw peptide counts and 9mer-normalized values are shown in Supplemental Table II. Strikingly, all of the HLAs presented more

Table I. Peptides identified in elution studies

Allele	Preferred Peptide Length for Binding[a]	Fixed C-Terminal Library Selected	No. Eluted Ligands in Final Dataset	Most Frequent Eluted Ligand Length[b]
HLA-A*01:01	10	F	2992	9
HLA-A*02:01	9	I	4456	9
HLA-A*24:02	9	F	3949	9
HLA-B*07:02	9	F	3314	9
HLA-B*51:01	8	I	1275	9

[a]The preferred peptide length identified from the MHC-I binding length preferences shown in Fig. 1.
[b]The preferred peptide length identified from the eluted MHC-I ligand profiles shown in Fig. 2.

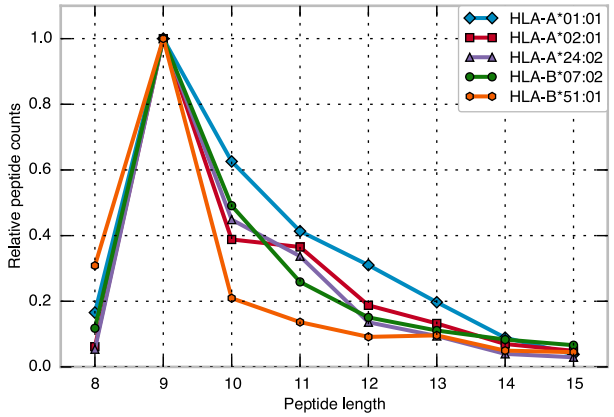

FIGURE 2. Length profiles of naturally presented peptides for five HLA molecules. Large datasets of HLA-I ligands were determined by the elution of ligands from secreted HLAs, followed by mass spectrometry identification of the peptide sequences. The number of ligands of each length was totaled from these ligand datasets. The *y*-axis indicates the number of ligands identified for a given length normalized by the number of peptides identified for the HLA at length 9.

9mers than peptides of any other lengths. This observation includes HLA-A*01:01 and HLA-B*51:01, which preferred binding 10mers and 8mers over 9mers, respectively. However, compared with HLA-A*02:01, HLA-A*24:02, and HLA-B*07:02, all of which prefer the binding of 9mers, HLA-A*01:01 presented the most 10mer ligands, and HLA-B*51:01 presented the most 8mers. This suggested that the HLA allele–dependent length preferences observed in the binding assay impacted the length distribution of naturally presented ligands in an allele-specific fashion, but that other factors dampened the allele-specific effects and led to a predominant presentation of 9mer peptides.

Length profiles of peptides available for MHC binding

We wanted to test the hypothesis that the discrepancy between allele-specific peptide length–binding preferences and naturally presented ligand repertoires was due to a fixed-length distribution of peptides available for binding to MHCs. To do this, we constructed a simple mathematical model that assumed that the available peptide-length distribution was the same for each MHC allele and that the observed eluted MHC ligand–length profiles (Fig. 2) were the result of this available peptide-length profile, in conjunction with the MHC-I binding–length preferences of each allele. Based on this, we calculated the available peptide-length profile (Fig. 3, Table II). By far, the most frequent peptide length available for binding based on this model was 9, which was expected given that this peptide length dominated in the eluted ligand profile for all alleles, even for HLA-A*01:01 and HLA-B*51:01, which preferred to bind peptides of different lengths.

To evaluate whether the available peptide-length profile that we calculated based on data from five alleles was robust, a leave-one-out experiment was carried out. Iteratively, each HLA allele was excluded from the training data, an available peptide-length profile was calculated using the remaining alleles, and the measured and predicted eluted peptide-length profiles for the excluded HLA alleles were compared. Comparisons of the predicted and measured MHC ligand–length profiles are shown in Fig. 4. For all alleles, there was an excellent fit of predicted and measured profiles (average root-mean-square deviation = 0.057 ± 0.021). This demonstrates that we are able to estimate one common available peptide-length profile, which, when combined with allele-specific binding preferences, is able to explain the differences between our five naturally processed MHC ligand–length profiles.

Benchmarking length profile–adjusted MHC-binding predictions

Next, we tested whether our modeled peptide-length distributions of naturally presented peptides could be used to improve the prediction of T cell epitopes and naturally processed MHC-I ligands. Rather than predicting candidate peptides based on binding affinity alone, we added a correction factor, resulting in a length-adjusted binding prediction for each peptide (see *Materials and Methods*). As shown in Supplemental Fig. 1, choosing peptides based on length-adjusted predictions resulted in a length distribution of peptides that mimicked that of naturally eluted ligands.

We evaluated the performance of the length-adjusted binding-affinity predictions on one T cell epitope dataset and three MHC-I ligand datasets. The T cell epitope dataset was retrieved by querying the IEDB for peptides that were recognized by human T cells in tetramer-staining assays, which we considered most reliable to identify exact epitopes. This dataset was further balanced to not overrepresent commonly studied alleles, such as HLA-A*02:01. The resulting dataset contained 185 epitopes, ranging in length from 8 to 13 residues (Supplemental Table III). These epitopes were considered positives, whereas all other 8–13mers from the same proteins were considered negatives. Peptides were pooled and sorted by predicted binding affinity (from strongest to weakest). This sorted peptide list was then used to determine the number of epitopes identified versus the number of peptides tested. Plots were created using three approaches for predicting T cell epitopes: pure MHC-binding predictions considering peptides of length 8–13 equally, MHC-binding predictions for 9mer peptides only, and the newly developed length-adjusted MHC-binding predictions. Our goal was to compare our novel length-correction approach with two other prediction strategies that are currently used: predicting epitopes of multiple lengths and treating each length equally and predicting epitopes for a single, optimal length. We opted to use 9 as the optimal length for all alleles because, for each allele in our study, this was the most common ligand length found. From the plots (Fig. 5, top left panel), it was apparent that the first approach of considering all peptide lengths equally had the worst performance, as, for example, approximately twice the number of peptides had to be considered to identify 60% of the epitopes in the benchmark. The other two approaches had very similar performances when considering the top 0.5% of peptides. However when the goal is to identify ≥80%

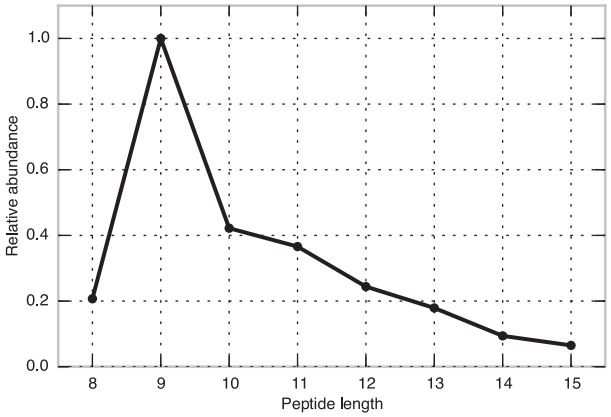

FIGURE 3. Model fit of the available peptide-length profile. The available peptide-length profile was fitted using MHC ligand–length profiles and HLA binding–length preferences for HLA-A*01:01, HLA-A*02:01, HLA-A*24:02, HLA-B*07:02, and HLA-B*51:01, as described in *Materials and Methods*. The optimal value for β associated with the fit was 0.30.

Table II. Fitted AP–length profile

Length	AP[a]
8	0.207
9	1.000
10	0.422
11	0.366
12	0.244
13	0.179
14	0.094
15	0.065

[a]Fitted AP–length profile using data from the five HLA alleles: HLA-A*01:01, HLA-A*02:01, HLA-A*24:02, HLA-B*07:02, and HLA-B*51:01. The β value associated with this profile was 0.3.

of the epitopes, the length-corrected approach is far superior to the 9mer-only prediction approach, which, by definition, misses epitopes of other lengths. Most importantly, this ability to comprehensively identify epitopes of all lengths comes with no significant additional cost, in contrast to the naive approach of considering all peptide lengths equally.

Next, we queried the IEDB for large-scale MHC-I ligand–elution datasets identified by different groups for which restrictions were determined and that represented peptides of different length. Three datasets were selected: the "Granados MHC-I ligands" (4433 ligands, lengths 8–11) (32), the "Marcilla MHC-I ligands" (2235 ligands, lengths 8–13), (33) and the "Thommen MHC-I ligands" (1041 ligands, lengths 8–13) (34). Note, that all ligand counts are after filtering for predicted nonbinders. Eluted MHC ligands were considered positives, and all other peptides from the same proteins were considered negatives. The results of these three benchmarks (Fig. 5, *upper right* and *lower panels*) were very similar to the IEDB dataset benchmark. Considering peptides of all lengths equally was the worst approach, whereas the 9mer-only approach and the weighted length approach performed equally well when the goal was to discover a subset of the eluted ligands (<50–60%). However, when the goal was to comprehensively predict eluted ligands (>60%), the length-weighted approach was far superior to the other two. Thus, the length-adjusted prediction approach developed in this study performed as well or better than the two other approaches in three independent benchmarks, and it was the most efficient approach for the comprehensive discovery of both epitopes and eluted ligands.

Discussion

MHC-binding predictions have facilitated T cell epitope discovery by narrowing the search space to a manageable number of likely peptide candidates. Compared with approaches that do not use predictions, such as screening overlapping peptides, a downside of the prediction approach was that peptides of noncanonical lengths were missed (35). Naively, one could simply extend binding predictions to peptides of any length and rank all peptides based on their predicted affinity. But, as demonstrated in our study, although peptides of noncanonical lengths might have similar or even better predicted binding affinities than the canonical 9mer peptides, they end up being underrepresented among the naturally presented ligands eluted from MHC molecules and, consequently, are recognized less frequently by T cells. In this study, we explained these similarities between the length profiles of naturally presented peptides by fitting a common, underlying peptide-length distribution. This common length distribution, which we

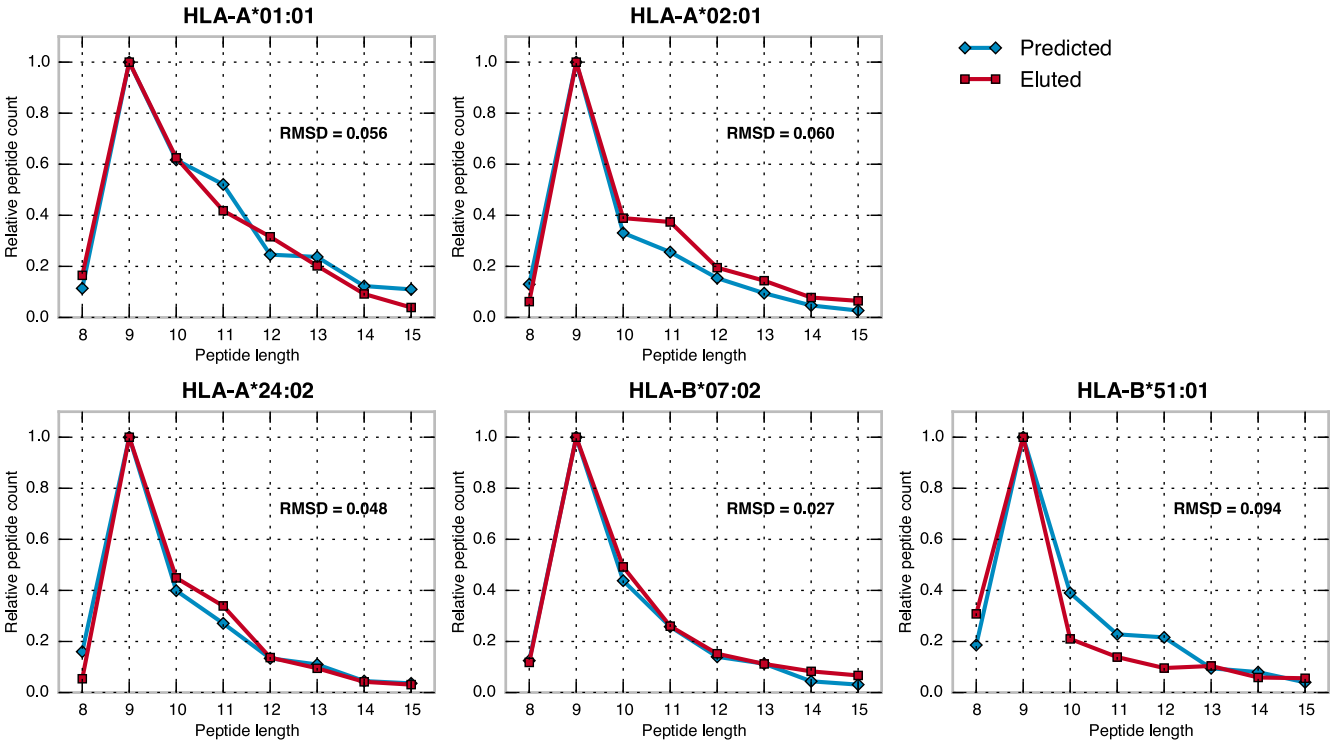

FIGURE 4. Predicted versus measured ligand-length profiles for five HLA molecules. A leave-one-out training was carried out by removing an HLA from the training dataset and then fitting the available peptide-length profile with the remaining four HLAs. The resulting available peptide-length profile was used in conjunction with the removed HLA's binding-length preference (Fig. 1) to predict the removed HLA's ligand–length profile. This predicted length profile was compared with the measured ligand-length profile of the removed HLA. As an example, in the HLA-A*01:01 plot, HLA-A*01:01 data were not used to fit an available peptide length profile (data not shown). This available peptide-length profile was then combined with the HLA-A*01:01 binding-length preference to determine the predicted ligand-length profile (blue line). This profile was compared with the measured HLA-A*01:01 ligand-length profile (red line).

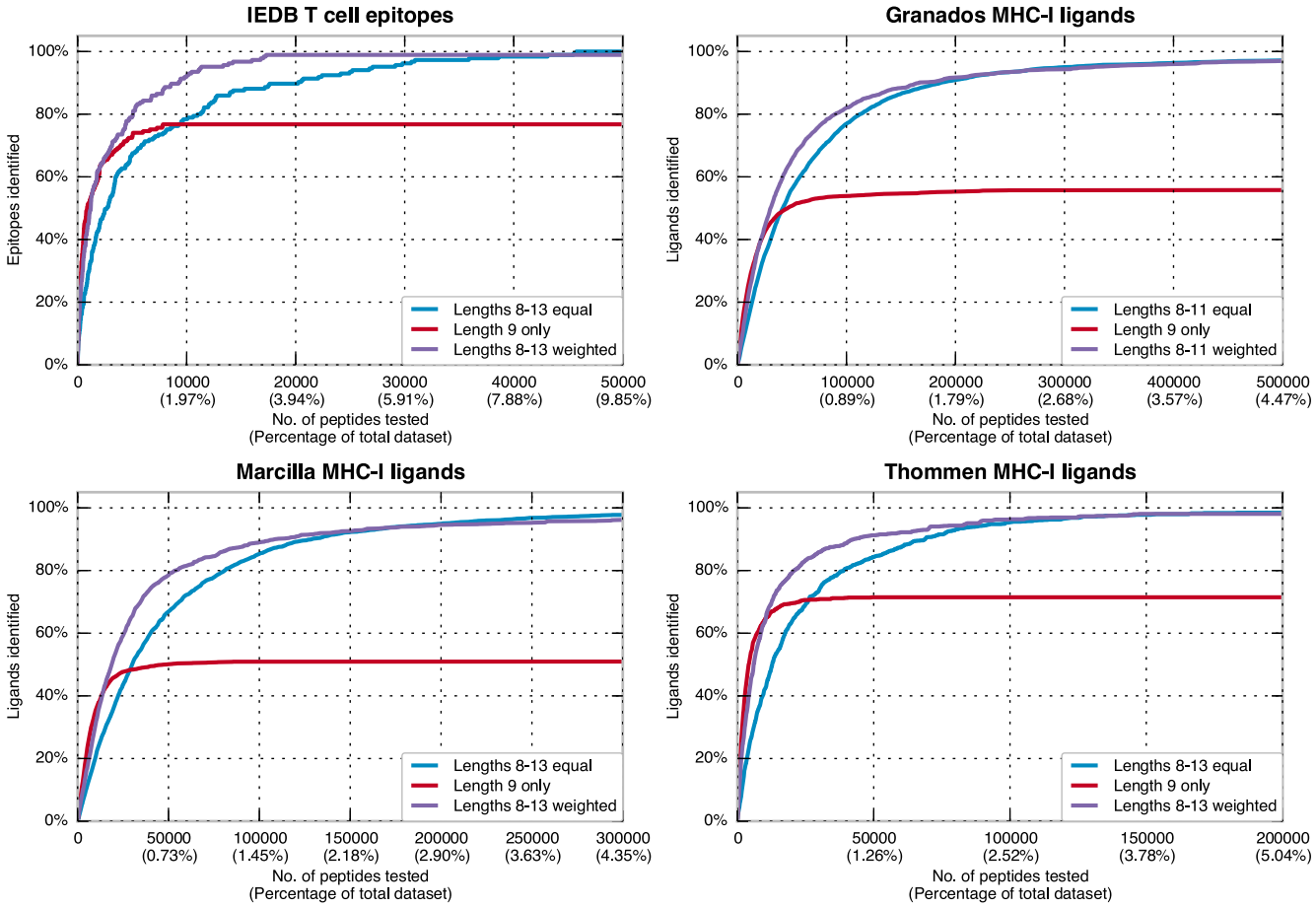

FIGURE 5. Benchmarks of T cell epitope and MHC-I ligand predictions. For each benchmark dataset, source proteins for each of the epitopes/ligands were downloaded and split into overlapping peptides of various lengths. The lengths of the overlapping peptides were determined by the lengths of the epitopes/ligands in the benchmark datasets; 8–13mer overlapping peptides for the IEDB, Marcilla, and Thommen datasets, and 8–11mers for the Granados dataset. For each dataset, three sorted peptide lists were created using the following approaches: predict affinities for all overlapping peptides and rank them based on their predicted IC_{50} value without taking length into account, predict affinities for all 9mer peptides and rank them based on their predicted IC_{50} values (peptides of other lengths are considered noncandidates), and predict length-corrected binding affinities for all overlapping peptides using the novel method described in this article and rank the peptides based on length-corrected predictions. The plots show the number of epitopes/ligand identified by each approach as a function of the number of peptides tested, had the peptides been selected using the sorted lists described above.

call the "available peptide-length profile," could be combined with allele-specific binding-length preferences to yield allele-specific MHC ligand–length profiles. However, at this point, we can only speculate with regard to the major factors behind the available peptide-length profile, as well as their relative contributions.

The available peptide-length profile (Fig. 3) suggests that 9mer peptides are the most common peptides available for MHC binding, with 8mers, 10mers, and longer peptides being far less frequent. We hypothesize that this is due to a combination of three Ag-processing mechanisms: peptide cleavage by the proteasome, transport into the endoplasmic reticulum (ER) by TAP, and peptide trimming by ERAP. Peptide fragments generated by the proteasome are generally 4–7 aa long, with the frequency of fragments longer than that decreasing as length increases (36–38). We see a similar decrease in the available peptide-length profile from length 10 and upward, which could be attributed to the proteasome. TAP was shown to preferentially transport peptides 9–16 aa long (39, 40), which would explain the low frequency of 8mers in the available peptide-length profile. Finally, peptides in the ER are trimmed by ERAP down to a minimum length of 9 aa (41), explaining the 9mer peak in the available peptide-length profile. Thus, although we did not experimentally verify the available peptide-length profile, our fitted profile is consistent with previous knowledge of Ag processing.

Although it is generally accepted that MHC-I molecules bind peptides with lengths 8–11, longer peptide binders were observed previously. These long peptide ligands can bind in a variety of configurations. Structural studies show that the majority of long ligands bind using the P2 and C-terminal amino acids, with a central bulge to accommodate the increased length of the peptide (42, 43). Thus far, this appears to be the primary mechanism by which long MHC-I ligands are bound. A second method for binding was suggested in the literature: a portion of the peptide binds in the groove with a C-terminal or N-terminal extension (44, 45). Indeed, there is a single structure of a peptide binding in an extended configuration (46). Although there are examples of peptide ligands binding with a C-terminal extension, it is unknown how frequently this occurs. In our eluted-ligand dataset, there is some evidence of C-terminal–extended peptides. However, in this study, we separated these putative extended ligands from the canonical binding peptides; the extended ligands will be investigated in detail in future studies.

We developed a simple, yet effective, approach to adjust the predicted binding affinity of a peptide based on its length and the corresponding availability for peptide binding to MHC. We show that this is much more effective for identifying epitopes compared with considering peptides of all lengths equally at any threshold. Also, our novel approach compares favorably with the approach of

considering only 9mer peptides when the goal is to comprehensively identify epitopes. Although, in our benchmark of tetramer-mapped epitopes, this effect was most pronounced when the goal was to identify >80% of all epitopes, it was already apparent when considering >50% of all ligands. Given that the tetramer-mapping dataset is biased because researchers preferably make 9mer peptide-based tetramers, we expect that the estimate based on the elution datasets is more accurate. Thus, we suggest that the length-based weighting of MHC binding predictions introduced in this article should be applied to any study aimed at comprehensively identifying MHC-I–restricted epitopes.

Acknowledgments

For so much more than her invaluable contributions to the performance of the peptide-binding assays described in this article, we dedicate this work to the memory of Carrie Moore (1982–2015).

Disclosures

The authors have no financial conflicts of interest.

References

1. Peters, B., H.-H. Bui, S. Frankild, M. Nielson, C. Lundegaard, E. Kostem, D. Basch, K. Lamberth, M. Harndahl, W. Fleri, et al. 2006. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLOS Comput. Biol.* 2: e65.
2. Lin, H. H., S. Ray, S. Tongchusak, E. L. Reinherz, and V. Brusic. 2008. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol.* 9: 8.
3. Zhang, G. L., H. R. Ansari, P. Bradley, G. C. Cawley, T. Hertz, X. Hu, N. Jojic, Y. Kim, O. Kohlbacher, O. Lund, et al. 2011. Machine learning competition in immunology - Prediction of HLA class I binding peptides. *J. Immunol. Methods* 374: 1–4.
4. Trolle, T., I. G. Metushi, J. A. Greenbaum, Y. Kim, J. Sidney, O. Lund, A. Sette, B. Peters, and M. Nielsen. 2015. Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 31: 2174–2181.
5. Erup Larsen, M., H. Kloverpris, A. Stryhn, C. K. Koofhethile, S. Sims, T. Ndung'u, P. Goulder, S. Buus, and M. Nielsen. 2011. HLArestrictor–a tool for patient-specific predictions of HLA restriction elements and optimal epitopes within peptides. *Immunogenetics* 63: 43–55.
6. Paul, S., D. Weiskopf, M. A. Angelo, J. Sidney, B. Peters, and A. Sette. 2013. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* 191: 5831–5839.
7. Wang, M., K. Lamberth, M. Harndahl, G. Røder, A. Stryhn, M. V. Larsen, M. Nielsen, C. Lundegaard, S. T. Tang, M. H. Dziegiel, et al. 2007. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. *Vaccine* 25: 2823–2831.
8. Sedegah, M., Y. Kim, H. Ganeshan, J. Huang, M. Belmonte, E. Abot, J. G. Banania, F. Farooq, S. McGrath, B. Peters, et al. 2013. Identification of minimal human MHC-restricted CD8+ T-cell epitopes within the *Plasmodium falciparum* circumsporozoite protein (CSP). *Malar. J.* 12: 185.
9. Chiu, C., M. McCausland, J. Sidney, F. M. Duh, N. Rouphael, A. Mehta, M. Mulligan, M. Carrington, A. Wieland, N. L. Sullivan, et al. 2014. Broadly reactive human CD8 T cells that recognize an epitope conserved between VZV, HSV and EBV. *PLoS Pathog.* 10: e1004008.
10. Rajasagi, M., S. A. Shukla, E. F. Fritsch, D. B. Keskin, D. DeLuca, E. Carmona, W. Zhang, C. Sougnez, K. Cibulskis, J. Sidney, et al. 2014. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* 124: 453–462.
11. Robbins, P. F., Y.-C. Lu, M. El-Gamil, Y. F. Li, C. Gross, J. Gartner, J. C. Lin, J. K. Teer, P. Cliften, E. Tycksen, et al. 2013. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* 19: 747–752.
12. Rizvi, N. A., M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, et al. 2015. Caner immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348: 124–128.
13. Chen, Y., J. Sidney, S. Southwood, A. L. Cox, K. Sakaguchi, R. A. Henderson, E. Appella, D. F. Hunt, A. Sette, and V. H. Engelhard. 1994. Naturally processed peptides longer than nine amino acid residues bind to the class I MHC molecule HLA-A2.1 with high affinity and in different conformations. *J. Immunol.* 152: 2874–2881.
14. Weiskopf, D., M. A. Angelo, E. L. de Azeredo, J. Sidney, J. A. Greenbaum, A. N. Fernando, A. Broadwater, R. V. Kolla, A. D. De Silva, A. M. de Silva, et al. 2013. Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proc. Natl. Acad. Sci. USA* 110: E2046–E2053.
15. Lidehall, A. K., F. Sund, T. Lundberg, B. M. Eriksson, T. H. Tötterman, and O. Korsgren. 2005. T cell control of primary and latent cytomegalovirus infections in healthy subjects. *J. Clin. Immunol.* 25: 473–481.
16. Motozono, C., N. Kuse, X. Sun, P. J. Rizkallah, A. Fuller, S. Oka, D. K. Cole, A. K. Sewell, and M. Takiguchi. 2014. Molecular basis of a dominant T cell response to an HIV reverse transcriptase 8-mer epitope presented by the protective allele HLA-B*51:01. *J. Immunol.* 192: 3428–3434.
17. Rist, M. J., A. Theodossis, N. P. Croft, M. A. Neller, A. Welland, Z. Chen, L. C. Sullivan, J. M. Burrows, J. J. Miles, R. M. Brennan, et al. 2013. HLA peptide length preferences control CD8+ T cell responses. *J. Immunol.* 191: 561–571.
18. Tey, S. K., F. Goodrum, and R. Khanna. 2010. CD8+ T-cell recognition of human cytomegalovirus latency-associated determinant pUL138. *J. Gen. Virol.* 91: 2040–2048.
19. Hassan, C., E. Chabrol, L. Jahn, M. G. Kester, A. H. de Ru, J. W. Drijfhout, J. Rossjohn, J. H. Falkenburg, M. H. Heemskerk, S. Gras, and P. A. van Veelen. 2015. Naturally processed non-canonical HLA-A*02:01 presented peptides. *J. Biol. Chem.* 290: 2593–2603.
20. Lundegaard, C., O. Lund, and M. Nielsen. 2008. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* 24: 1397–1398.
21. Blum, J. S., P. A. Wearsch, and P. Cresswell. 2013. Pathways of antigen processing. *Annu. Rev. Immunol.* 31: 443–473.
22. Sidney, J., S. Southwood, C. Moore, C. Oseroff, C. Pinilla, H. M. Grey, and A. Sette. 2013. Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr. Protoc. Immunol.* Chapter 18: Unit 18.3.
23. Cheng, Y., and W. H. Prusoff. 1973. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* 22: 3099–3108.
24. Gulukota, K., J. Sidney, A. Sette, and C. DeLisi. 1997. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* 267: 1258–1267.
25. Yaciuk, J. C., M. Skaley, W. Bardet, F. Schafer, D. Mojsilovic, S. Cate, C. J. Stewart, C. McMurtrey, K. W. Jackson, R. Buchli, et al. 2014. Direct interrogation of viral peptides presented by the class I HLA of HIV-infected T cells. *J. Virol.* 88: 12992–13004.
26. Carreno, B. M., V. Magrini, M. Becker-Hapak, S. Kaabinejadian, J. Hundal, A. A. Petti, A. Ly, W. R. Lie, W. H. Hildebrand, E. R. Mardis, and G. P. Linette. 2015. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* 348: 803–808.
27. Zhang, J., L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. a. Lajoie, and B. Ma. 2012. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* 11: M111.010587.
28. Nielsen, M., C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund, and S. Buus. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2: e796.
29. Hoof, I., B. Peters, J. Sidney, L. E. Pedersen, A. Sette, O. Lund, S. Buus, and M. Nielsen. 2009. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 61: 1–13.
30. Jørgensen, K. W., M. Rasmussen, S. Buus, and M. Nielsen. 2014. NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* 141: 18–26.
31. Vita, R., J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. 2015. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 43: D405–D412.
32. Granados, D. P., D. Sriranganadane, T. Daouda, A. Zieger, C. M. Laumont, O. Caron-Lizotte, G. Boucher, M.-P. Hardy, P. Gendron, C. Côté, et al. 2014. Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat. Commun.* 5: 3600.
33. Marcilla, M., A. Alpízar, M. Lombardía, A. Ramos-Fernandez, M. Ramos, and J. P. Albar. 2014. Increased diversity of the HLA-B40 ligandome by the presentation of peptides phosphorylated at their main anchor residue. *Mol. Cell. Proteomics* 13: 462–474.
34. Thommen, D. S., H. Schuster, M. Keller, S. Kapoor, A. O. Weinzierl, C. S. Chennakesava, X. Wang, L. Rohrer, A. von Eckardstein, S. Stevanovic, and B. C. Biedermann. 2012. Two preferentially expressed proteins protect vascular endothelial cells from an attack by peptide-specific CTL. *J. Immunol.* 188: 5283–5292.
35. Kotturi, M. F., B. Peters, F. Buendia-Laysa, Jr., J. Sidney, C. Oseroff, J. Botten, H. Grey, M. J. Buchmeier, and A. Sette. 2007. The CD8+ T-cell response to lymphocytic choriomeningitis virus involves the L antigen: uncovering new tricks for an old virus. *J. Virol.* 81: 4928–4940.
36. Wenzel, T., C. Eckerskorn, F. Lottspeich, and W. Baumeister. 1994. Existence of a molecular ruler in proteasomes suggested by analysis of degradation products. *FEBS Lett.* 349: 205–209.
37. Nussbaum, A. K., T. P. Dick, W. Keilholz, M. Schirle, S. Stevanović, K. Dietz, W. Heinemeyer, M. Groll, D. H. Wolf, R. Huber, et al. 1998. Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. *Proc. Natl. Acad. Sci. USA* 95: 12504–12509.
38. Kisselev, A. F., T. N. Akopian, K. M. Woo, and A. L. Goldberg. 1999. The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. *J. Biol. Chem.* 274: 3363–3371.
39. van Endert, P. M., R. Tampé, T. H. Meyer, R. Tisch, J. F. Bach, and H. O. McDevitt. 1994. A sequential model for peptide binding and transport by the transporters associated with antigen processing. *Immunity* 1: 491–500.

40. Schumacher, T. N., D. V. Kantesaria, M. T. Heemels, P. G. Ashton-Rickardt, J. C. Shepherd, K. Fruh, Y. Yang, P. A. Peterson, S. Tonegawa, and H. L. Ploegh. 1994. Peptide length and sequence specificity of the mouse TAP1/TAP2 translocator. *J. Exp. Med.* 179: 533–540.
41. Chang, S.-C., F. Momburg, N. Bhutani, and A. L. Goldberg. 2005. The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism. *Proc. Natl. Acad. Sci. USA* 102: 17107–17112.
42. Tynan, F. E., N. A. Borg, J. J. Miles, T. Beddoe, D. El-Hassen, S. L. Silins, W. J. van Zuylen, A. W. Purcell, L. Kjer-Nielsen, J. McCluskey, et al. 2005. High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I. Implications for T-cell receptor engagement and T-cell immunodominance. *J. Biol. Chem.* 280: 23900–23909.
43. Burrows, S. R., J. Rossjohn, and J. McCluskey. 2006. Have we cut ourselves too short in mapping CTL epitopes? *Trends Immunol.* 27: 11–16.
44. Samino, Y., D. López, S. Guil, L. Saveanu, P. M. van Endert, and M. Del Val. 2006. A long N-terminal-extended nested set of abundant and antigenic major histocompatibility complex class I natural ligands from HIV envelope protein. *J. Biol. Chem.* 281: 6358–6365.
45. Hörig, H., A. C. Young, N. J. Papadopoulos, T. P. DiLorenzo, and S. G. Nathenson. 1999. Binding of longer peptides to the H-2Kb heterodimer is restricted to peptides extended at their C terminus: refinement of the inherent MHC class I peptide binding criteria. *J. Immunol.* 163: 4434–4441.
46. Collins, E. J., D. N. Garboczi, and D. C. Wiley. 1994. Three-dimensional structure of a peptide extending from one end of a class I MHC binding site. *Nature* 371: 626–629.