nature biotechnology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41587-019-0289-6

Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes

Julien Racle[1,2], Justine Michaux[1,3], Georg Alexander Rockinger[1], Marion Arnaud[1,3], Sara Bobisse[1,3], Chloe Chong[1,3], Philippe Guillaume[1,3], George Coukos[1,3], Alexandre Harari[1,3], Camilla Jandus[1], Michal Bassani-Sternberg[1,3*] and David Gfeller[1,2*]

Predictions of epitopes presented by class II human leukocyte antigen molecules (HLA-II) have limited accuracy, restricting vaccine and therapy design. Here we combined unbiased mass spectrometry with a motif deconvolution algorithm to profile and analyze a total of 99,265 unique peptides eluted from HLA-II molecules. We then trained an epitope prediction algorithm with these data and improved prediction of pathogen and tumor-associated class II neoepitopes.

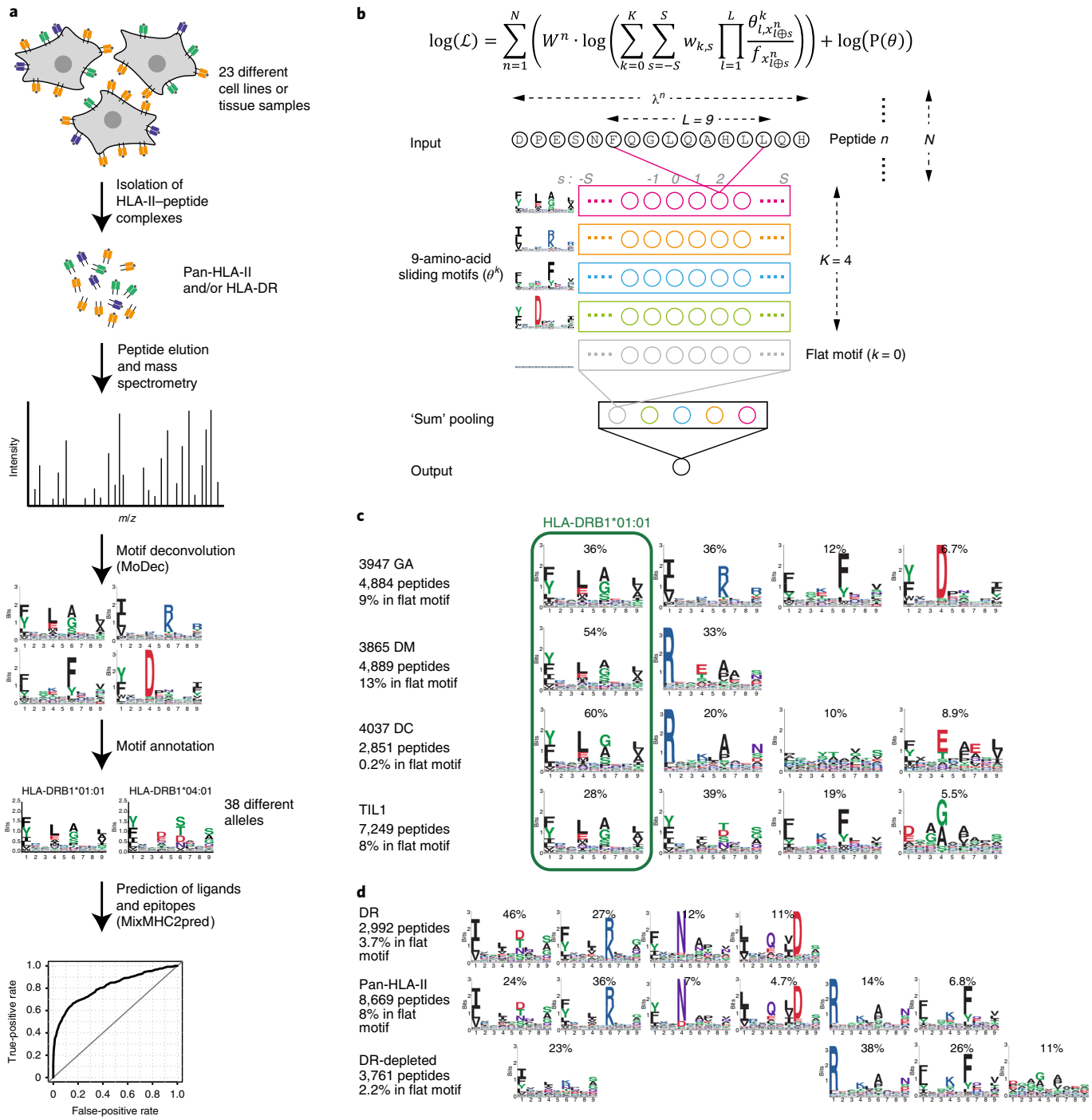
Antigen-presenting cells (APCs) display peptides bound to HLA-II on their surface. In infections or cancer, interactions between the T cell receptor on CD4+ T cells and HLA-II–peptide complex on APCs, presenting non-self peptides or tumor-associated antigens, are key to initiate and sustain an immune response[1–4]. Prediction and analysis of HLA-II epitopes is critical to vaccine design and targeted therapy development in immunology and cancer immunotherapy, but is challenging because HLA-II are highly polymorphic and the size of the peptides presented varies. For this reason, the core binding region of HLA-II ligands is difficult to determine, especially in mass spectrometry (MS) datasets of peptides eluted from HLA-II, as peptides from the same sample come from multiple HLA-II alleles. This issue becomes important when training HLA-II epitope prediction algorithms, which display low accuracy[5,6]. Recent developments in HLA peptidomics[7–9], allowing fast and reliable measurements of thousands of HLA ligands per sample, can improve epitope predictions, as shown for HLA-I molecules[10–14]. However, similar improvements have not been observed for HLA-II, and previous studies based on high-throughput peptidomics have been restricted to a few HLA-II alleles[8,15] or failed to demonstrate improvements in epitope predictions[16].

Here we profiled the HLA-II peptidome of 13 different cell lines or tissue samples and identified 40,864 unique HLA-II ligands. We combined this dataset with another one recently generated by our lab[7] to reach a total of 77,189 unique peptides from 23 different samples (Methods; Fig. 1a, Supplementary Table 1 and Supplementary Data 1), making it the largest dataset of HLA-II ligands available to date. To analyze these data, we developed MoDec, a motif deconvolution algorithm (Methods; Supplementary Code 1). Unlike previous approaches that attempted to align peptides[17,18], MoDec is a fully probabilistic framework that allows motifs to be found everywhere on the peptide sequences and learns both the motifs as well as their weights and preferred binding core position offsets (Fig. 1b). MoDec shows conceptual similarity with convolutional neural networks (each motif can be thought of as a filter) but provides direct interpretation and visualization of the results as sequence logos (Fig. 1b). Applying MoDec to our data, we found many motifs (Supplementary Fig. 1). HLA-II motifs identified across samples with shared HLA-II alleles displayed high similarity (Fig. 1c and Supplementary Fig. 2). This demonstrates high reproducibility of our motif deconvolution approach and enabled us to unambiguously annotate the different motifs to their respective alleles (Supplementary Fig. 2; Methods). Comparison with motifs from the Immune Epitope Database (IEDB)[19] or predictions from NetMHCIIpan[20] showed some similarity but also some important differences, including clearer anchor residues (Supplementary Fig. 2).

The most frequent motifs that we identified corresponded to HLA-DR alleles. To further validate them, we sequentially purified the HLA-DR molecules with an anti-HLA-DR antibody and then the remaining HLA-II molecules with a pan-HLA-II antibody (42,903 and 27,692 peptides, respectively, for a total of 99,265 unique peptides across all our samples; Supplementary Data 2). We observed that the motifs deconvolved from HLA-DR peptidomes are identical to those assigned to HLA-DR alleles from the pan-HLA-II peptidomes (Fig. 1d and Supplementary Fig. 3). In addition, in the HLA-DR-depleted samples, all motifs previously predicted to correspond to HLA-DP or HLA-DQ alleles could be found, indicating that our motif deconvolution approach is not restricted to HLA-DR motifs (Fig. 1d and Supplementary Fig. 3). In comparison to other motif deconvolution methods[18,21], MoDec shows improved resolution and is 100–10,000 times faster (Supplementary Fig. 4), making it particularly appropriate for large HLA-II peptidomics datasets.

We next investigated properties of naturally presented HLA-II ligands beyond the specificity of HLA-II alleles. We first grouped all of our HLA-II ligands and used MoDec to identify motifs occurring three amino acids upstream and downstream of the N or C terminus (Methods). Multiple motifs appeared with a specificity that was in general stronger for amino acids inside the peptides (Supplementary Fig. 5a,b). Some of these amino acid preferences had been suggested to represent different peptide processing and cleavage pathways[15,22,23], such as aminopeptidase trimming of class II ligands with a preference for proline near the N terminus[22]. All alleles showed proline enrichment two residues downstream of the N terminus and upstream of the C terminus (Supplementary Fig. 5c,d). Binding assays demonstrated similar binding between peptides containing proline or alanine at the second position (Supplementary Fig. 5e), validating the hypothesis that proline enrichment reflects peptide processing and loading[22]. We also observed conserved peptide length distributions across HLA-II alleles (Supplementary Fig. 6a),

[1]Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland. [3]Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University Hospital of Lausanne, Lausanne, Switzerland. *e-mail: michal.bassani@chuv.ch; david.gfeller@unil.ch



$$\log(\mathcal{L}) = \sum_{n=1}^{N}\left(W^n \cdot \log\left(\sum_{k=0}^{K}\sum_{s=-S}^{S} w_{k,s}\prod_{l=1}^{L}\frac{\theta^k_{l,x^n_{l\oplus s}}}{f_{x^n_{l\oplus s}}}\right)\right) + \log(\mathrm{P}(\theta))$$

Fig. 1 | Motif deconvolution in HLA-II peptidomics data. a, Description of our pipeline for HLA-II ligand isolation, motif deconvolution and training of an HLA-II ligand predictor. **b**, Top: log-likelihood optimized in MoDec (Methods). Bottom: a graphical interpretation of the model, including *K* sliding 9-amino-acid motifs (θ^k) and a sum pooling step over all positions (*s*). **c**, Motifs identified in four samples sharing exactly one allele (HLA-DRB1*01:01) and showing exactly one highly conserved motif. **d**, A comparison of the motifs found in HLA-II peptidomics, HLA-DR peptidomics and HLA-DR-depleted peptidomics in the same sample (3830-NJF).

unlike for HLA-I alleles[24], and conserved binding core offsets (Supplementary Fig. 6b).

We then took advantage of our deconvolved HLA-II peptidomics datasets to train a predictor of HLA-II ligands (MixMHC2pred; Methods; Supplementary Code 2). This predictor combines allele-specific motifs and allele-independent peptide N- and C-terminal motifs, peptide length and binding core offset preferences (Methods). We first performed predictions on multiple HLA-II peptidomics datasets from independent studies (Supplementary Table 2) and observed improved accuracy as compared to NetMHCIIpan[20] (Fig. 2a and Supplementary Fig. 7). The improvement of our predictor was particularly compelling when considering the positive predictive value for the top 2% of predictions, with an average of 0.90 for MixMHC2pred versus 0.67 for NetMHCIIpan (Supplementary Fig. 7). These results also show the superiority of the full predictor that combines the HLA-II motifs, N- and

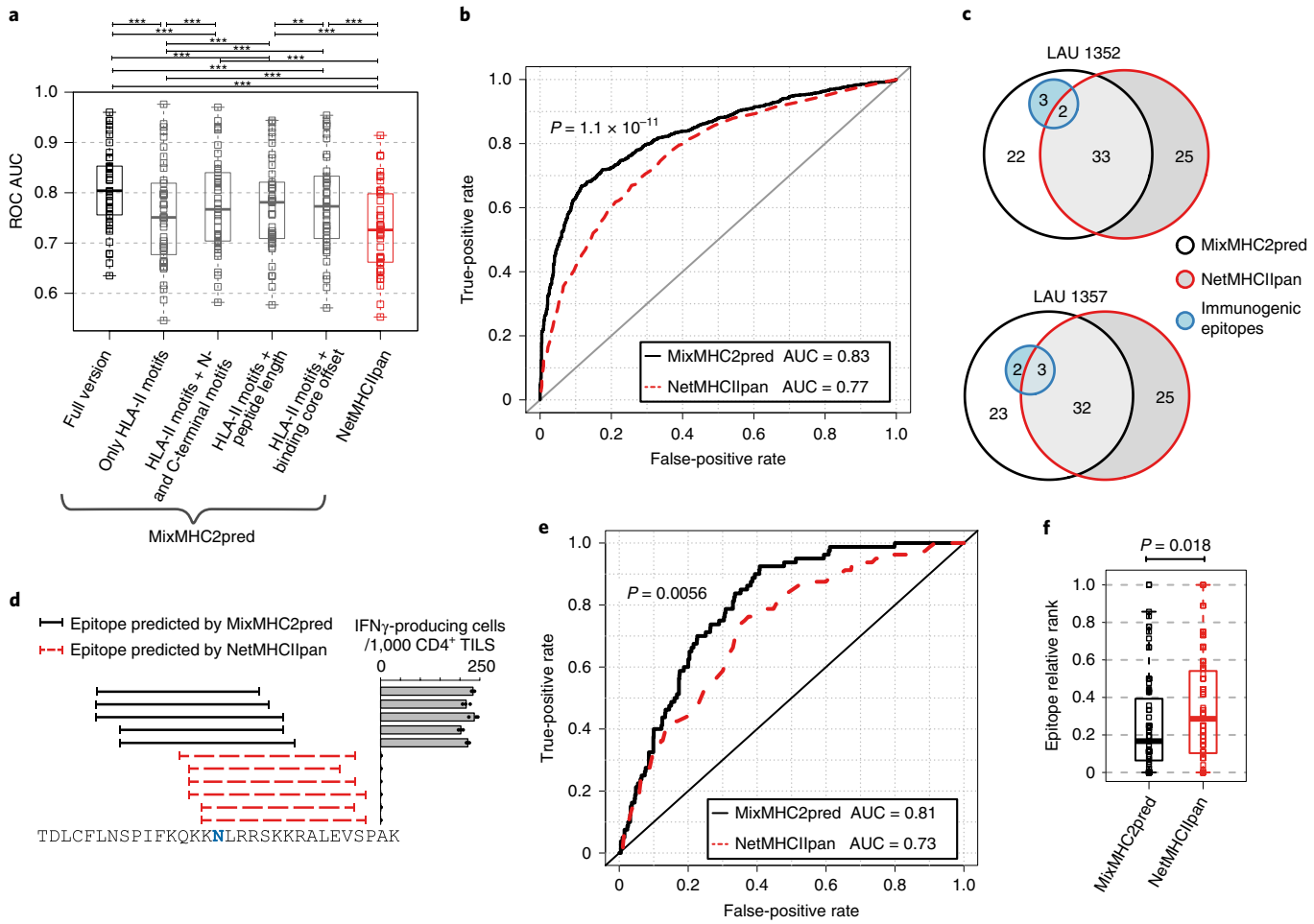


Fig. 2 | MixMHC2pred improves class II epitope prediction. a, Comparison of prediction accuracy of MixMHC2pred (and multiple variants) and NetMHCIIpan for HLA-II ligands ($n = 41$ biologically independent samples). **b**, Receiver operating characteristic (ROC) curve for the predictions of all class II epitopes from CD4+ T cell tetramer assays ($n = 2{,}359$ positive and negative epitopes). **c**, Euler diagram showing the number of tested and validated (blue circle) epitopes from viral, bacterial and melanoma-associated antigens in two patients with melanoma. **d**, Predictions of epitopes in a minigene encoding an ovarian cancer class II immunogenic mutation ($SGOL1_{D246N}$). The bar plot shows CD4+ T cell responses toward these peptides (mean values from $n = 3$ technical repeats, overlaid with each data point). **e**–**f**, Benchmarking of neoepitope predictions (Supplementary Data 3), based on ROC (**e**, $n = 929$ tested peptides from 13 different patients) and the epitope relative rank (**f**, $n = 80$ independent neoepitopes from 13 different patients). In **a** and **f**, box plots indicate the median, upper and lower quartiles; the results of a paired two-sided Wilcoxon signed rank-test are indicated (**$P < 0.01$, ***$P < 0.001$). In **b** and **e**, P values are based on a two-sided DeLong's test.

C-terminal motifs, peptide length and binding core offset, as compared to a predictor based only on the HLA-II motifs or a combination of the HLA-II motifs with any of the other allele-independent characteristics (Fig. 2a and Supplementary Fig. 7). To investigate whether MixMHC2pred was also appropriate for class II epitopes, we compiled all CD4+ T cell tetramer assays from the IEDB[19]. MixMHC2pred was significantly more accurate than NetMHCIIpan ($P = 1.1 \times 10^{-11}$; Fig. 2b and Supplementary Fig. 8a). The use of neural networks did not improve predictions, which is consistent with the very few interactions observed between positions in the binding core of HLA-II ligands (Supplementary Fig. 8b, Supplementary Table 3 and Supplementary Note). We further surveyed known melanoma-associated antigens and viral and bacterial proteins (Supplementary Table 4a). Top hits from MixMHC2pred and from NetMHCIIpan were tested for class II immunogenicity in two patients with melanoma (Supplementary Table 4b). Our results show a higher fraction of true positives among the predictions of MixMHC2pred versus NetMHCIIpan (Fig. 2c; Matthews correlation coefficients of 0.16 and 0.16 for MixMHC2pred versus −0.17 and −0.058 for NetMHCIIpan). The same peptides were also tested with CD4+ T cells from a healthy donor resulting in many correctly predicted epitopes and similar yield between the two predictors (Supplementary Fig. 9a). We next took advantage of a class II immunogenic substitution (D246N in SGOL1) that was recently identified in a patient with ovarian cancer by screening tumor-infiltrating lymphocytes (TILs) with minigenes (Methods; Supplementary Fig. 10). To determine the actual epitope, we applied MixMHC2pred and NetMHCIIpan on the 31-amino-acid oligomer encoded by the minigene. The results indicate that MixMHC2pred could predict the actual epitope (Fig. 2d and Supplementary Fig. 11). To further assess the use of MixMHC2pred for neoepitope predictions, we compiled recent class II neoantigen studies with available HLA-II typing (Supplementary Data 3). Here again, we observed significant improvements in predictions (Fig. 2e; $P = 0.0056$), with true epitopes in general in the top 16.7% relative to the tested non-immunogenic peptides for MixMHC2pred, but only in the top 28.6% for NetMHCIIpan (Fig. 2f and Supplementary Fig. 12; Methods).

By combining in-depth HLA-II peptidomics with a motif deconvolution algorithm, we could capitalize on unbiased MS profiling of HLA-II ligands for class II epitope predictions. The very high

similarity between HLA-DR motifs identified in pan-HLA-II and HLA-DR peptidomes shows that HLA-DR motifs can be accurately resolved in the pan-HLA-II samples with MoDec. This suggests that monoallelic samples are not needed to determine HLA-DR motifs. Whether increased detection efficacy could be obtained by using anti-HLA-DP and anti-HLA-DQ antibodies remains to be seen, and our motif deconvolution tool may prove highly valuable to analyze such data. Our approach may not capture the full complexity of class II antigen presentation[1], and does not include any information about immunogenicity. Moreover, we cannot exclude the possibility that the use of ligand data also has impacts on our definition of HLA-II motifs, and the latter may therefore not perfectly model the experimental binding stability of peptides, which is important for immunogenicity[25]. This may explain the remaining false positives in our predictions. Despite these limitations, the use of MS data enabled us to integrate unbiased HLA-II motifs together with N- and C-terminal motifs, as well as peptide length and binding core offset preferences, which led to enrichment in true positives among candidate epitopes. The large allele coverage (especially for HLA-DR) makes our predictor suitable for a wide range of applications in infectious diseases, autoimmunity and cancer immunotherapy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0289-6.

Received: 15 February 2019; Accepted: 11 September 2019;
Published online: 14 October 2019

References

1. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
2. Khodadoust, M. S. et al. *Nature* **543**, 723–727 (2017).
3. Linnemann, C. et al. *Nat. Med.* **21**, 81–85 (2015).
4. Kreiter, S. et al. *Nature* **520**, 692–696 (2015).
5. Andreatta, M. et al. *Immunogenetics* **67**, 641–650 (2015).
6. Andreatta, M. et al. *Bioinformatics* **34**, 1522–1528 (2018).
7. Chong, C. et al. *Mol. Cell. Proteomics* **17**, 533–548 (2018).
8. Ritz, D. et al. *Proteomics* **18**, 1700246 (2018).
9. Bassani-Sternberg, M. et al. *Nat. Commun.* **7**, 13404 (2016).
10. Bassani-Sternberg, M. & Gfeller, D. *J. Immunol.* **197**, 2492–2499 (2016).
11. Bassani-Sternberg, M. et al. *PLoS Comput. Biol.* **13**, e1005725 (2017).
12. Abelin, J. G. et al. *Immunity* **46**, 315–326 (2017).
13. Jurtz, V. et al. *J. Immunol.* **199**, 3360–3368 (2017).
14. Bulik-Sullivan, B. et al. *Nat. Biotechnol.* **37**, 55–63 (2019).
15. Barra, C. et al. *Genome Med.* **10**, 84 (2018).
16. Garde, C. et al. *Immunogenetics* **71**, 445–454 (2019).
17. Nielsen, M. & Andreatta, M. *Nucleic Acids Res.* **45**, W344–W349 (2017).
18. Andreatta, M., Alvarez, B. & Nielsen, M. *Nucleic Acids Res.* **45**, W458–W463 (2017).
19. Vita, R. et al. *Nucleic Acids Res.* **47**, D339–D343 (2019).
20. Jensen, K. K. et al. *Immunology* **154**, 394–406 (2018).
21. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. in *Proc. Second International Conference on Intelligent Systems for Molecular Biology* (eds Altman, R., Brutlag, D., Karp, P., Lathrop, R., & Searls, D.) 28–36 (AAAI, 1994).
22. Falk, K., Rötzschke, O., Stevanovíc, S., Jung, G. & Rammensee, H.-G. *Immunogenetics* **39**, 230–242 (1994).
23. Ciudad, M. T. et al. *J. Leukoc. Biol.* **101**, 15–27 (2017).
24. Gfeller, D. et al. *J. Immunol.* **201**, 3705–3716 (2018).
25. Yin, L., Calvo-Calle, J. M., Dominguez-Amorocho, O. & Stern, L. J. *J. Immunol.* **189**, 3983–3994 (2012).

Acknowledgements

We thank the Center of Experimental Therapeutics team for providing us with the patient-derived tissue samples and T cells. We thank P. Romero from the University of Lausanne for sharing the B cell lines with us. We thank R. T. Daniel and M. Hegi from the University Hospital of Lausanne for providing us with the collection of meningioma tissues. We thank M. Solleder for help with the visualization of motifs with ggseqlogo, F. Marino for technical support with sample preparation, H.-S. Pak for MS measurements and R. Genolet for HLA typing. This work was supported by the Swiss Cancer League (grant KFS-4104-02-2017 to D.G. and J.R.), the Ludwig Institute for Cancer Research, the ISREC Foundation thanks to a donation from the Biltema Foundation (to J.M., C.C. and M.B.-S.) and by the MEDIC foundation (to G.A.R. and C.J.).

Author contributions

J.R. developed the computational methods; J.R. and D.G. analyzed the data; J.M., C.C. and M.B.-S. generated the MS peptidomics data; G.A.R., M.A., S.B., P.G., A.H. and C.J. performed the binding and T cell assays; G.C., A.H., C.J. and M.B.-S. provided reagents; J.R., M.B.-S. and D.G. designed the study; and J.R., M.B.S. and D.G. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0289-6.

Correspondence and requests for materials should be addressed to M.B.-S. or D.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cells and patient material. Epstein–Barr-virus-transformed human B-cell lines JY (ATCC, 77442), CD165, PD42, CM467, RA957, BP455, GD149 (a gift from P. Romero (Ludwig Cancer Research Lausanne)), were maintained in RPMI-1640 + GlutaMAX medium (Life Technologies) supplemented with 10% heat-inactivated FBS (Dominique Dutscher) and 1% penicillin–streptomycin solution (BioConcept). Cells were grown to the required cell numbers, collected by centrifugation at 1,200 rpm for 5 min, washed twice with ice cold PBS and stored as dry cell pellets at −20 °C until use.

T cells were expanded from two melanoma tumors as previously described[7,11] and following established protocols[26,27]. In brief, fresh tumor samples were cut in small fragments and placed in 24-well plates containing RPMI CTS grade (Life Technologies), 10% human serum (Valley Biomedical), 0.025 M HEPES (Life Technologies), 55 µmol l^{-1} 2-mercaptoethanol (Life Technologies) and supplemented with a high concentration of interleukin (IL)-2 (Proleukin, 6,000 IU ml^{-1}; Novartis) for 3–5 weeks. TILs (25×10^6) were stimulated with irradiated feeder cells, anti-CD3 (OKT3, 30 ng ml^{-1}; Miltenyi biotec) and high-dose IL-2 (3,000 IU ml^{-1}) for 14 d. The cells were washed using a cell harvester (LoVo, Fresenius Kabi). Finally, the cells were washed with PBS on ice, aliquoted to a cell count of 1×10^8 and stored as dry pellets at −80 °C until use.

Snap-frozen meningioma tissues from patients (3808-HMC, 3830-NJF, 3849-BR, 3865-DM, 3869-GA, 3911-ME, 3912-BAM, 3947-GA, 3971-ORA, 3993, 4001, 4021, 4037-DC and 4052-BA) were obtained from the Centre Hospitalier Universitaire Vaudois (CHUV, Lausanne, Switzerland). Informed consent of the participants was obtained following requirements of the Institutional Review Board (Ethics Commission, CHUV). Protocol F-25/99 has been approved by the local ethics committee and the biobank of the Lab of Brain Tumor Biology and Genetics. Protocol 2017-00305 for antigen and T cell discovery in tumors has been approved by the local ethics committee.

HLA typing. Genomic DNA was extracted using DNeasy kit from Qiagen and 500 ng of genomic DNA was used to amplify HLA genes by PCR. High-resolution 4-digit HLA typing was performed with the TruSight HLA v2 Sequencing Panel from Illumina on a MiniSeq instrument (Illumina). Sequencing data were analyzed with the Assign TruSight HLA v.2.1 software (Illumina) and are provided as Supplementary Table 1.

Generation of antibody-crosslinked beads. Anti-pan-HLA-II and anti-HLA-DR monoclonal antibodies were purified from the supernatant of HB145 (ATCC, HB-145) and HB298 cells (ATCC, HB-298), respectively, grown in CELLLine CL-1000 flasks (Sigma-Aldrich) using protein A-sepharose 4B beads (pro-A beads; Invitrogen). Antibodies were crosslinked to pro-A beads at a concentration of 5 mg of antibodies per milliliter of beads. For this purpose, the antibodies were incubated with pro-A beads for 1 h at room temperature. Chemical crosslinking was performed by addition of dimethyl pimelimidate dihydrochloride (Sigma-Aldrich) in 0.2 M sodium borate buffer, pH 9 (Sigma-Aldrich) at a final concentration of 20 mM for 30 min. The reaction was quenched by incubation with 0.2 M ethanolamine, pH 8 (Sigma-Aldrich) for 2 h. Crosslinked antibodies were kept at 4 °C until use.

Purification of HLA-II and HLA-DR peptides. Cells were lysed in PBS containing 0.25% sodium deoxycholate (Sigma-Aldrich), 0.2 mM iodoacetamide (Sigma-Aldrich), 1 mM EDTA, 1:200 protease inhibitors cocktail (Sigma-Aldrich), 1 mM phenylmethylsulfonylfluoride (Roche) and 1% octyl-beta-D-glucopyranoside (Sigma-Aldrich) at 4 °C for 1 h. The lysis buffer was added to cells at a concentration of 1×10^8 cells per milliliter. Cell lysates were cleared by centrifugation with a table-top centrifuge (Eppendorf Centrifuge) at 4 °C and 14,200 rpm for 50 min. Meningioma tissues were placed in tubes containing the same lysis buffer and homogenized on ice in three to five short intervals of 5 s each using an Ultra Turrax homogenizer (IKA) at maximum speed. For 1 g of tissue, 10–12 ml of lysis buffer was required. Cell lysis was performed at 4 °C for 1 h. Tissue lysates were cleared by centrifugation at 20,000 rpm in a high-speed centrifuge (Beckman Coulter, JSS15314) at 4 °C for 50 min. The cells and tissue lysates were loaded on stacked 96-well single-use microplates (3-µm glass fiber, 10-µm polypropylene membranes; cat no. 360063, Seahorse Bioscience). Purification of pan-HLA-II peptides was performed following depletion of HLA-I as previously described[7]. For the sequential purification of HLA-DR and HLA-II from tissues, three plates were used. The first plate contained pro-A beads (Invitrogen) for depletion of antibodies (pre-clear plate), the second plate contained the same beads crosslinked to the anti-HLA-DR monoclonal antibodies and the third plate contained the beads crosslinked to anti-HLA-II monoclonal antibodies. For the sequential purification of HLA-DR and HLA-II from cells, only the last two plates were used. The Waters Positive Pressure-96 processor (Waters) was employed. The second and third plates were washed separately four times with 2 ml of 150 mM sodium chloride (Carlo-Erba) in 20 mM Tris-HCl, pH 8, four times with 2 ml of 400 mM NaCl in 20 mM Tris-HCl, pH 8 and four times again with 2 ml of 150 mM NaCl in 20 mM Tris-HCl, pH 8. Finally, the plates were washed twice with 2 ml of 20 mM Tris-HCl, pH 8. Each affinity plate was stacked on top of a Sep-Pak tC_{18} 100 mg Sorbent 96-well plate (cat. no. 186002321, Waters) pre-equilibrated with 1 ml of 80% acetonitrile (ACN) in 0.1 % trifluoroacetic acid (TFA) and with 2 ml of 0.1% TFA. The HLA and peptides were eluted with 500 µl of 1% TFA into the Sep-Pak plate and then we washed this plate with 2 ml of 0.1% TFA. Thereafter, HLA-II and HLA-DR peptides were eluted with 500 µl of 32% ACN in 0.1% TFA into a collection plate. Recovered HLA-II and HLA-DR peptides were dried using vacuum centrifugation (Concentrator plus, Eppendorf) and stored at −20 °C.

The data from the third plate in the sequential purification (that we termed 'HLA-DR-depleted') are meant to be enriched in HLA-DP- and HLA-DQ-bound peptides, but not to lead to a full depletion of HLA-DR-bound peptides owing to the amount of antibodies used in the purification and the incubation time during the affinity purification. This is confirmed by the clear presence of HLA-DR motifs in the deconvolution (Supplementary Fig. 3).

LC–MS/MS analyses of HLA-II peptides. Before MS analysis, HLA-II and HLA-DR peptide samples were resuspended in 10 µl of 2% ACN in 0.1% formic acid (FA) and aliquots of 3 µl for each MS run were placed in the Ultra HPLC autosampler. HLA peptides were separated by nanoflow HPLC (Proxeon Biosystems, Thermo Fisher Scientific) coupled on-line to a Q Exactive HF or HFX mass spectrometers (Thermo Fisher Scientific) with a nanoelectrospray ion source (Proxeon Biosystems). We packed a 20-cm-long, 75-µm-inner-diameter column with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH) in buffer A (0.5% acetic acid). Peptides were eluted with a linear gradient of 2–30% buffer B (80% ACN and 0.5% acetic acid) at a flow rate of 250 nl min^{-1} over 90 min. Data were acquired using a data-dependent 'top 10' method. Full-scan MS spectra were acquired at a resolution of 70,000 at 200 m/z with an auto gain control (AGC) target value of 3×10^6 ions. The ten most abundant ions were sequentially isolated, activated by higher-energy collisional dissociation and accumulated to an AGC target value of 1×10^5 with a maximum injection time of 120 ms. In the case of assigned precursor ion charge states of one, and from six and above, no fragmentation was performed. MS/MS resolution was set to 17,500 at 200 m/z. Selected ions were dynamically excluded for additional fragmentation for 20 s. The peptide match option was disabled. The raw files and MaxQuant output tables have been deposited to the ProteomeXchange Consortium via the PRIDE[28] partner repository with the dataset identifier PXD012308.

Peptide identification. We employed the MaxQuant platform[29] v.1.5.5.1 to search the peak lists against a fasta file containing the human proteome (Homo_sapiens_UP000005640_9606, the reviewed part of UniProt, with no isoforms, including 21,026 entries downloaded in March 2017) and a list of 247 frequently observed contaminants. Peptides with a length between 8 and 25 amino acids were allowed. The second peptide identification option in Andromeda was enabled. The enzyme specificity was set as unspecific. A false-discovery rate of 1% was required for peptides and no protein false-discovery rate was set. 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. Methionine oxidation and N-terminal acetylation were set as variable modifications. The peptide output files summarizing MaxQuant result files are provided as Supplementary Data 1 and 2.

Motif deconvolution algorithm for HLA-II peptidomics. In HLA-II peptidomics data, the HLA-II ligands are coming from different alleles, are of different lengths and their binding core positions are a priori unknown. To account for this, and building upon the successful application of the mixture model to HLA-I peptidomics[10], we developed a probabilistic framework able to learn multiple motifs anywhere on the peptides, as well as the weights and binding core offsets of these motifs. The log-likelihood is given by the following equation (see also Fig. 1b):

$$\log(\mathcal{L}) = \sum_{n=1}^{N} \left(W^n \cdot \log \left(\sum_{k=0}^{K} \sum_{s=-S}^{S} w_{k,s} \prod_{l=1}^{L} \frac{\theta^k_{l,x^n_{l\oplus s}}}{f_{x^n_{l\oplus s}}} \right) \right) + \log(P(\theta)) \qquad (1)$$

where N is the number of peptides; W^n is the similarity weight of the nth peptide (see below); K is the number of motifs; S is the maximal binding core offset ($S = \left[\frac{max_n(\lambda^n)-L}{2}\right]$, where λ^n is the length of the nth peptide); $w_{k,s}$ is the weight of motif k with binding core offset s ($\sum_k \sum_s w_{k,s} = 1$); L is the motif length (equal to nine here as HLA-II ligands are known to bind with a 9-amino-acid core (Supplementary Note and Supplementary Fig. 13)); $\theta^k_{l,i}$ represents the motifs (with $\sum_i \theta^k_{l,i} = 1$; $k = 0$ is a special case of a flat motif, $\theta^0_{l,i}\text{def} = h_i$, where h_i are the amino acid frequencies in the human proteome (this motif is used to model potential contaminant peptides)); x^n_j indicates which amino acid is found in peptide n at the position j (when x^n_j is not defined (that is, $j < 1$ or $j > \lambda^n$), then $\theta^k_{l,x^n_j}\text{def} = 0$); f_i is the expected background frequency in HLA-II peptidomics data for amino acid i; and P(θ) is a Dirichlet prior term (with the hyperparameter equal to 0.1)[30]. The '$l \oplus s$' in $x^n_{l\oplus s}$ is a 'special sum' that ensures that the binding core offsets are symmetric around 0 for each peptide (see Supplementary Note). A peptide similarity weight is given by $W^n = 1/S^n_{sim}$, where S^n_{sim} represents the average number of times each 9-amino-acid oligomer from the nth peptide is observed in the full dataset. This is useful as multiple overlapping HLA-II ligands from the same source region are typically found by MS.

Unlike the previous approaches for HLA-I[10,11], our model does not need previous peptide alignment, learns the binding core offsets of the motifs and includes peptide similarity weights. MoDec estimates the parameters θ and w based on expectation maximization (Supplementary Note). Although our framework is fully probabilistic, peptide responsibilities are derived during the expectation maximization, and these can be used to predict to which motif each peptide is most likely associated and with which binding core offset. Multiple runs (250 in this study) are performed by MoDec to optimize the log-likelihood of equation (1), starting from different initial conditions, considering all peptides of length 12 or more. As HLA-DR ligands have preference for hydrophobic amino acids at position 1, we implemented the possibility to include such a bias in a subset of the initial conditions used in the optimization by MoDec.

HLA-II motifs determined by MoDec are visualized with ggseqlogo[31].

The optimal number of motifs (K) was first determined using the Akaike information criterion (AIC):

$$\begin{aligned}\mathrm{AIC} = 2 \cdot n_{\mathrm{params}} - 2 \cdot \log(\mathcal{L}) = 2 \cdot (K \cdot (n_{\mathrm{aa}} - 1) \cdot L \\ +(2 \cdot S + 1) \cdot (K + 1) - 1) - 2 \cdot \log(\mathcal{L})\end{aligned} \quad (2)$$

where n_{params} is the number of free parameters, n_{aa} is the number of different amino acids (20) and the other parameters have been defined earlier. This AIC is commonly used in information theory to determine the information gained from using a model with more parameters over a simpler model (the smaller the AIC value the better). However, as with HLA-I peptidomics data[10] (and more generally with many clustering approaches), the optimal number of motifs is difficult to determine in a fully unsupervised way. We therefore explored additional motifs, tried further splitting specific motifs and manually curated each dataset for consistency across samples. Comparison between the numbers of motifs manually curated or determined by the AIC showed that in most cases the correct number of motifs would have been found with the AIC and in 80% of the cases the error when using this criterion would be at most of one motif (Supplementary Fig. 14), suggesting that AIC is a good starting point for the selection of the optimal number of motifs.

Assignment of motifs to alleles. To annotate the different motifs to their respective alleles, we used an iterative approach: we considered all samples that share a given allele and determined whether a motif was shared between all these samples. To decide which motifs are shared, we used the Kullback–Leibler divergence (KLD) between the motifs $\left(\mathrm{KLD}(k,m) = -\sum_l \sum_i \theta^k_{l,i} \cdot \log\left(\frac{\theta^m_{l,i}}{\theta^k_{l,i}}\right) < \mathrm{KLD}_{\mathrm{thr}}\right)$ for a given threshold ($\mathrm{KLD}_{\mathrm{thr}}$ varied between 1 and 1.75 depending on the iteration). Each iteration consists of checking for each allele, one after the other, if we can assign a motif to this allele. In the first five iterations, at least 75% of the samples containing the given allele had to share a motif to assign this motif to the given allele (in later iterations, the threshold of samples is decreased to 60%), with the additional requirement that no other allele was shared by these samples. By repeating these iterations multiple times, various motifs could be annotated to their respective allele. All annotations were further manually curated, allowing, for example, the annotation of HLA-DRB1*01:02, which is highly similar to HLA-DRB1*01:01 but with only the PD42 sample expressing this allele (Supplementary Fig. 2). We could then observe that the KLDs between each pair of motifs was significantly lower between motifs of the same allele or allele supertype than between different alleles (Supplementary Fig. 15a; $P < 0.001$).

We also explored an alternative approach to determine the motif corresponding to each allele, merging the data from multiple samples sharing a given allele together instead of using MoDec on each sample separately. Although this alternative approach allows more peptides per allele, it also requires deconvolving a larger number of specificities, which adds complexity to the deconvolution. In practice, we could nicely recover most of the HLA-DR motifs, but failed to accurately identify most HLA-DP and HLA-DQ motifs, and did not find any additional or improved motifs (Supplementary Note and Supplementary Fig. 15b).

For the binding motifs from the IEDB[19] database, we downloaded the full MHC ligand data (http://www.iedb.org/database_export_v3.php, 28 January 2018) and filtered this data to remove peptides obtained from MS (as for many of them, allele restriction information is based on predictions) and keep only the peptides described as 'Positive-High' binders. MoDec was run considering a single motif on the resulting list of peptides per allele. The corresponding motifs are shown in Supplementary Fig. 2.

Binding motifs from NetMHCIIpan[20] were determined in the following way: 16,000 random human peptides (2,000 of each length between 12 and 19 amino acids) were input into NetMHCIIpan. For each allele, the peptides with a '%Rank' better than five were kept and ggseqlogo[31] was used to draw the motifs from the corresponding 'Core' sequences returned by NetMHCIIpan.

Comparison to other motif deconvolution methods. We compared the motifs found by MoDec with those predicted by Gibbscluster (v.2.0 (ref. [18])) and MEME v.4.12 (ref. [21]) (Supplementary Fig. 4). For a comparison of the timing, the tools were launched on a single 3.3 GHz CPU with 2 GB of RAM, searching for one to eight motifs in various samples.

Gibbscluster was run with the recommended parameters for HLA-II (five seeds for the initial conditions, an initial MC temperature of 1.5, using a trash cluster with a threshold of two for this cluster, the rest being left unchanged). Gibbscluster suggests using a Kullback–Leibler criterion to select the optimal number of clusters from their deconvolution. In some cases, manual curation allowed us to find additional clusters that had similarity with known HLA-II motifs, and we show the results of Gibbscluster including these additional clusters in Supplementary Fig. 4.

MEME was run setting a motif width of nine, a maximum dataset size of 10,000,000 (needed owing to the size of the samples) and the rest was left at default.

Investigation of the properties of HLA-II ligands other than binding specificity. Analysis of N- and C-terminal flanking motifs was done with MoDec by taking the three amino acids upstream and downstream of the N and C termini of the peptides that could be assigned to alleles (that is, the peptides were extended on the basis of their protein of origin to include the three amino acids upstream of the N terminus and downstream of the C terminus).

The peptide length distributions, binding core offset distributions and frequencies of proline two residues downstream of the N terminus and upstream of the C terminus, were computed for all peptides associated to each allele.

Binding affinity assays. Peptide binding affinity (Supplementary Fig. 5e) was assessed by peptide competition assay. For each peptide, eight wells of a v-bottom 96-well plate (Greiner Bio-One) were filled with 100 µl of each recombinant 'empty' DR1, DR4 or DR7 protein (1 µg) in a citrate saline buffer (100 mM citrate, pH 6.0), with 0.2% β-octyl-glucopyranoside (Calbiochem), 1× complete protease inhibitors (Roche) and 2 µM FLAG-$\mathrm{HA}_{307-319}$ peptide. Competitor peptides (10 mM DMSO solution) were added to each well to a final concentration of 100, 33, 11, 3.7, 1.2, 0.4, 0.1 and 0 µM for DR1 and DR4 or 100, 33, 11, 3.7, 1.2, 0.4, 0.1 and 0 nM for DR7. After incubation at 37 °C overnight, 100 µl was transferred to a plate coated with avidin (2 µg ml^{-1}) and previously blocked. After 1 h of incubation at room temperature and three washes with 1× PBS, pH 7.4 and 0.05% Tween 20, anti-FLAG-alkaline phosphatase conjugate (Sigma) was added as 1:5,000. After 1 h, the plate was washed as previously described and developed with pNPP SigmaFAST substrate and absorbance was read with a 405-nm filter.

MixMHC2pred, a predictor of HLA-II ligands. We trained a predictor of HLA-II ligands (MixMHC2pred) using all our HLA-II peptidomics data (including the pan-HLA-II, HLA-DR and HLA-DR-depleted peptidomics data). For a given allele, a, and peptide, n, the binding score is given by:

$$B^a_n = \left(\sum_{s=-S}^{S} w_s \prod_{l=1}^{L} \frac{\bar{\theta}^a_{l,x^n_{l+s}}}{f_{x^n_{l+s}}}\right) \cdot \left(\sum_{k=0}^{3} w^N_k \prod_{l=1}^{3} \frac{\bar{\nu}^k_{l,x^n_l}}{f_{x^n_l}}\right) \cdot \left(\sum_{k=0}^{3} w^C_k \prod_{l=\lambda^n-2}^{\lambda^n} \frac{\bar{\gamma}^k_{l,x^n_l}}{f_{x^n_l}}\right) \quad (3)$$

where w_s represents the global binding core offset preference (computed by combining all peptides associated to an allele); $\bar{\theta}^a_{l,i}$ is the position probability matrix for allele a (computed from all peptides associated to this allele with their respective binding core offset on the basis of the highest responsibility value, and adding pseudocounts on the basis of the BLOSUM62 substitution matrix with a parameter $\beta = 200$ (ref. [32])); $\bar{\nu}^k_{l,i}$ and $\bar{\gamma}^k_{l,i}$ are similar matrices representing the N- and C-terminal motifs (Supplementary Fig. 5a,b; including here only the amino acids within the peptides); w^N_k and w^C_k represent the relative contributions of the N- and C-terminal motifs (that is, the fraction of peptides assigned to each of these motifs). See equation (1) for the definition of other terms.

This binding score is then transformed to a percentile rank per peptide length by comparing it to the score of 10,000 random human peptides of the same length, and then further transformed to a global percentile rank by ensuring that the top 1% of random human peptides follow the same peptide length distribution as the global peptide length distribution observed in our HLA-II peptidomics data. Finally, when the score among multiple alleles is requested, the score from each peptide is taken as its best percentile rank among all the alleles.

Benchmarking HLA-II ligand predictions. The accuracy of MixMHC2pred was tested in 41 samples from seven independent HLA-II peptidomics datasets[2,8,33–37] (Supplementary Table 2). The positives were the peptides of lengths between 12 and 19 amino acids observed in these samples (removing all the peptides that were also part of the training data from MixMHC2pred for any of the alleles from a given sample). For each sample we then added four times more negatives by randomly sampling human peptides of lengths between 12 and 19 amino acids.

Predictions from MixMHC2pred and its different variants (see Fig. 2a) were compared with those from NetMHCIIpan (v.3.2 (ref. [20]) with default parameters) on the basis of the HLA-II typing provided in these studies (Supplementary Table 2). The area under the curve (AUC) of the receiver operating characteristic (ROC) curve was computed for each sample separately (Fig. 2a and Supplementary Fig. 7). The positive predictive value for the top x% of predictions ($\mathrm{PPV}_{x\%}$) was also computed for each sample (Supplementary Fig. 7; computed for $x = 2\%$ and $x = 20\%$). For this, a threshold was determined for the predictor score to have x% of peptides considered to be ligands. PPV at this threshold was then obtained as $\mathrm{PPV}_{x\%} = \frac{\mathrm{True\ positives}_{x\%}}{\mathrm{True\ positives}_{x\%} + \mathrm{False\ positives}_{x\%}}$.

In addition to the HLA-II peptidomics data, we tested the predictors on binding affinity data. This included all binding affinity data obtained from the IEDB database[19] (as of 5 May 2019). These data was filtered to remove peptides with non-standard amino acids or that were tested against an allele absent from MixMHC2pred. Only data from 2017 onward was included, because NetMHCIIpan had been trained on all the binding affinity data up to 2016. This analysis showed that the motifs obtained from HLA-II peptidomics data also helped in predicting binding affinities (Supplementary Fig. 16a,b). Other features than HLA-II motifs did not help here (Supplementary Fig. 16), which is expected as these characteristics would be more related to processing and cleavage biases than binding affinities.

Benchmarking predictions of epitopes from tetramer assays. All the multimer and tetramer assay data for human CD4$^+$ T cells from the IEDB database[19] were downloaded (as of 20 July 2018). We then filtered these data to remove peptides with non-standard amino acids and to keep peptides of length 12 and longer that were associated to a known allele based on the following 'Allele evidence codes': 'MHC binding assay' or 'T cell assay -Single MHC type present'. Only interactions involving alleles available in MixMHC2pred were considered.

Predictions with MixMHC2pred and NetMHCIIpan v.3.2 (ref. [20]) were performed for each peptide with its associated HLA-II allele, both for the positive (1,319 peptides) and the negative (1,040 peptides) cases. The corresponding ROC curve and its AUC are shown in Fig. 2b. The epitope relative rank is defined as the fraction of negative peptides in the dataset that had a better predicted score than a given true epitope, and was computed separately for each true epitope (Supplementary Fig. 8a).

Selection of candidate viral, bacterial and tumor-associated epitopes. To further benchmark MixMHC2pred, we retrieved a list of known viral, bacterial and melanoma-associated proteins (Supplementary Table 4a) and tested their immunogenicity in two HLA-DRB1*07:01-positive patients with melanoma and one HLA-DRB1*07:01-positive healthy donor. Each protein was cut into 20-amino-acid oligomers overlapping by 10 amino acids to cover all possible 9-amino-acid cores. These 20-amino-acid peptides were then ranked according to the predicted affinity to HLA-DRB1*07:01 (considering the highest predicted affinity from the 15-amino-acid subsequences in each peptide). We then selected the 30 best scoring potential epitopes from the viral and bacterial proteins and the 30 best scoring potential epitopes from the tumor-associated antigens for experimental validation, both for the predictions from MixMHC2pred and NetMHCIIpan.

Matthews correlation coefficients were computed on the basis of the epitopes tested experimentally (for example, the true negatives from MixMHC2pred are the peptides that had been predicted in the top 60 by NetMHCIIpan but not by MixMHC2pred and that are not immunogenic).

Peptide synthesis. Peptides were synthesized at the Protein and Peptide Chemistry Facility at the University of Lausanne by standard solid phase chemistry on a multiple peptide synthesizer (Applied Biosystem). All peptides were >90% pure as indicated by analytic HPLC. Lyophilized peptides were diluted in pure DMSO at 10 mg ml^{-1} and aliquots at 1 mg ml^{-1} in 10% DMSO were prepared and stored at −80 °C.

In vitro peptide stimulation. Peripheral blood mononuclear cells (PBMCs) from two HLA-DRB1*07:01-positive patients with malignant melanoma and from one HLA-DRB1*07:01-positive healthy donor were thawed and CD4$^+$ T cells were enriched using anti-CD4 microbeads and MiniMACS magnetic separation columns (Miltenyi Biotec). CD4$^+$ T cells were resuspended in RPMI 1640 (Gibco) supplemented with 2 mM glutamine, 1% (vol/vol) non-essential amino acids, 50 µM 2-β-mercaptoethanol, penicillin (50 U ml^{-1}) and streptomycin (50 µg ml^{-1}) (Gibco), and 8% human serum (Blood Transfusion Center, Bern) (complete medium) and seeded (0.5×10^6/well) in 48-well plates to which autologous irradiated (30grey) CD4$^-$ T cells were added at a 1:1 ratio. Pools of the selected viral, bacterial or tumor-associated peptides (20-amino-acid oligomers; Supplementary Table 4b) were added to the wells at a final concentration of 2 µM each. After an overnight period in culture, 500 µl of medium was replaced by fresh medium containing 100 IU ml^{-1} final of human recombinant IL-2. Every 2 d the medium was refreshed. After 10 d of in vitro expansion, cultures were tested for the presence of antigen-reactive CD4$^+$ T cells. Aliquots of 10^5 cells were transferred to individual wells of a 96-well plate and stimulated overnight with a mix of multiple peptides distributed in different pools according to a specific matrix (Supplementary Fig. 9b). Brefeldin A at 2.5 µg ml^{-1} (Sigma-Aldrich) was added to each well. A non-stimulated control was added as well as a positive control where cells were stimulated with phorbol 12-myristate 13-acetate (PMA) (Sigma-Aldrich) and ionomycine (Sigma-Aldrich) at 50 ng ml^{-1} and 500 ng ml^{-1}, respectively. The following day, cells were collected and stained using anti-CD3-APC (clone UCHT1, Beckman Coulter), anti-CD4-FITC antibodies (clone RPA-T4, Biolegend) and live–dead feasible Aqua dead-cell stain (Invitrogen) for 20 min at 4 °C. Cells were then washed with PBS, fixed and permeabilized using the FOXP3/transcription kit (Invitrogen) for 30 min at room temperature. Finally, the cells were stained for intracellular markers using anti-interferon-γ (IFNγ)-PE (clone 4SB3, BD Biosciences) and anti-tumor necrosis factor-α (TNFα)-AF700 antibodies (clone Mab11, BD Biosciences) for 20 min at 4 °C. Cells were acquired with the CYTOFLEX analyzer (Beckman Coulter) and data were analyzed with Flowjo software. Positive wells for IFNγ and/or TNFα were identified (Supplementary Fig. 9c,d). As each individual peptide was only contained in two different pools, by matching the positive wells in the matrix, individual immunogenic peptide were selected and evaluated individually. In a similar procedure that was used to evaluate multiple peptides in a matrix format, individual selected peptides were added to newly seeded CD4$^+$ T cells and after an overnight incubation the cells were evaluated for their expression of IFNγ and TNFα using the same method as described above (Supplementary Fig. 9e–f).

Patient and neoantigen description. Patient CTE-0007 is a patient with recurrent ovarian cancer. Clinical data and all methodologies for the identification of non-synonymous somatic mutations were already described[38].

Identification and validation of neoepitope-specific CD4$^+$ TILs. Neoantigen (mutation D246N in SGOL1)-specific CD4$^+$ TILs were identified in patient CTE-0007 upon co-culture with tandem-minigene-transfected autologous B cells. TILs were derived from tumor single-cell suspensions and expanded with high-dose IL-2 (Proleukin, 6,000 IU ml^{-1}) for 15 d as previously reported[38]. In parallel, CD19$^+$ cells were isolated from PBMCs using magnetic beads (Miltenyi) and expanded for 14 d with multimeric CD40L (Adipogen; 1 µg ml^{-1}) and IL-4 (Miltenyi; 200 IU ml^{-1}). CD40-activated B cells were electroporated using a Neon system (Invitrogen) with 1 µg of in vitro transcribed RNA (Ambion) coding for 31-amino-acid oligomers centered on the specific mutations. Following 16 h of resting after electroporation, 10^5 cells per well of RNA-transfected B cells were co-cultured with TILs at a ratio of 1:1 and incubated overnight in precoated enzyme-linked immune absorbent spot (ELISpot) plates (Mabtech). Subsequently, T cell activation was validated by intracellular cytokine staining as described[38]. Either RNA-transfected B cells or B cells loaded with peptides were used as APCs. Neoepitope-reactive CD4$^+$ TILs were sorted using a FACSAria IIu, on the basis of CD154 upregulation, as described[39]. CD154-sorted cells were expanded with irradiated feeder cells (PBMCs from two donors) in the presence of OKT3 (Miltenyi; 30 ng ml^{-1}) and IL-2 and further interrogated to identify the predicted candidate epitopes by IFNγ ELISpot. Additionally, HLA-DR-blocking (clone L243, in-house production) antibody was added together with cognate peptides. For HLA-restriction analysis, HLA-matched or mismatched CD40-activated B cells were loaded with 2 µM peptide SPIFKQKKNLRRS for 2 h before co-culture.

The candidate epitopes have been selected on the basis of the top five predictions from MixMHC2pred and NetMHCIIpan among all 13- to 16-amino-acid oligomers of the minigene (TDLCFLNSPIFKQKKNLRRSKKRALEVSPAK).

Benchmarking neoepitope predictions. The list of neoantigens that were tested experimentally for CD4$^+$ T cell immunogenicity was retrieved from the literature[40–46]. All patients for which HLA typing was publicly available and that had allele(s) defined in MixMHC2pred were included in the benchmark (Fig. 2e–f and Supplementary Data 3). Neoantigens tested experimentally were usually sequences of 20–25 amino acids. We gave a score to these sequences on the basis of the highest score from their 15-amino-acid oligomer subsequences with either MixMHC2pred or NetMHCIIpan. We then computed the ROC curve and corresponding AUC considering only the peptides tested experimentally for each patient (that is, no artificial negatives). Figure 2e shows the results obtained by grouping the data from all patients and Supplementary Fig. 12 shows the data for each patient separately. Additionally, for each neoepitope, we determined its epitope relative rank (that is, the fraction of negative antigens tested experimentally for a given patient that had a better predicted score than a given true epitope; Fig. 2f).

Statistics. The relevant statistical test, sample size, replicate type and *P* values for each figure and table are found in the figure or table and/or the corresponding figure or table legends. Statistical analyses were performed with R.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw datasets generated during the current study are available in the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD012308 and the corresponding peptide output files are provided as Supplementary Data 1 and 2. Additional datasets generated during the current study are available as Supplementary Tables 1 and 4. In addition, public datasets were analyzed in this study, obtained from the IEDB database[19] and from the studies listed in Supplementary Table 2, as well as from multiple neoantigen studies listed in Supplementary Data 3.

Code availability

MoDec and MixMHC2pred are freely available as C++ executables (https://github.com/GfellerLab/ and Supplementary Code 1 and 2) for academic non-commercial

research purposes. MixMHC2pred is also freely available for academic non-commercial research purposes as a web application (http://mixmhc2pred.gfellerlab.org/).

References

26. Dudley, M. E. et al. *Clin. Cancer Res.* **16**, 6122–6131 (2010).
27. Donia, M., Larsen, S. M., Met, Ö. & Svane, I. M. *Cytotherapy* **16**, 1117–1120 (2014).
28. Vizcaíno, J. A. et al. *Nucleic Acids Res.* **44**, D447–D456 (2016).
29. Cox, J. & Mann, M. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
30. Gfeller, D. et al. *Mol. Syst. Biol.* **7**, 484 (2011).
31. Wagih, O. *Bioinformatics* **33**, 3645–3647 (2017).
32. Nielsen, M. et al. *Bioinformatics* **20**, 1388–1397 (2004).
33. Clement, C. C. et al. *J. Biol. Chem.* **291**, 5576–5595 (2016).
34. Collado, J. A. et al. *Eur. J. Immunol.* **43**, 2273–2282 (2013).
35. Ooi, J. D. et al. *Nature* **545**, 243–247 (2017).
36. Wang, Q. et al. *J. Proteome Res.* **16**, 122–136 (2017).
37. Bergseng, E. et al. *Immunogenetics* **67**, 73–84 (2015).
38. Bobisse, S. et al. *Nat. Commun.* **9**, 1092 (2018).
39. Chattopadhyay, P. K., Yu, J. & Roederer, M. *Nat. Protoc.* **1**, 1–6 (2006).
40. Ott, P. A. et al. *Nature* **547**, 217–221 (2017).
41. Tran, E. et al. *Science* **350**, 1387–1390 (2015).
42. Veatch, J. R. et al. *J. Clin. Invest.* **128**, 1563–1568 (2018).
43. Veatch, J. R et al. *Cancer Immunol. Res.* **7**, 910–922 (2019).
44. Yossef, R et al. *JCI Insight* **3**, e122467 (2018).
45. Zacharakis, N. et al. *Nat. Med.* **24**, 724–730 (2018).
46. Sahin, U. et al. *Nature* **547**, 222–226 (2017).

nature research

Corresponding author(s): David Gfeller and Michal Bassani-Sternberg

Last updated by author(s): Sep 2, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	HLA typing information was obtained with the Assign TruSight HLA v2.1 software. MaxQuant platform version 1.5.5.1 was used to identify the peptides in HLA-II peptidomics.
Data analysis	Data was analyzed with our custom codes MoDec v1.1 and MixMHC2pred v1.1. Comparisons were also performed with Gibbscluster version 2.0, MEME version 4.12, NetMHCIIpan version 3.2 and NNalign version 2.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw datasets generated during the current study are available in the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD012308 (https://www.ebi.ac.uk/pride/archive/projects/PXD012308) and the corresponding peptide output files are provided as Supplementary Data 1 and 2. Additional datasets generated during the current study are available as Supplementary Tables 1 and 4. In addition, public datasets were analyzed in this study, obtained from the IEDB database, from the studies listed in Supplementary Table 2, as well as from multiple neoantigen studies listed in Supplementary Data 3.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed; the goal of this study was to have as a high allele coverage as possible, performing thus experiments for as many samples as possible.
Data exclusions	We did not exclude data from the datasets generated for this study. In the public datasets we excluded a sample when the corresponding allele was not part of the alleles defined for our predictor.
Replication	The motif deconvolution was performed on multiple different samples sharing some alleles, showing the replicability of the motifs obtained for the alleles. For the immunogenicity of viral, bacterial and tumor-associated epitopes, we repeated the experiment for two patients and a healthy donor sharing a same allele, showing similar results in all cases. For the validation of a neoepitope, no replication of the experiment could be performed in other patients as a neoepitope is patient-specific by definition and another patient would thus not have the same neo-epitope.
Randomization	This is irrelevant for our study because we did not compare multiple treatment options but analyzed all data with all tools.
Blinding	For the T cell immunogenicity assays, the experimenters were blinded as to which predictor (MixMHC2pred or NetMHCIIpan) had predicted which peptide.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Anti-pan-HLA-II antibody : from hybridoma “HB-145”. Company name : ATCC. Catalog number : HB-145. Clone name: iva12. Antigenic determinant: HLA DR, DP, DQ. Isotype: IgG1; kappa light chain. Host: mouse. Cell type: Hybridoma: B lymphocyte. Clonality: monoclonal. Anti-HLA-DR antibody: from hybridoma “HB-298”. Company name : ATCC. Catalog number : HB-298. Clone name: LB3.1. Antigenic determinant: HLA DR alpha chain. Isotype: IgG2a. Host: mouse. Cell type: Hybridoma: B lymphocyte. Clonality: monoclonal. CD3-APC, FC, Beckman Coulter, ref: IM2467, clone UCHT1, lot60. FITC anti-human CD4 Antibody, FC, Biolegend, ref: 300538, clone RPA-T4, lotB256063. PE Mouse Anti-Human IFN-γ, FC, BD, ref: 554552, clone 4SB3, lot3144873. Alexa Fluor® 700 Mouse Anti-Human TNF, FC, BD, ref: 557996, clone Mab11, lot7125816. LIVE/DEAD™ Fixable Aqua Dead Cell Stain Kit, FC, Invitrogen, for 405 nm excitation, ref: L34966, Lot2031176.
Validation	We produce the anti-pan-HLA-II and anti-HLA-DR antibodies ourselves through the hybdridoma cells as described in the manuscript. CD3-APC: Target Species: Human, Application: flow cytometry. FITC anti-human CD4 Antibody: Reactivity: Human, Chimpanzee, Cynomolgus, Application: FC - Quality tested, each lot of this antibody is quality control tested by immunofluorescent staining with flow cytometric analysis. For flow cytometric staining, the

suggested use of this reagent is 5 µl per million cells in 100 µl staining volume or 5 µl per 100 µl of whole blood.
PE Mouse Anti-Human IFN-γ: Reactivity: Human (QC Testing) Rhesus, Cynomolgus, Baboon (Tested in Development), Application: Intracellular staining (flow cytometry) (Routinely Tested).
Alexa Fluor® 700 Mouse Anti-Human TNF: Reactivity: Human (QC Testing) Rhesus, Cynomolgus, Baboon (Tested in Development), Application: Intracellular staining (flow cytometry) (Routinely Tested).
LIVE/DEAD™ 405 nm excitation: Compatible Cells: Eukaryotic Cells, For Use With (Equipment): Flow Cytometer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	JY: ATCC, 77442. CD165, PD42, CM467, RA957, BP455, GD149: a gift from Pedro Romero, Ludwig Cancer Research Lausanne.
Authentication	The cell lines were not authenticated but they were HLA-II typed which is the relevant information needed for this study.
Mycoplasma contamination	The cell lines were negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Human research participants

Policy information about studies involving human research participants

Population characteristics	All samples were typed for HLA-II which is the only characteristcs relevant for this study.
Recruitment	Existing samples from patients at CHUV were used after informed consent of the patients.
Ethics oversight	Ethics Comission, CHUV.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Experiment with the candidate viral, bacterial and tumor-associated epitopes: Peripheral blood mononuclear cells (PBMCs) from two HLA-DRB1*07:01 positive malignant melanoma patients and one healthy HLA-DRB1*07:01 positive donor were thawed and CD4+ T cells enriched using anti-CD4 microbeads and MiniMACS magnetic separation columns (Miltenyi Biotec, Bergisch Gladbach, Germany). CD4+ T cells were resuspended in RPMI 1640 (Gibco, Dublin, Ireland) supplemented with 2 mM glutamine, 1% (vol/vol) nonessential amino acids, 50 µM 2β-mercaepthanol, penicillin (50 U/ml) and streptomycin (50 µg/ml) (Gibco, Dublin, Ireland), and 8% human serum (Blood transfusion center, Bern) (complete medium) and seeded (0.5x106/well) in 48 well plates to which autologous irradiated (30grey) CD4− T cells were added at a 1:1 ratio. Pools of the selected viral/bacterial or tumor associated peptides were added to the wells at a final concentration of 2 µM each. Experiment with a neoepitope: TILs were plated in the presence or absence of the 31-mer peptide containing the identified neoepitope and brefeldinA (BD biosciences, USA). After 16–18 h, cells were harvested and stained with anti-CD3, anti-CD8, anti-CD4, anti-IL-2, anti-TNFα, anti-IFNγ and with viability dye (Life technologies).
Instrument	Experiment with the candidate viral, bacterial and tumor-associated epitopes: CYTOFLEX analyzer (Beckman Coulter). Experiment with a neoepitope: acquired on a four-lasers Fortessa (BD biosciences).
Software	Flowjo, LLC software (Oregon, USA). Version 10.5.3.
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	Experiment with the candidate viral, bacterial and tumor-associated epitopes:

Gating strategy

Gating was done on the lymphocyte population identified by forward (FSC-A) and side scatter (SSC-A), single events are selected by evaluating the FSC-A against the FSC-H. CD3 positive and LIVE/DEAD negative cells are selected followed by CD3 positive and CD4 positive cell selection. The cells are displayed in the figures with the markers IFN-γ-PE and TNF-α-AF700.

Experiment with a neoepitope:
Cells of interest were gated on lymphocytes identified by FSC-A/SSC-A, doublets were excluded using FSC-A against FSC-H. Live cells (LIVE/DEAD negative) and finally CD3 positive cells were selected. CD4 positive and CD8 positive cells against the three cytokines (IFNγ, TNFα and IL-2) are displayed in the supplementary figure 10.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.