Title:

Deciphering HLA motifs across HLA peptidomes correctly predicts neo-antigens and identifies allostery in HLA specificity

Running title: Deciphering HLA motifs across HLA peptidomes

Authors

Michal Bassani-Sternberg[1,2,*], Chloé Chong[1,2], Philippe Guillaume[1,2], Marthe Solleder[1,3], HuiSong Pak[1,2], Philippe O Gannon[2], Lana E Kandalaft[1,2], George Coukos[1,2], David Gfeller[1,2,3,*]

Affiliations

[1]Ludwig Centre for Cancer Research, University of Lausanne, 1066 Epalinges, Switzerland. [2]Department of Fundamental Oncology, University Hospital of Lausanne, Lausanne, Switzerland. [3]Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland.

[*]To whom correspondence should be sent: David.Gfeller@unil.ch and Michal.Bassani@chuv.ch.

Abstract

The precise identification of Human Leukocyte Antigen class I (HLA-I) binding motifs plays a central role in our ability to understand and predict (neo-)antigen presentation in infectious diseases and cancer. Here, by exploiting co-occurrence of HLA-I alleles across ten newly generated as well as forty publicly available in-depth HLA peptidomics datasets, we show that we can rapidly and accurately identify HLA-I binding motifs and map them to their corresponding alleles without any *a priori* knowledge of HLA-I binding specificity. Our novel approach uncovers new motifs for several alleles that up to now had no known ligands. HLA-ligand predictors trained on such data substantially improve neo-antigen predictions in four melanoma and two lung cancer patients, indicating that unbiased HLA peptidomics data are ideal for *in silico* identification of (neo-)antigens. The new motifs further reveal allosteric modulation of HLA-I binding specificity and we unravel the underlying mechanisms by protein structure analysis, mutagenesis and *in vitro* binding assays.

Introduction

HLA-I molecules play a central role in defence mechanisms against pathogens and immune recognition of cancer cells. Their main functionality is to bind short peptides (9-11 mers) coming from degradation products of endogenous or viral proteins. The peptides are cleaved in the proteasome, transported by the transporter associated with antigen processing (TAP) complex, loaded on the HLA-I molecules in the ER and presented at the cell surface[1]. Non-self peptides presented on HLA-I molecules, such as those derived from degradation of viral proteins or mutated and other cancer specific proteins (referred to as neo-antigens), can then be recognized by CD8 T cells and elicit cytolytic activity. Neo-antigens have recently emerged as promising targets for development of personalized cancer immunotherapy[2].

Human cells express three HLA-I genes (HLA-A/B/C). These genes are the most polymorphic of the human genome and currently more than 8,000 different alleles have been observed[3]. Such a high polymorphism makes it challenging to model the different binding specificities of each allele and predict antigens presented at the cell surface. Information about binding motifs of HLA-I molecules has been mainly obtained from biochemical assays where chemically synthesized peptides are tested *in vitro* for binding. Currently, the most frequent HLA-I alleles have many known ligands and these data have been used to train machine learning algorithms for HLA-peptide interaction predictions[4-9]. However, the vast majority (>90%) of HLA-I alleles still lack documented ligands and despite algorithmic developments to generalize prediction methods to any allele[10], it is more challenging to accurately describe motifs of alleles without known ligands.

Mass-spectrometry (MS) analysis of HLA binding peptides eluted from cell lines or tissue samples is a promising alternative to the use of HLA-ligand interaction predictions. Despite technical and logistic challenges (typically 1cm^3 of tissue material is required), MS is increasingly used to directly identify viral[11,12] or cancer-specific (neo-)antigens[13-18], when experimentally feasible. Tens of thousands of endogenous peptides naturally presented on HLA-I molecules are identified in such HLA peptidomics studies, providing a unique opportunity to collect very large numbers of HLA ligands that could be used to better understand the binding properties of HLA-I molecules. The challenge in studying HLA-I motifs based on

such pooled peptidomics data from unmodified cell lines or tissue samples is to determine the allele on which each peptide was displayed. While the most widely used approach is to predict binding affinity of each peptide to each allele present in a sample[19], recent studies have suggested that HLA-I motifs could be identified in HLA peptidomics datasets in an unsupervised way by grouping peptides based on their sequence similarity[15,20-22]. However, this strategy still relies on previous information about HLA-I binding specificity when associating predicted motifs with HLA-I alleles and is therefore restricted to alleles whose motifs have been already characterized.

Here, we show for the first time that dozens HLA-I motifs can be identified without any *a priori* information about HLA-I binding specificity by taking advantage of co-occurrence of HLA-I alleles across both newly generated and publicly available HLA peptidomics datasets. Our approach uncovers new motifs for several alleles for which, until this study, no known ligands had been documented and significantly improves neo-antigen predictions in five out of six tumor samples with experimentally determined neo-antigens. Our large and unbiased collection of HLA-I ligands further allows us to unravel some of the molecular determinants of HLA-I binding motifs and reveals allosteric modulation of HLA-I binding specificity. To elucidate the underlying molecular mechanisms, we show how a single point mutation in HLA-B14:02 outside of the P2 binding site significantly changes the amino acid preferences at P2 in the ligands.

Results

Fully unsupervised identification of HLA-I binding motifs

To study the properties of HLA-I alleles at a large scale and without relying on *a priori* knowledge about their binding specificity, we reasoned that HLA-I binding motifs might be identified across samples with in-depth and accurate HLA peptidomics data by taking advantage of co-occurrence of HLA-I alleles. To this end, we measured the HLA peptidome eluted from six B cell lines, two *in vitro* expanded tumor-infiltrating lymphocytes (TILs) samples and two leukapheresis samples (peripheral blood mononuclear cells) selected based on their high diversity of HLA-I alleles (see Methods and Supplementary Dataset 1). We identified 47,023 unique

peptides displayed on 32 HLA molecules. To expand the coverage of HLA-I alleles, we further collected 40 publicly available high-quality HLA peptidomics datasets[15,16,20,21,23-25] (see Methods and Supplementary Dataset 2). Our final data consists of a total of 50 HLA peptidomics datasets covering 66 different HLA-I alleles (18 HLA-A, 32 HLA-B and 16 HLA-C alleles, see Supplementary Table 1). The number of unique HLA-ligand interactions across all samples reaches 252,165 for a total of 119,035 unique peptides (9- to 14-mers), which makes it, to our knowledge, the largest currently available collection of HLA peptidomics datasets both in terms of number of peptides and diversity of HLA-I molecules. Binding motifs in each HLA peptidomics dataset were identified for 9- and 10-mers using a motif discovery algorithm initially developed for multiple specificity analyses[26,27] and recently applied to the analysis of seven HLA peptidomics datasets[22] (see Supplementary Fig. 1). Importantly, this method does not rely on HLA-peptide interaction predictions (see Methods).

To assign each motif to its allele even in the absence of *a priori* information about the alleles' binding specificity, we developed the strategy illustrated in Fig. 1a. In this example coming from our newly generated HLA peptidomics data, one allele was shared between two samples (HLA-B38:01). Remarkably, exactly one identical motif was shared between the two samples. As such, one can predict that this motif corresponds to the shared allele. Similarly, if one sample shares all but one allele with another sample, it can be inferred that the motif that is not shared corresponds to the unshared allele (Fig. 1b). These two ideas can then be recursively applied to identify HLA-I motifs across large HLA peptidomics datasets from all our different samples (see Methods). Of note, motifs identified in distinct samples that have some alleles in common show very high similarity (Fig. 1 and Supplementary Fig. 1) and our new approach builds upon this remarkable reproducibility of in-depth and accurate HLA peptidomics data.

We applied our algorithm to our 50 HLA peptidomics datasets. In total, 44 different motifs could be associated with their corresponding allele without relying on any *a priori* knowledge of HLA-I binding specificity (Fig. 2a). To validate our predictions, we compared the motifs predicted by our fully unsupervised method with known motifs derived from IEDB[28]. Despite some differences (e.g. HLA-A25:01 motif at P9), we observed an overall high similarity for most common alleles, confirming the very high accuracy of our predictions (Fig. 2a). Moreover, seven alleles (HLA-

B13:02, HLA-B14:01, HLA-B15:11, HLA-B15:18, HLA-B18:03, HLA-B39:24 and HLA-C07:04) for which our approach uncovered a clear motif did not have known ligands in IEDB, and an additional one (HLA-B56:01) had only three known ligands.

Semi-supervised approach

For the most frequent HLA-I alleles, a reasonable description of their binding motifs can be obtained from existing databases[28]. To further expand our collection of HLA-I binding motifs, we used similarity to the binding motifs derived from IEDB ligands to annotate motifs that could not be assigned to their corresponding allele by the fully unsupervised approach (see Methods). We then re-ran our algorithm on the remaining motifs. This enabled us to determine the binding motifs of eight additional alleles (Supplementary Fig. 2), including one without known ligands until this study (HLA-A02:20, Fig. 2b). The same approach was applied on 10-mers identified by MS and revealed six new motifs for poorly characterized alleles in IEDB (Supplementary Fig. 3).

Investigating technical biases in HLA peptidomics studies

Different technical biases may affect MS data, which could undermine their use for training HLA-peptide interaction predictors. To investigate this potential issue, we computed amino acid frequencies at non-anchor positions (P4 to P7) in our HLA peptidomics data (see Methods and Supplementary Table 2). We observed a high correlation between amino acid frequencies at non-anchor positions in our HLA peptidomics data and in the human proteome (r=0.85) (Fig. 3 and Supplementary Fig. 4). The most important difference was found for cysteine, which is rich in post-translational modifications that are typically not included in database searches and was observed at very low frequency in the HLA peptidomics data. Other amino acids were not much under- or over-represented, and no clear pattern emerged from these data with respect to amino acid biophysical properties (e.g., charge, hydrophobicity, size). Overall, our results suggest that HLA peptidomics data do not show strong technical biases, apart from under-representation of cysteine, and therefore provide ideal data for training HLA-peptide interaction predictors, especially for ligands coming from human cells like neo-antigens.

Predictions of neo-antigens in tumors

To test whether our unique dataset of naturally presented peptides could help predicting neo-antigens in tumors, we trained a predictor of HLA-ligands with these data. As MS only includes positive examples, we built Position Weight Matrices (PWMs) for each allele based on the peptides assigned to it across all our HLA peptidomics datasets. To account for the main bias in HLA peptidomics data, we further renormalized our predictions by amino acid frequencies observed at non-anchor positions (see Methods).

We then collected all available datasets that included direct identification with mass spectrometry of neo-antigens displayed on cancer cells as well as exome sequencing (Mel5, Mel8, Mel15 from[15] and 12T from[18], for a total of ten 9- and 10-mers neo-antigens, see Table 1). This dataset has the unique advantage of not being restricted to neo-antigens selected based on *in silico* predictions, and is therefore ideal for benchmarking neo-antigen prediction methods. Moreover, as these studies are quite recent, the neo-antigens are not part of the training set of any existing algorithm. We retrieved all possible 9- and 10-mer peptides that encompassed each missense mutation (Supplementary Dataset 3). We then ranked all these potential neo-antigens based on the score of our predictor (see Methods). Of note, HLA peptidomics data used to train our predictor only consist of wild-type human peptides and did not include any mutated peptide identified in our previous report[15]. Remarkably, eight out of the ten neo-antigens were best predicted by our predictor and seven of them fell among the top 20, indicating that by testing as few as 20 peptides per sample, we could identify more than two thirds of the neo-antigens found by MS (Table 1). Considering that the total number of potential neo-antigens (i.e. 9- and 10-mers containing a missense mutation) can be as large as 25,000 for tumors with high mutational load, our predictor trained on naturally presented human HLA ligands clearly enabled us to significantly reduce the number of peptides that would need to be experimentally tested to identify *bona fide* neo-antigens from exome sequencing data. Importantly, even if we did not include in the training of our predictor MS data from the samples in which the neo-antigens were identified, six neo-antigens were still well predicted (see Supplementary Table 3). When comparing with standard tools that are widely used to narrow-down the list of predicted neo-antigens[6,10,29], our method trained on HLA peptidomics data showed clear improvement (Table 1), even when restricting the number of potential neo-antigens to those predicted by NetMHC[6] (Fig. 4a). Nevertheless, both our predictor and standard prediction tools failed to

identify some neo-antigens (e.g., KLILWRGLK from NCAPG2 P333L mutation, see Table 1)[15]. This suggests that, when enough tumor material is available for immunopeptidomics analyses, direct identification of neo-antigens with MS should still be performed to optimally enrich in true positives the list of ligands to be experimentally tested for immunogenicity[17]. As a second independent validation for our predictions we used the neo-antigens recently identified in two lung cancer patients[30] (see Methods). Although this study was restricted to peptides containing missense mutations pre-selected based on binding affinities predicted with existing tools[10], our predictor enabled us to further improve the prioritization of neo-antigens that elicit T-cell reactivity with an average Area Under the ROC Curve (AUC) of 0.70 compared to 0.60 for NetMHC[6], 0.64 for NetMHCpan[10] and 0.66 for NetMHCstabpan[29] (see Fig. 4b, Supplementary Table 4 and Methods).

Analysis of the newly identified motifs

One of our novel HLA-I motifs describes the binding specificity of HLA-A02:20 (Fig. 2b). HLA-A02 binding motifs have been widely studied. However, HLA-A02:20 motif differs from standard HLA-A02 motifs at P1, with a clear preference for charged residues (Fig. 5a). Interestingly, HLA-A02:20 is among the very few (<2%) HLA-A02 alleles that do not have a conserved lysine pointing towards P1 at position 90. Instead an asparagine is found there (Fig. 5a). To explore whether the absence of lysine at position 90 may explain the observed difference in binding specificity, we collected all HLA-I alleles showing preference for charged amino acids at P1 (see Supplementary Fig. 5). All of them had either asparagine or isoleucine at position 90. We then explored available crystal structures of HLA-peptide complexes with charged residues at P1. HLA-B57:03 was crystalized with such a ligand (KAFSPEVI)[31]. Superposing the crystal structure of this complex with the structure of HLA-A02:01 provides a possible mechanism for understanding the change in binding specificity at P1. In HLA-A02:01, lysine at position 90 interacts with the hydroxyl group of serine at P1 (Fig. 5a, green sidechains). Such a conformation would not be compatible with a longer residue. Reversely, when asparagine is found at position 90, it does not point towards P1 (Fig. 5a, pink sidechains), thereby freeing space for larger sidechains like lysine or arginine at P1. Overall, our analysis indicates that the presence of asparagine at residue 90 may be responsible for the change in binding specificity between HLA-A02:01 and HLA-

A02:20. More generally, our results suggest that lysine at residue 90 in HLA-I alleles strongly disfavours charged residues at P1.

The new motif identified for HLA-B15:18 (Fig. 2a) displayed strong preference for histidine at P2, which is not often observed in HLA-I ligands. To gain insights into the mechanisms underlying this uncommon binding motif, we surveyed all alleles that show preference for histidine at P2 (Supplementary Fig. 6a). Sequence and structure analysis showed that all of them have a conserved P2 binding site, including cysteine at position 91 (Fig. 5a and 5c), suggesting that this amino acid may play a role in accommodating histidine in the ligands. However, a few other alleles also have the conserved cysteine at position 91 but show specificity for arginine at P2 (Supplementary Fig. 6b). Among them, HLA-B14:02 had the highest sequence similarity to HLA-B15:18, with only 8 different residues in the peptide binding domain. Structural inspection of the non-conserved residues showed that none make any contact with arginine at P2 in the crystal structure of HLA-B14:02 (orange residues in Fig. 5c). This suggests that the difference in binding specificity at P2 between HLA-B14:02 and HLA-B15:18 is likely explained by allosteric mechanisms. Of particular interest is residue 121 (W in HLA-B14:02 and R in HLA-B15:18), which is more than 7Å away from the arginine sidechain at P2 and is part of a network of aligned aromatic residues (Y33, W121 and F140) in HLA-B14:02 (Fig. 5c). We hypothesized that mutating this residue into arginine in HLA-B14:02 may modify the binding specificity to accommodate histidine at P2.

Molecular mechanism underlying allosteric modulation of HLA-I binding specificity

To test our hypothesis, we generated a construct for HLA-B14:02 wild-type (wt) and W121R mutant. We tested several ligands of HLA-B15:18 identified in our HLA peptidomics data with histidine at P2, which were predicted to show enhanced binding to HLA-B14:02 W121R. As expected, a strong decrease in binding stability was observed between HLA-B14:02 W121R and HLA-B14:02 wt (Fig. 5d). Reversely, when testing the same peptides with arginine at P2, a significant increase in stability was observed between HLA-B14:02 W121R and HLA-B14:02 wt (Fig. 5d). For instance, binding of the peptide AHTKPRPAL was fully abolished in HLA-B14:02 wt, but was rescued when changing histidine to arginine at P2. Although other

residues may also play a role in the binding specificity differences between of HLA-B14:02 and HLA-B15:18, all of them are further away from P2, which supports the allosteric hypothesis. Overall, our results show that HLA-I binding specificity at P2 can be modulated by amino acids outside of the P2 binding site.

Discussion

Despite decades of work to characterize the binding motifs of the most common HLA-I alleles, in-depth unbiased peptide screening approaches are not commonly used. This is mainly because both the N- and the C-terminus of the peptides are engaged in binding to HLA-I molecules, thereby preventing the use of high-throughput techniques for peptide screening like phage display or peptide arrays. Here, for the first time, we show that in-depth and accurate HLA peptidomics data from unmodified cell lines and tissue samples together with novel machine learning algorithms enable us to rapidly identify thousands of new HLA-I ligands in a fully unsupervised way and characterize the binding properties of HLA-I alleles that had no known ligands until this study. Remarkably, our predicted motifs displayed high similarity with known motifs for common alleles and very little technical biases. This suggests that HLA peptidomics data are optimal to train HLA-ligand interaction predictors, as confirmed by our ability to accurately predict many neo-antigens identified by mass spectrometry in tumor samples. Although mass spectrometry may miss a fraction of the actually presented and immunogenic neo-antigens, those detected by MS are likely presented at high level on cancer cells. Therefore, accurately predicting such dominant neo-antigens is promising to prioritize targets for cancer immunotherapy. Importantly, the improvement in neo-antigen prediction accuracy appears to come primarily from better motifs. For instance, differences are observed at P9 between the motif of HLA-A25:01 derived from our study (preference for F/W/Y/L) or from IEDB ligands (preference for Y/L/M/F) (Fig. 2a). This likely explains why the neo-antigen `ETSKQVTRW` was poorly predicted by standard tools[6,10] (predicted IC_{50} > 3,000nM). Moreover, as our predictor is only trained on naturally presented ligands, it may also capture some features of antigen presentation of

endogenous peptides beyond the sole binding to HLA molecules. Along this line, it is interesting to note that a less important improvement in predictions had been observed when attempting to predict epitopes from the SYFPEITHI database[32] (including a large fraction of viral peptides) with HLA peptidomics data[22]. Although the dataset used in this previous study was significantly smaller, this observation suggests that HLA peptidomics data may be especially well suited for training predictors of human endogenous or mutated HLA ligands. Overall, our work highlights the importance of carefully determining HLA-I motifs based on unsupervised analysis of naturally presented human HLA ligands for neo-antigen discovery.

Results shown in Fig. 5 further emphasize the power of in-depth sampling of the HLA-I ligand space to inform us about molecular mechanisms underlying HLA-I binding properties[33]. Considering the rapid expansion of HLA peptidomics experiments performed in cancer immunotherapy research[14-16,18,24,25], we anticipate that our novel approach for HLA-I motif identification will enable similar analyses in the future to uncover other molecular determinants of HLA-I binding specificity.

Our algorithm does not require motifs corresponding to each allele to be identified. For instance in Fig. 1a, only five motifs were found in the first sample (GD149), while this sample had six different alleles. This is a frequent situation when analysing HLA peptidomics data with unsupervised approaches[22,34]. As previously observed, it affects especially HLA-C alleles which are often poorly expressed and are more redundant[22,35] (Supplementary Fig. 1).

Overall, our work shows for the first time that HLA-I motifs can be reliably identified across in-depth and accurate HLA peptidomics datasets without relying on peptide-HLA interaction prediction tools or *a priori* knowledge of HLA binding specificity. This fully unsupervised and unbiased approach not only recapitulates and improves known HLA-I binding motifs but also expands our understanding of HLA-I binding specificities to alleles without documented ligands. As such, our work is a powerful alternative to chemically synthesizing every peptide for *in vitro* binding assays, or to genetically modifying or transfecting cell lines with soluble HLA alleles[36-38]. Our results further contribute to our global understanding of HLA-I binding properties and significantly improve neo-antigen predictions. This work may therefore facilitate identification of clinically relevant targets for cancer immunotherapy, especially when direct identification of neo-antigens with MS cannot be experimentally done.

Methods

Cell lines and antibodies

We carefully selected ten donors expressing a broad range of HLA-I alleles and generated novel HLA peptidomics data (Supplementary Table 1). EBV-transformed human B-cell lines CD165, GD149, PD42, CM467, RA957 and MD155 were maintained in RPMI 1640 + GlutaMAX medium (Gibco, Paisley, UK) supplemented with 10% FBS (Gibco) and 1% Penicillin/Streptomycin Solution (BioConcept, Allschwil, Switzerland). TIL were expanded from two melanoma tumors following established protocols[39,40]. Informed consent of the participants was obtained following requirements of the institutional review board (Ethics Commission, University Hospital of Lausanne (CHUV)). Briefly, fresh tumor samples were cut in small fragments and placed in 24-well plate containing RPMI CTS grade (Life Technologies Europe BV, Switzerland), 10% Human serum (Valley Biomedical, USA), 0.025 M HEPES (Life Technologies Europe BV, Switzerland), 55 µmol/L 2-Mercaptoethanol (Life Technologies Europe BV, Switzerland) and supplemented with a high concentration of IL-2 (Proleukin, 6,000 IU/mL, Novartis, Switzerland) for three to five weeks. Following this initial pre-REP, TIL were then expanded in using a REP approach. To do so, 25 x10^6 TIL were stimulated with irradiated feeder cells, anti-CD3 (OKT3, 30 ng/mL, Miltenyi biotech) and high dose IL-2 (3,000 IU/mL) for 14 days. The final cell product was washed and prepared using a cell harvester (LoVo, Fresenius Kabi). Leukapheresis samples (Apher1 and 6) were obtained from blood donors from the *Service régional vaudois de transfusion sanguine, Lausanne*. Upon receival of TIL and leukapheresis samples, the cells were washed with PBS on ice, aliquoted and stored as dry pellets at -80°C until use. High resolution 4-digit HLA-I typing was performed at the Laboratory of Diagnostics, Service of Immunology and Allergy, CHUV, Lausanne.

W6/32 monoclonal antibodies were purified from the supernatant of HB95 cells grown in CELLLine CL-1000 flasks (Sigma-Aldrich, Missouri, USA) using Protein-A Sepharose (Invitrogen, California, USA).

Purification of HLA-I complexes

We extracted the HLA-I peptidome from 2-5 biological replicates per cell line or patient material. The cell counts ranged from 1 x 10^8 to 3 x 10^8 cells per replicate. Lysis was performed with 0.25% sodium deoxycholate (Sigma-Aldrich), 0.2 mM iodoacetamide (Sigma-Aldrich), 1 mM EDTA, 1:200 Protease Inhibitors Cocktail (Sigma, Missouri, USA), 1 mM Phenylmethylsulfonylfluoride (Roche, Mannheim, Germany), 1% octyl-beta-D glucopyranoside (Sigma) in PBS at 4°C for 1 hr. The lysates were cleared by centrifugation with a table-top centrifuge (Eppendorf Centrifuge 5430R, Schönenbuch, Switzerland) at 4°C at 14200 rpm for 20 min. Immuno-affinity purification was performed by passing the cleared lysates through Protein-A Sepharose covalently bound to W6-32 antibodies. Affinity columns were then washed with at least 6 column volumes of 150 mM NaCl and 20 mM Tris HCl (buffer A), 6 column volumes of 400 mM NaCl and 20 mM Tris HCl and lastly with another 6 column washes of buffer A. Finally, affinity columns were washed with at least 2 column volumes of 20 mM Tris HCl, pH 8. HLA-I complexes were eluted by addition of 1% trifluoroacetic acid (TFA, Merck, Darmstadt, Switzerland) for each sample.

Purification and concentration of HLA-I peptides

HLA-I complexes with HLA-I peptides were loaded on Sep-Pak tC18 (Waters, Massachusetts, USA) cartridges which were pre-washed with 80% acetonitrile (ACN, Merck) in 0.1% TFA and 0.1 % TFA only. After loading, cartridges were washed twice with 0.1% TFA before separation and elution of HLA-I peptides from the more hydrophobic HLA-I heavy chains with 30 % ACN in 0.1 % TFA. The HLA-I peptides were dried using vacuum centrifugation (Eppendorf Concentrator Plus, Schönenbuch, Switzerland) and re-suspended in a final volume of 12 uL 0.1% TFA. For MS analysis, we injected 5 uL of these peptides per run.

LC-MS/MS analysis of HLA-I peptides

Measurements of HLA-I peptidomics samples were acquired using the nanoflow UHPLC Easy nLC 1200 (Thermo Fisher Scientific, Germering, Germany) coupled online to a Q Exactive HF Orbitrap mass spectrometer (Thermo Fischer Scientific, Bremen, Germany) or with Dionex Ultimate RSLC3000 nanoLC (Thermo Fischer Scientific, *Sunnyvale, CA)* coupled online to an Orbitrap Fusion Mass Spectrometer (Thermo Fischer Scientific, San Jose, CA), both with a nanoelectrospray ion source.

We packed an uncoated PicoTip 8µm tip opening with diameter of 50 cm x 75 um with a ReproSil-Pur C18 1.9 µm particles and 120 Å pore size resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) re-suspended in Methanol. The analytical column was heated to 50°C using a column oven. Peptides were eluted with a linear gradient of 2–30% buffer B (80% ACN and 0.1% formic acid) at a flow rate of 250 nl/min over 90 min.

Data was acquired with data-dependent "top10" method, which isolates the ten most intense ions and fragments them by higher-energy collisional dissociation (HCD) with a normalized collision energy of 27% and 32% for the Q Exactive HF and Fusion instruments, respectively. For the Q Exactive HF instrument the MS scan range was set to 300 to 1,650 *m/z* with a resolution of 60,000 (200 *m/z*) and a target value of 3e6 ions. The ten most intense ions were sequentially isolated and accumulated to an AGC target value of 1e5 with a maximum injection time of 120 ms and MS/MS resolution was 15,000 (200 *m/z*). For the Fusion, a resolution of 120,000 (200 *m/z*) and a target value of 4e5 ions were set. The ten most intense ions accumulated to an AGC target value of 1e5 with a maximum injection time of 120 ms and MS/MS resolution was 15,000 (200 *m/z*). The peptide match option was disabled. Dynamic exclusion of fragmented *m/z* values from further selection was set for 20 or 30 seconds with the Q Exactive HF and Fusion instruments, respectively.

Data analysis of HLA-I peptides

We employed the MaxQuant computational proteomics platform[41] version 1.5.3.2 to search the peak lists against the UniProt databases (Human 85,919 entries, May 2014) and a file containing 247 frequently observed contaminants. N-terminal acetylation (42.010565 Da) and methionine oxidation (15.994915 Da) were set as variable modifications. The second peptide identification option in Andromeda was enabled. A false discovery rate of 0.01 was required for peptides and no protein false discovery rate was set. The enzyme specificity was set as unspecific. Possible sequence matches were restricted to 8 to 15 amino acids, a maximum peptides mass of 1,500 Da and a maximum charge state of three. The initial allowed mass deviation of the precursor ion was set to 6 ppm and the maximum fragment mass deviation was set to 20 ppm. We enabled the 'match between runs' option, which allows matching of identifications across different replicates of the same biological sample in a time window of 0.5 min and an initial alignment time window of 20 min.

Publicly available HLA-I peptidomics data

To expand the number of samples and survey an even broader range of HLA-I alleles, we included in this study forty publicly available HLA peptidomics data from seven recent studies[15,16,20,21,23-25]. Only samples with HLA-I typing were used. Peptides identified in the recent study[16] in different repeats and under different treatments were pooled together to generate one list of unique peptides per sample. Since the published peptidomics datasets from Pearson et al.[25] were filtered to include only peptides with predicted affinity scores of less or equal to 1250 nM, we re-processed the mass spectrometer raw data using MaxQuant with similar settings as mentioned above except that peptide length was set to 8-25 mers (Supplementary Dataset 2).

IEDB data

Known HLA-I ligands were retrieved from IEDB (mhc_ligand_full.csv file, as of October 5, 2016)[28]. All ligands annotated as positives with a given HLA-I allele (i.e., "Positive-High", "Positive-Intermediate", "Positive-Low" and "Positive") were used to build the IEDB reference motifs (Fig. 2). Ligands coming from HLA peptidomics studies analysed in this work were not considered to prevent circularity in the motif comparisons and because the HLA-I alleles to which these peptides bind were not experimentally determined. Position Weight matrices (PWMs) representing binding motifs in IEDB and used to compare with motifs derived from our deconvolution of HLA peptidomics datasets were built by computing the frequency of each amino acid at each position and using a random count of 1 for each amino acid at each position.

Mixture model for HLA-I binding motifs identification in HLA peptidomics data

An algorithm based on mixture models and initially developed for multiple specificity analysis in peptide ligands[26,27] was used to identify binding motifs in each dataset analysed in this work. Briefly, all peptides pooled by mass spectrometry analysis of eluted peptide-HLA complexes in a given sample were first split into different groups according to their size (9-10 mers). All 9- and 10-mers ligands were then modelled using multiple PWMs[22]. The results of such analysis consist of a set of PWMs that describe distinct motifs for each HLA peptidomics datasets (see Supplementary Fig. 1) and probabilities (i.e., responsibilities) for each peptide to be associated with each motif.

Fully unsupervised deconvolution of HLA peptidomics datasets

The availability of high-quality HLA peptidomics data from several samples with diverse HLA-I alleles suggest that one could infer which motifs correspond to which HLA-I alleles without relying on comparison with known motifs. For instance, if two samples share exactly one HLA-I allele, it is expected that the shared motif will originate from the shared allele (Fig. 1a). To exploit this type of patterns of shared HLA-I alleles, we designed the following algorithm:

1) For each allele present in at least two samples, find all samples that share only this allele (e.g., HLA-B38:01 in Fig. 1a). Identify the shared motif. If a shared motif is found, map this motif to the corresponding allele in each sample.
2) Find samples that share all except one allele with another sample (e.g., HLA-A24:02 in Fig. 1b). Find the motif that is not shared and map it to the HLA-I allele that is not shared.
3) Check if some samples contain exactly one motif and one allele that have not yet been associated. Annotate the remaining motif to the corresponding allele.
4) Use the motifs mapped to HLA-I alleles in steps 1), 2) and 3) to identify them in other samples that contain these alleles based on motif similarity.
5) Go back to 1) until no new motif can be mapped to HLA-I alleles.

Comparison of motifs was performed using Euclidean distance between the corresponding PWMs: $D = \frac{1}{L}\sum_{i=1}^{20}\sum_{A=1}^{L}(M_{iA} - M'_{iA})^2$, where M and M' stands for two PWMs (i.e., *20* x L matrices) to be compared and L is the peptide length. A threshold of $T = 0.078$ was used to define similar motifs based on visual inspection. Cases of inconsistencies (i.e., distances larger that T) between motifs mapped to the same allele were automatically eliminated. The final binding motifs for each HLA-I allele (Fig. 2a) were built by combining peptides from each sample that had been associated with the corresponding allele.

In practice, HLA-A and HLA-B alleles tend to be more expressed and therefore give rise to a stronger signal in HLA peptidomics data. We therefore used first our deconvolution method[22], setting the number of motifs equal to the number of HLA-A and HLA-B alleles and identified HLA-A and HLA-B motifs with the algorithm

introduced above (Step 1). We then ran the deconvolution method of [22] without restricting the number of clusters (Step 2) and identified motifs corresponding to HLA-A and HLA-B alleles based on the similarity with those identified in Step 1. The remaining motifs were then analysed across all samples with the algorithm introduced above.

Semi-supervised deconvolution of HLA peptidomics datasets

To expand the identification of binding motifs for alleles without known ligands, we used data from IEDB for HLA-I alleles with well-described binding motifs. In practice, for all HLA-I alleles in our samples that had not been mapped to motifs in the fully unsupervised approach and have more than twenty different ligands in IEDB, PWMs were built from IEDB data. These PWMs were used to scan the remaining motifs in each sample that contained the corresponding alleles. Motifs were mapped to HLA-I alleles if exactly one PWM obtained with the mixture model was found to be similar to the IEDB-derived motif (i.e., Euclidean distance smaller than T, as before). The unsupervised procedure described above was then applied to the remaining motifs to identify new motifs for alleles without ligands in IEDB.

Amino acid frequencies at non-anchor positions in HLA peptidomics datasets

To have reliable estimates of the potential technical biases due to MS, amino acid frequencies were computed at non-anchor positions (P4 to P7) for alleles in our HLA peptidomics datasets (9-mers). Alleles showing some specificity at these positions (A02:01, A02:05, A02:06, A02:20, A25:01, A26:01, A29:02, B08:01, B14:01, B14:02, C03:03, and C07:04, see Supplementary Fig. 2) were excluded from this analysis. The average frequencies of amino acids across alleles were then compared against the human proteome using Pearson correlation coefficient (Fig. 3). We also performed the same analysis with HLA-I ligands (9-mers) from IEDB splitting between those obtained by MS and those obtained by other assays ("non-MS data") (see Supplementary Fig. 4). To enable meaningful comparison between these datasets, only alleles present in our HLA peptidomics data, with more than 100 ligands in both IEDB MS and non-MS data were considered in this analysis (14 alleles in total, see Supplementary Fig. 2).

Prediction of neo-antigens

For each HLA-I allele, PWMs were built from all peptides associated to this allele across all samples where the binding motif could be identified, using the highest responsibility values of the mixture model[22]. The frequency of each amino acid was first computed. Pseudocounts were added using the approach described in[42], based on the BLOSUM62 substitution matrix with parameter β=200. The score of a given peptide ($X_1,\ldots X_N$) was computed by summing the logarithm of the corresponding PWM entries, including renormalization by expected amino acid frequencies: $S = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{p_{X_i,i}}{q_{X_i}}\right)$. Here q_A stands for frequency of amino acid A at non-anchor positions (Fig. 3 and Supplementary Table 2), $p_{A,i}$ stands for the PWM entry corresponding to amino acid A at position i, and N stands for the length of the peptide (N=9,10). The final score of a peptide was taken as the maximal score across all alleles present in a given sample and a p-value estimate was computed by comparing with distribution of scores obtained from 100,000 randomly selected peptides from the human proteome.

To test our ability to predict neo-antigens, we used four melanoma samples in which ten neo-antigens (9- and 10-mers) have been directly identified with in-depth immunopeptidomics analyses of the tumor samples: Mel5, Mel8, Mel15 from [15] and 12T [18]. Missense mutations (i.e. cancer specific non-synonymous point mutations) identified by exome sequencing in those four melanoma samples[15] were retrieved and a list of all possible 9- and 10-mer peptides encompassing each mutation was built (Supplementary Dataset 3). Multiple transcripts corresponding to the same genes were merged so that each mutated peptide appears only once in the list. The total number of potential neo-antigens in each sample is shown in Table 1. Predictions for each HLA-I allele of each sample were carried out with the model described above. Peptides were ranked based on the highest score over the different alleles present in their sample. In parallel, affinity predictions with NetMHC (v4.0)[6] and NetMHCpan (v2.8) [10] and stability predictions with NetMHCstabpan (v1.0)[29] were performed for the same peptides and peptides were ranked based on predicted affinity using the highest value (i.e., lowest Kd) over all alleles. Only HLA-A and HLA-B alleles were considered since HLA-C alleles are known to show much lower expression and NetMHC could not be run for some HLA-C alleles in these melanoma patients. Ranking of the neo-antigens compared to all possible peptides containing a missense mutation is shown in Table 1. To compare with standard approaches, Area Under the

Curves were also calculated by restricting to potential neo-antigens that passed the widely used 500nM threshold on NetMHC predictions[6] (Fig. 4a). The experimentally identified neo-antigens ETS**K**QVTRW (Mel5 predicted best IC_{50} = 3231nM with HLA-A25:01) and DAN**S**FLQSV (12T, predicted best IC_{50} = 6235nM with HLA-B51:01) did not pass this affinity threshold and were added to the lists.

The same analysis was applied to study neo-antigens (9- and 10-mers) recently identified in two lung cancer patients[30] (L011: FAFQEYDSF, GTSAPRKKK, SVTNEFCLK, RSMRTVYGLF, GPEELGLPM and L013: YSNYYCGLRY, ALQSRLQAL, KVCCCQILL). In this study, only mutated peptides with a predicted binding affinity < 500nM were tested for T-cell reactivity. From the list of missense mutations, we therefore extracted all such mutated peptides potentially binding to one HLA-I allele and ranked with our predictor as well as other tools those experimentally found to be immunogenic (Fig. 4b).

The standalone command-line programme MixMHCpred to run these predictions is provided free of charge for academic users to run predictions with the model trained on HLA peptidomics data, including 19 additional alleles with MS data in IEDB but not present in our datasets to further expand the coverage of HLA alleles to 71.

Predictions of neo-antigens in melanoma samples – cross-patient validation

To assess how much our improved predictions of neo-antigens in the three melanoma samples of [15] depend on HLA peptidomics data generated from these samples, we performed a careful cross-sample validation. For each of the three samples where neo-eptiopes had been identified (Mel15, Mel8, Mel5) [15], we re-run our entire pipeline (i.e., identification of HLA-I motifs across HLA peptidomics datasets + construction of PWMs for each allele) without the HLA peptidomics data coming from this sample. The PWMs were then used to rank all possible peptides (9- and 10-mers) encompassing each mutation. Overall, the predictions change very little, especially for neo-antigens with verified immunogenicity (Supplementary Table 3). In one case (`ETS`**`K`**`QVTRW` for Mel5), our method failed. This is because HLA-A25:01 was only present in Mel5 and therefore no longer identified by our algorithm when removing HLA peptidomics data from Mel5. Of note, this peptide was not well predicted by other methods[6,10,29] either.

Analysis of HLA sequences and structures

HLA-I sequences were retrieved from IMGT database[3]. All protein structures analysed in this work were downloaded from the PDB. Residues forming the P2 binding site in HLA-B14:02 (PDB: 3BVN[43]) were determined using a standard cut-off of 5Å from any heavy atoms of arginine at P2.

Binding stability assays for HLA-B14:02 wt and W121R mutant

W121R mutation was introduced into HLA-B14:02 wt by overlap extension PCR and confirmed by DNA sequencing. BL21(DE3)pLys bacterial cells were used to produce HLA-B14:02 wt and W121R as inclusion bodies. Four peptides with histidine or arginine at P2 (A[H/R]TKPRPAL, G[H/R]YDRSKSL, A[H/R]FAKSISL, H[H/R]FEKAVTL) were synthesized at the Peptide Facility (UNIL, Lausanne) with free N and C-termini (1mg of each peptide, > 80% purity). Peptides with histidine at P2 come from our HLA peptidomics data and were assigned to HLA-B15:18 by our mixture model algorithm. Based on our analysis of HLA sequence and structure, peptides with histidine at P2 are predicted to interact with HLA-B14:02 W121R, while peptides with arginine at P2 are predicted to bind better HLA-B14:02 wt.

Synthetic peptides were incubated separately with denaturated HLA-B14:01 wt and HLA-B14:01 W121R mutant 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 and samples were retrieved at time t=0h, 8h, 24h, 48h and t=72h. The known HLA-B14:02 ligand IRHENRMVL was used for positive controls (measured half-live of 248h). Negative controls consist of absence of peptides. k_{off} were determined by fitting exponential curves to the light intensity values obtained by ELISA at different time points. Half-lives were computed as $\ln(2)/k_{off}$. Values shown in Fig. 5d correspond to the average over two replicates. For two peptides showing exceptionally high binding stability, only lower bounds on half-lives could be determined (dashed lines in Fig. 5d).

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD005231.

The code to predict HLA ligands is available upon request at DG for non-commercial use.

Acknowledgment

We thank Nicholas MacGranahan and Charles Swanton for sharing with us the list of somatic mutations for the two lung cancer samples[30]. We are thankful to Camilla Jandus and Pedro Romero for sharing the B cell lines with us. We thank the Protein Analysis Facility of the University of Lausanne for technical help and access to MS instrumentation during the first period of this study. We thank Julien Racle and Santiago Carmona for insightful discussions about the manuscript. M.S. and D.G. acknowledge funding from CADMOS. All mass spectrometry analyses were supported by the Ludwig Institute for Cancer Research.

Author contributions

M.B.-S. contributed to the design of the study, performed HLA peptidomics experiments, analysed the data and wrote the manuscript. C.C. performed HLA peptidomics experiments. P.G. performed the mutagenesis and *in vitro* binding experiments. M.S. analysed the data. H.P. performed MS measurements. P.G. provided reagents. L.E.K provided reagents. G.C. provided reagents and contributed to the manuscript. D.G designed the study, developed and implemented the algorithms, analysed the data and wrote the manuscript.

Conflict of interest.

None.

References

1. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11,** 823–836 (2011).
2. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348,** 69–74 (2015).
3. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43,** D423–31 (2015).
4. Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* **6,** 132 (2005).

5. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12,** 1007–1017 (2003).
6. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv639
7. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64,** 177–186 (2012).
8. Sidney, J. *et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res* **4,** 2 (2008).
9. Gfeller, D., Bassani-Sternberg, M., Schmidt, J. & Luescher, I. F. Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology* e1177691 (2016). doi:10.1080/2162402X.2016.1177691
10. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* **8,** 33 (2016).
11. Yaciuk, J. C. *et al.* Direct interrogation of viral peptides presented by the class I HLA of HIV-infected T cells. *J. Virol.* **88,** 12992–13004 (2014).
12. McMurtrey, C. *et al.* Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. *Elife* **5,** 246 (2016).
13. Carreno, B. M. *et al.* A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* (2015). doi:10.1126/science.aaa3828
14. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515,** 572–576 (2014).
15. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* **7,** 13404 (2016).
16. Shraibman, B., Kadosh, D. M., Barnea, E. & Admon, A. Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. *Mol. Cell Proteomics* **15,** 3058–3070 (2016).
17. Bassani-Sternberg, M. & Coukos, G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr. Opin. Immunol.* **41,** 9–17 (2016).
18. Kalaora, S. *et al.* Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* **7,** 5110–5117 (2016).
19. Caron, E. *et al.* An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife* **4,** O111.011833 (2015).
20. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell Proteomics* **14,** 658–673 (2015).
21. Ritz, D. *et al.* 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* n/a–n/a (2016). doi:10.1002/pmic.201500445
22. Bassani-Sternberg, M. & Gfeller, D. Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions. *J. Immunol.* **197,** 2492–2499 (2016).

23. Mommen, G. P. M. *et al.* Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci. U.S.A.* **111,** 4507–4512 (2014).
24. Gloger, A., Ritz, D., Fugmann, T. & Neri, D. Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. *Cancer Immunol. Immunother.* **65,** 1377–1393 (2016).
25. Pearson, H. *et al.* MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126,** (2016).
26. Gfeller, D. *et al.* The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol.* **7,** 484–484 (2011).
27. Kim, T. *et al.* MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res.* **40,** e47 (2012).
28. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43,** D405–12 (2015).
29. Rasmussen, M. *et al.* Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J. Immunol.* **197,** 1517–1524 (2016).
30. Bentzen, A. K. *et al.* Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* **34,** 1037–1045 (2016).
31. Stewart-Jones, G. B. E. *et al.* Structures of three HIV-1 HLA-B*5703-peptide complexes and identification of related HLAs potentially associated with long-term nonprogression. *J. Immunol.* **175,** 2459–2468 (2005).
32. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. & Stevanović, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50,** 213–219 (1999).
33. Mester, G., Hoffmann, V. & Stevanovic, S. Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands. *Cell. Mol. Life Sci.* **68,** 1521–1532 (2011).
34. Andreatta, M., Lund, O. & Nielsen, M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* **29,** 8–14 (2013).
35. Rasmussen, M. *et al.* Uncovering the peptide-binding specificities of HLA-C: a general strategy to determine the specificity of any MHC class I molecule. *J. Immunol.* **193,** 4790–4802 (2014).
36. Schittenhelm, R. B., Dudek, N. L., Croft, N. P., Ramarathinam, S. H. & Purcell, A. W. A comprehensive analysis of constitutive naturally processed and presented HLA-C*04:01 (Cw4)-specific peptides. *Tissue Antigens* **83,** 174–179 (2014).
37. Giam, K. *et al.* A comprehensive analysis of peptides presented by HLA-A1. *Tissue Antigens* **85,** 492–496 (2015).
38. Trolle, T. *et al.* The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. *J. Immunol.* **196,** 1480–1487 (2016).
39. Dudley, M. E. *et al.* CD8+ enriched 'young' tumor infiltrating lymphocytes can mediate regression of metastatic melanoma. *Clin. Cancer Res.* **16,** 6122–6131 (2010).
40. Donia, M., Larsen, S. M., Met, O. & Svane, I. M. Simplified protocol for

clinical-grade tumor-infiltrating lymphocyte manufacturing with use of the Wave bioreactor. *Cytotherapy* **16,** 1117–1120 (2014).

41. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).
42. Nielsen, M. *et al.* Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20,** 1388–1397 (2004).
43. Kumar, P. *et al.* Structural basis for T cell alloreactivity among three HLA-B14 and HLA-B27 antigens. *J. Biol. Chem.* **284,** 29784–29797 (2009).
44. Chen, J.-L. *et al.* Structural and kinetic basis for heightened immunogenicity of T cell vaccines. *J. Exp. Med.* **201,** 1243–1255 (2005).

Figures

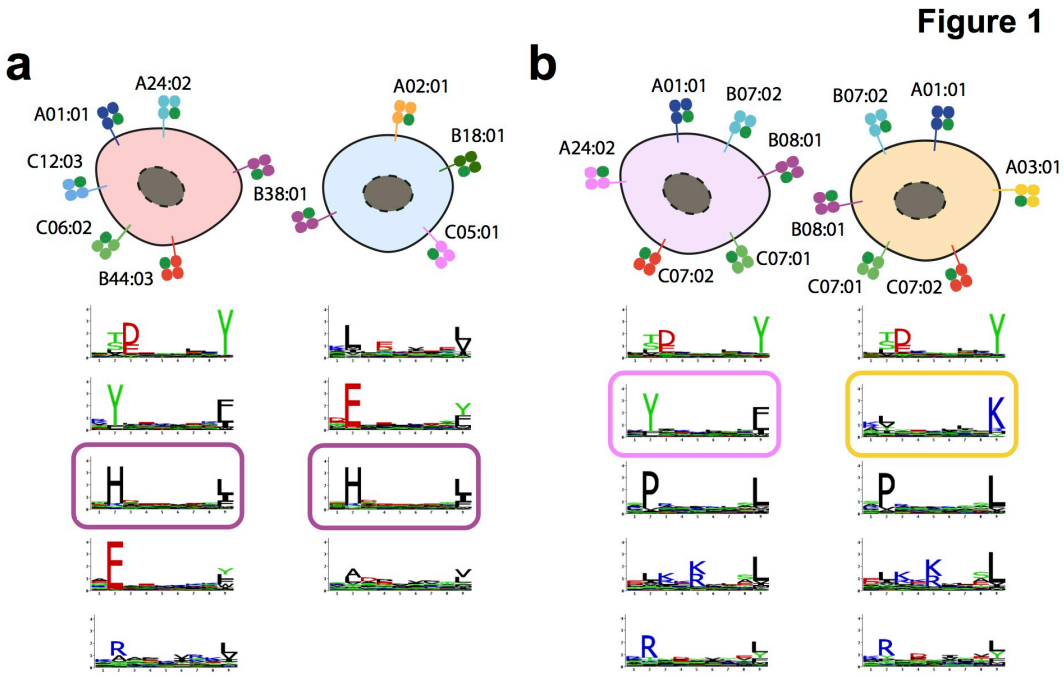


Figure 1: Exploiting co-occurrence of HLA-I alleles across samples to identify HLA-I binding motifs. **a**: Example of a pair of samples that share only one allele (HLA-B38:01). The sample on the left is a B cell line (GD149) and the sample on the right is a TIL cell line (TIL1) analysed in this work. **b**: Example of a pair of samples that share all except one allele in each sample (HLA-A24:02 and HLA-A03:01). The left and right samples come from melanoma tumors (Mel16 and Mel8, respectively).

Figure 2

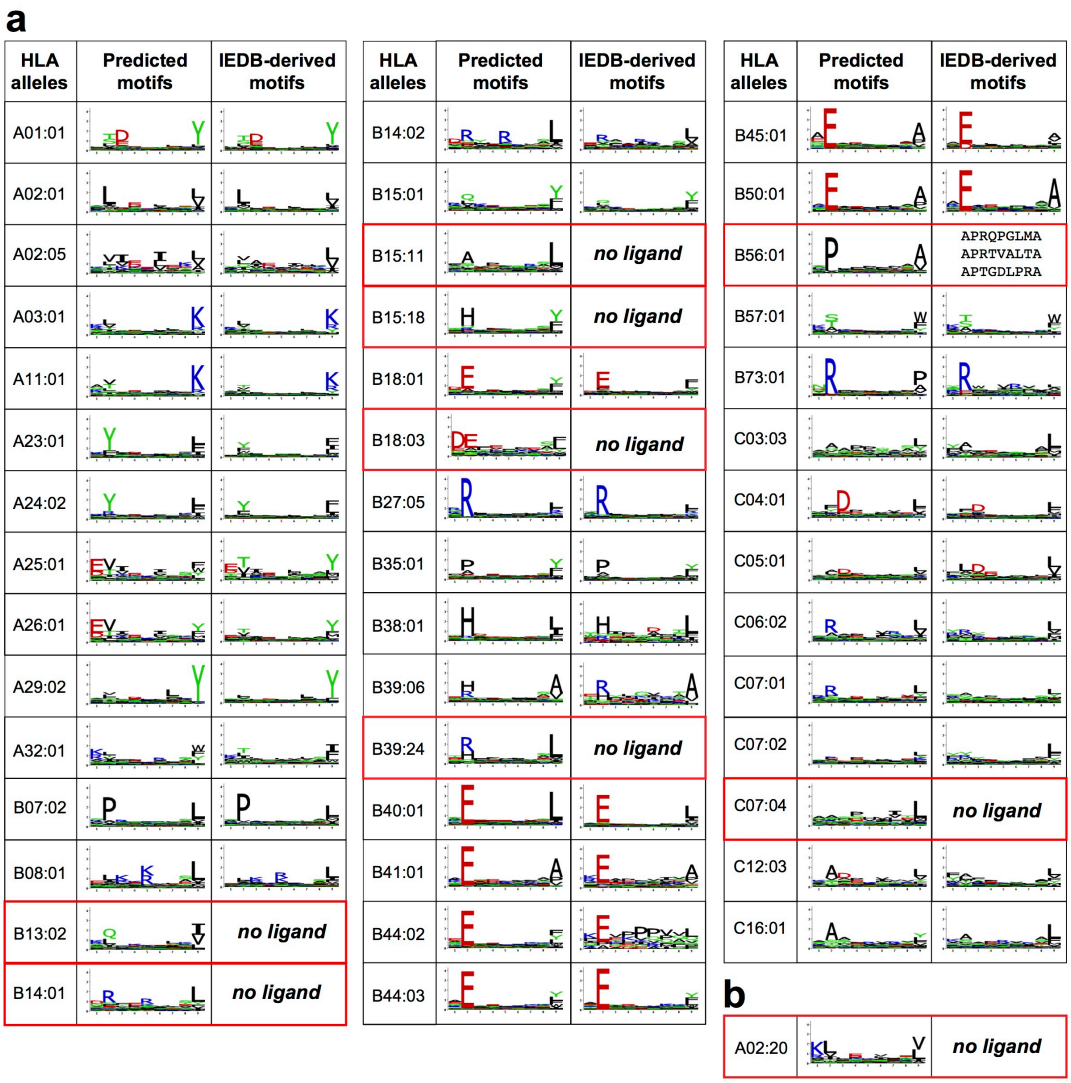


Figure 2: Comparison between motifs predicted by our algorithm and those from IEDB. **a**: 44 HLA-I binding motifs identified with the fully unsupervised approach. Alleles without previously documented ligands are highlighted in red. For HLA-B56:01, the three known ligands are shown. **b**: New motif identified with the semi-supervised approach for an additional allele without ligands in IEDB.

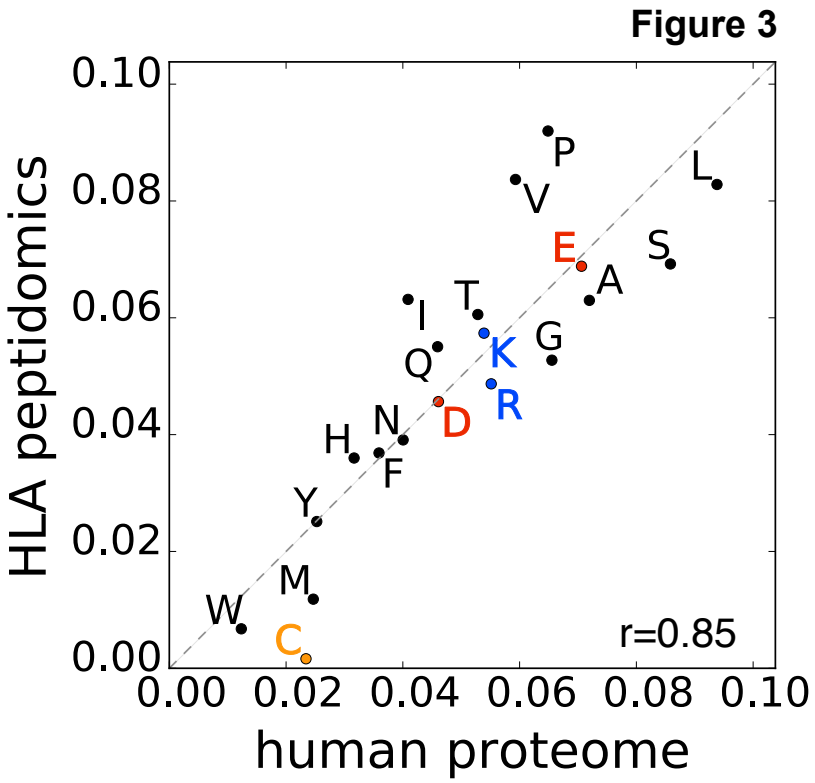


Figure 3: Correlation between amino acids frequencies at positions P4 to P7 in our HLA peptidomics data and in the human proteome.

Figure 4

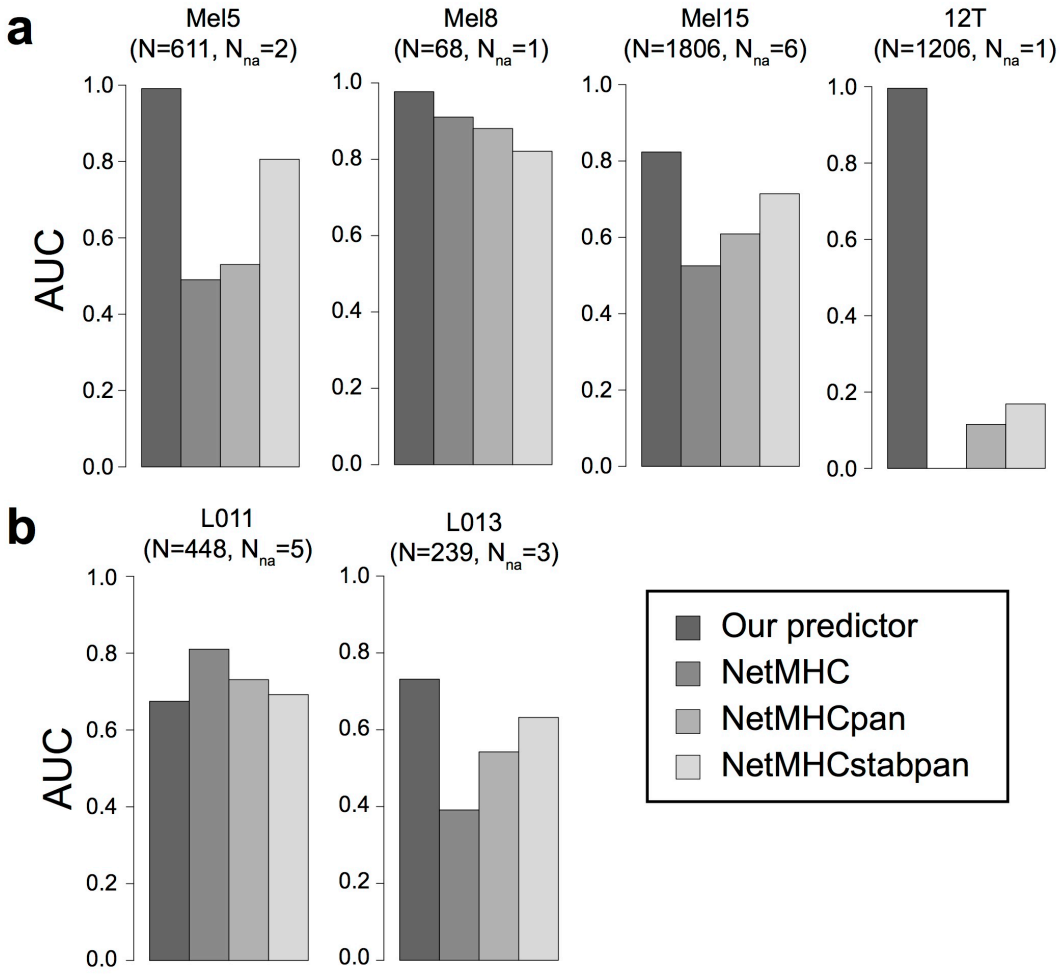
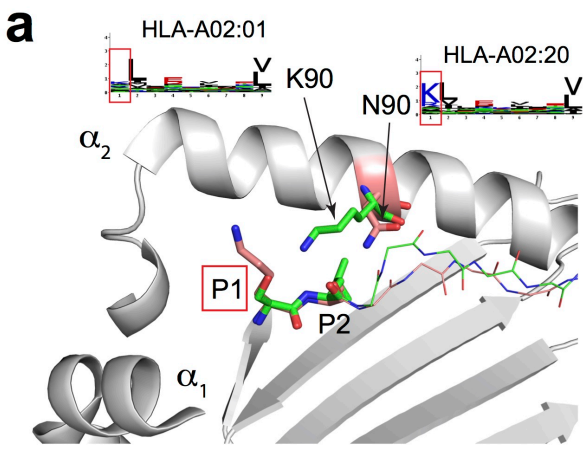


Figure 4: Comparison between neo-antigen predictions with our predictor trained on HLA peptidomics data and with other existing tools. **a**: AUC values for predictions of neo-antigens identified in four melanoma samples [15,18]. The neo-antigens ETS**K**QVTRW (Mel5, best predicted IC_{50} = 3231nM) and DAN**S**FLQSV (12T, best predicted IC_{50} = 6235nM) did not pass the standard NetMHC affinity threshold and were therefore poorly predicted by this tool. **b:** AUC values for predictions of neo-antigens identified in two lung cancer samples [30]. N indicates the number of mutated peptides that passed the NetMHC threshold (500nM). N_{na} shows the number of detected neo-antigens.

Figure 5



b

	P2 binding site										
Positions	31	33	48	50	58	60	69	87	90	91	121
HLA-B15:09	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B15:10	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B15:18	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B38:01	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B39:01	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B39:05	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B39:06	Y	Y	S	G	V	F	E	N	I	C	T
HLA-B39:09	Y	Y	S	G	V	F	E	N	I	C	R
HLA-B14:02	Y	Y	S	G	V	F	E	N	I	C	W

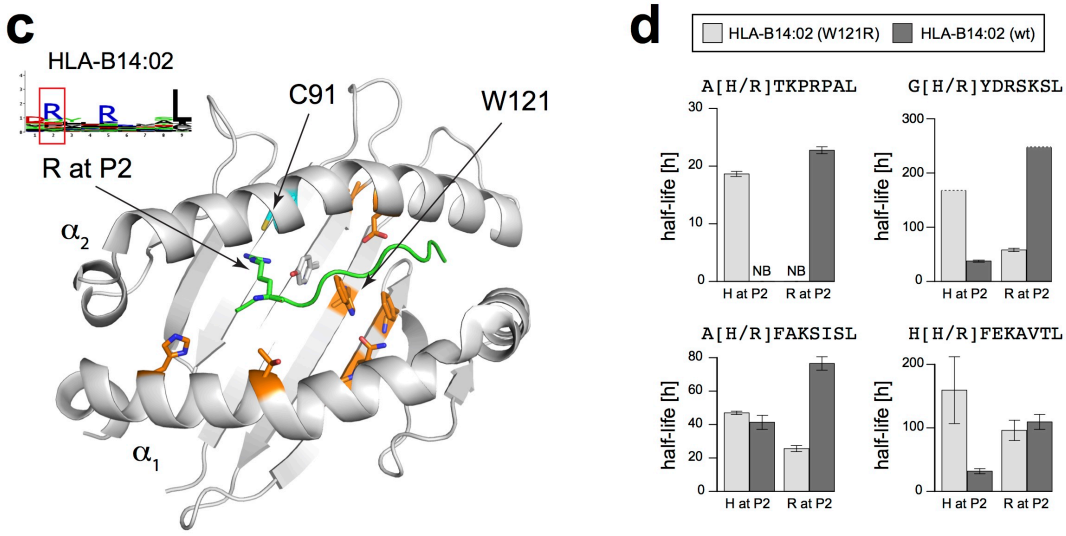


Figure 5: Analysis of newly identified HLA-I motifs. **a**: Structural view of two different HLA alleles with N90 (PDB: 2BVQ [31], pink sidechains) or K90 (PDB: 2BNR [44], green sidechains). For clarity, the α_1 helix has been truncated. **b:** P2 binding site residues' conservation across HLA-I alleles displaying preference for histidine at P2. The last line shows the sequence of HLA-B14:02, which does not show histidine preference at P2 (see motif in **c**). The last column shows amino acids at position 121, which is not part of the P2 binding site. **c:** Structural view of HLA-B14:02 in complex with a peptide with arginine at P2 (PDB: 3BVN [43]). Residues not conserved between HLA-B15:18 and HLA-B14:02 are displayed in orange. None of them are making direct contact with the arginine residue at P2. **d:** Stability values (half-lives) obtained for peptides with H or R at P2 for both HLA-B14:02 wt and W121R mutant. NB stands for no binding. Dashed lines indicate lower bounds for half-lives values.

Tables

Table 1: Ranking of the neo-antigens identified in four melanoma samples [15,18]. Column 5 shows the ranking based on our predictions. Column 6 shows the P-value estimate. Column 7 to 9 show the ranking based on NetMHC [6], NetMHCpan [10] and NetMHCstabpan [29], respectively. The last column shows the total number of neo-antigen candidates (i.e., all possible 9- and 10-mers encompassing each missense mutation).

Sample	Sequence	Protein	Muta-tion	Rank	P-value	Net-MHC	Net-MHC-pan	Net-MHC-stabpan	# Candi-dates
Mel8	SPGPVKLE**L**	NOP16	P169L	**2**	0.004	7	9	14	1340
Mel5	YIDE**R**FERY	SEPT2	Q125R	**3**	0.0002	13	20	148	25807
Mel15	GRIAF**F**LKY	SYTL4	S363F	**4**	0.0002	138	334	480	24766
12T	DAN**S**FLQSV	MED15	P747S	**6**	0.002	4186	2945	3102	15750
Mel15	**L**PIQYEPVL	SEC23A	P52L	**7**	0.0007	1715	882	36	24766
Mel5	ETS**K**QVTRW	GABPA	E161K	**13**	0.0008	2079	2682	576	25807
Mel15	KLKLP**I**IMK	AKAP6	M1482I	**20**	0.002	196	183	44	24766
Mel15	GRTGAGKS**F**L	ABCC2	S1342F	**261**	0.009	1154	1813	3111	24766
Mel15	K**L**ILWRGLK	NCAPG2	P333L	519	0.03	300	200	**199**	24766
Mel15	ASWVVPIDI**K**	MAP3K9	E689K	3944	0.18	**1645**	1667	2998	24766