nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-022-01464-2

Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors

Assaf Kacen[1,8], Aaron Javitt[1,8], Matthias P. Kramer[1], David Morgenstern[2], Tomer Tsaban[3], Merav D. Shmueli[1], Guo Ci Teo[4], Felipe da Veiga Leprevost[4], Eilon Barnea[5], Fengchao Yu[4], Arie Admon[5], Lea Eisenbach[1], Yardena Samuels[6], Ora Schueler-Furman[3], Yishai Levin[2], Alexey I. Nesvizhskii[4,7] and Yifat Merbl[1]✉

Post-translational modification (PTM) of antigens provides an additional source of specificities targeted by immune responses to tumors or pathogens, but identifying antigen PTMs and assessing their role in shaping the immunopeptidome is challenging. Here we describe the Protein Modification Integrated Search Engine (PROMISE), an antigen discovery pipeline that enables the analysis of 29 different PTM combinations from multiple clinical cohorts and cell lines. We expanded the antigen landscape, uncovering human leukocyte antigen class I binding motifs defined by specific PTMs with haplotype-specific binding preferences and revealing disease-specific modified targets, including thousands of new cancer-specific antigens that can be shared between patients and across cancer types. Furthermore, we uncovered a subset of modified peptides that are specific to cancer tissue and driven by post-translational changes that occurred in the tumor proteome. Our findings highlight principles of PTM-driven antigenicity, which may have broad implications for T cell-mediated therapies in cancer and beyond.

Targeting tumor antigens that are bound to major histocompatibility complex (MHC) molecules holds great promise for T cell therapies and immunotherapies. Peptides derived from foreign pathogens and self-proteins that have undergone disease-related changes, such as mutations[1–5], may elicit an immune response. Similarly, post-translational modifications (PTMs) such as phosphorylation, citrullination or glycosylation[6–12] may also occur on presented antigens, which have been reported to modulate antigen presentation and recognition[13]. For example, such changes in the antigenic landscape were reported in clinical phospho-proteomic analysis of breast and lung cancer, uncovering differential activation of cellular pathways[14,15]. However, with more than 200 different types of PTMs and the technical difficulties in detecting them, whether and to what extent such PTM-driven alterations expand our landscape of antigenic targets in cancer remained under-explored.

Current approaches for neoantigen discovery rely mostly on genomic or transcriptomic data[16], combined with computational prediction tools for human leukocyte antigen class I (HLA I) binding[17–21]. Such approaches are geared toward identifying neo-antigens generated by mutations or noncanonical amino acid sequences[22]. Because they are focused on the pretranslational level, they lack direct information on the state of modification of the peptides. Another approach relies on the identification of HLA I-bound peptides by immunoprecipitation of the MHC/HLA-peptide complex from the surface of cells and eluting the bound peptides before mass spectrometry (MS)-based analysis (that is, immunopeptidomics). MS analysis and the identification of peptides are done by comparing the peptides detected by the instrument to a reference dataset containing all possible theoretical peptides across the proteome[23]. Thus, to detect PTMs on such peptides, it is necessary to have the relevant reference sequence that contains the same mass shift imparted by the modification. As each additional modification increases the number of theoretical peptide possibilities in the search space exponentially, the search time becomes a limiting factor. Many approaches, to cope with the exponential growth of the search space when searching for PTMs, have recently been implemented (open search[24], de novo[25]) in various MS analysis tools (MetaMorpheus[26], PEAKS PTM)[25–28]. To date, however, the vast majority of PTMs, and combinations thereof, have not been examined in the immunopeptidome.

To address these challenges and examine the potential landscape of modified peptides bound to HLA I molecules in a systematic and unbiased manner, we developed a Protein Modification Integrated Search Engine (PROMISE). Our computational pipeline allows for combinatorial detection of multiple PTMs without prior biochemical enrichment. By examining data generated from 210 samples, including patient-derived tumor samples and cancer cell lines, we found thousands of new modified HLA I-bound peptides. Notably, some of these modified peptides reside within known cancer-associated antigens or cancer driver genes, offering a new class of antigens, that may be further examined in the context of

[1]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [2]De Botton Institute for Protein Profiling, Nancy and Stephen Grand Israel National Center for Personalized Medicine, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Microbiology and Molecular Genetics, Institute for Medical Research Israel-Canada, Faculty of Medicine, The Hebrew University, Jerusalem, Israel. [4]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [5]Faculty of Biology, Technion—Israel Institute of Technology, Haifa, Israel. [6]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [7]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [8]These authors contributed equally: Assaf Kacen, Aaron Javitt. ✉e-mail: yifat.merbl@weizmann.ac.il

immunotherapy. By systematically analyzing the locations of PTMs on MHC-eluted peptides, we uncovered PTM-driven motifs across many haplotypes, in many cases altering the anchoring positions or the middle region of the peptide, which is associated with the T cell recognition region. We further confirmed these observations by using structural three-dimensional (3D) modeling. We compared the immunopeptidome of mouse colorectal cancer to the healthy tissue immunopeptidome, finding peptides that were unique to the tumors and passed spectrum validation. Finally, by extending our analysis to a breast cancer cohort from the Clinical Proteomic Tumor Analysis Consortium (CPTAC[14]) database, we revealed cancer- and site-specific modifications. Such sites were identical to the ones we found in modified antigens, bringing insight into metabolic changes and the altered modification landscape induced by the cancerous state. Collectively, our systematic identification of modified peptides and their impact on HLA I binding and recognition should broaden our understanding of the effects PTMs may have on defining tumor–host interactions.

Results

Establishment of PROMISE. The assignment of peptides that were detected by the MS instrument to their cognate amino acid sequence is based on the reference proteome that is provided to the analysis software. As such, peptides that are eluted from HLA I molecules may only be identified if they match a specific 'theoretical' peptide in a defined search space (that is, reference proteome). Peptides that are detected by MS but cannot be matched or assigned to any sequence are considered as the 'dark matter of the proteome' (ref. [29]). The dark matter of the proteome may include all sequences that deviate from the encoded amino acid sequence of proteins, such as mutations, noncanonical translation products, fusion proteins and spliced peptides or PTMs[22,30–34]. For the last of these, identifying modified peptides that may be presented on HLA I molecules, in a systematic manner, remains a challenge due to the huge search space of endogenous peptides with the numerous possibilities of protein modifications. In recent years, several approaches were implemented to cope with this challenge (MetaMorpheus[26], PEAKS PTM), offering state-of-the-art solutions to peptide assignment. Here we developed PROMISE, which relies on the ultrafast MSFragger[35] software for comparison between theoretical peptides and the peptide captured in the instrument (for details, see Supplementary Information). PROMISE simultaneously searches HLA immunopeptidomics data against multiple modification types that are not identified by standard analysis (Fig. 1a). Modifications identified by PROMISE can indicate either PTMs that represent the altered protein state in the cell or modifications that may have been introduced during sample processing (for example, carbamidomethylation[36] and deamidation[37,38]). Nevertheless, incorporation of diverse types of modifications to the search space allows us to choose the best match for the detected peptides and assign peptides that would otherwise not be identified. Only peptides that match better to a theoretical peptide with a modification than all other possible matches to the encoded amino acid sequences in the proteome (Methods) are defined as modified peptides by PROMISE. To ascertain that the distributed architecture of PROMISE does not alter the peptide spectrum assignments of MSFragger, we chose a small set of modifications and analyzed them by both PROMISE and stand-alone MSFragger. The spectral assignments from the same subset of data were 99.2% identical (Fig. 1b).

To use the full potential of PROMISE, we defined a broad range of PTMs comprising 29 modification combinations of 12 modification types (36 mass shifts; Table 1) on 16 different amino acids and protein termini (termed hereafter 'multi-modification search'). These include modifications such as methylation, acetylation, phosphorylation, citrullination, ubiquitination, SUMoylation, oxidation, deamidation, cysteinylation and carbamidomethylation. We then ran this multi-modification search to analyze previously published high-resolution HLA I immunopeptidomics data (PRIDE identifications: PXD004894 (ref. [7]), PXD000394 (ref. [39]), PXD006939 (ref. [40]), PXD003790 (ref. [41]) and PXD009738 (ref. [42]) of patient tumor tissues ($n=35$) or healthy adjacent tissue ($n=5$), cancer cell lines ($n=13$) and tumor-infiltrating lymphocytes ($n=2$). However, because a multi-modification search increases the search space (Methods and Supplementary Information), we needed to ensure that we are not increasing identifications merely by altering the false positive rate. Therefore, we used a subgroup false discovery rate (FDR) at 5% by splitting spectra into different groups based on the modification state, thereby using a stricter FDR cutoff on the modified peptides than on the unmodified ones (Fig. 1c). Indeed, the probability of true peptide spectrum matches (PSMs), as calculated by PeptideProphet[43], was higher in the group of modified peptides compared to their unmodified counterparts (Fig. 1d).

We next set out to characterize the enrichment in peptide identification with PROMISE compared to the original search criteria used in the previous studies, which in most of our datasets included methionine oxidation and protein N-terminus acetylation (termed hereafter 'standard search'). The multi-modification search identified 32,798 modified peptides (Supplementary Data 1), of which 12,228 were unique to the multi-modification search, thereby enriching the total pool of immunopeptides identified by 3.7% (Fig. 1e and Extended Data Fig. 1). While the amino acid compositions of the immunopeptidome were similar between the standard search and PROMISE, we saw an enrichment in amino acids that can carry modifications when comparing the modified and unmodified peptide subsets (Fig. 1f,g). For example, as previously described[44], cysteines are consistently under-represented in immunopeptidomics analyses, yet these constituted 2% of the modified immunopeptidome (Fig. 1g). As expected, most of the modified

Fig. 1 | Computational pipeline for global search of PTMs on HLA I-bound peptides enriches identifications. a, PROMISE allows for the systematic detection of modifications on HLA I-bound peptides. **b**, Scatter plot of the intensity of PSMs as identified in stand-alone MSFragger or PROMISE distributed search. Spearman correlation, $\rho=0.98$. **c**, Global FDR relies on the distribution of decoy PSMs compared to the correctly assigned (true) hits (light gray; left distribution compared to the right distribution in black). Subgroup FDR divides all the PSMs and counterpart decoys into unmodified (middle; shades of blue) and modified (bottom; shades of red) peptides. The FDR cutoff is calculated for each group separately, resulting in a more strict cutoff for modified PSMs. **d**, PeptideProphet probability score binned for unmodified (gray) and modified (dark red) PSMs. The subgroup FDR cutoff results in a higher probability of true peptide matches for modified than unmodified PSMs. **e**, PROMISE identification enrichment. Top, 10% of the immunopeptidome contains modifications and 3.7% are unique peptides with modifications that were identified through PROMISE. Bottom, pie chart of the 10% of modified peptides identified in the standard and multi-modification searches performed. Of 32,798 modified peptides identified in the analysis, 37.29% were unique to PROMISE (dark red). **f**,**g**, The amino acid composition of the peptides identified compared for the standard and PROMISE search (**f**) or the unmodified and modified subsets of peptides in the PROMISE search (**g**). Circle size and color indicate the $\log_2$ transformed ratio of amino acid abundance between the two subsets. **h**, Peptides identified in PROMISE binned by number of modifications. **i**, When viewed by modification site, 19,630 positions were uniquely identified by PROMISE in the immunopeptidomics datasets analyzed. These sites are presented in a pie chart divided by modification type and amino acid modified. **j**, Peptide length distribution (density as a percentage of total peptides) per modification type.

peptides carried only one modification (Fig. 1h). In total, we identified 19,630 modification sites (from 12,228 peptides) that were unique to PROMISE, 88% of which contained modification types that are not included in a standard search (Fig. 1i). We next analyzed the length distribution per modification type and observed that acetylation, citrullination, dimethylation, SUMOylation and ubiquitination were longer on average than in the unmodified subset (Fig. 1j).

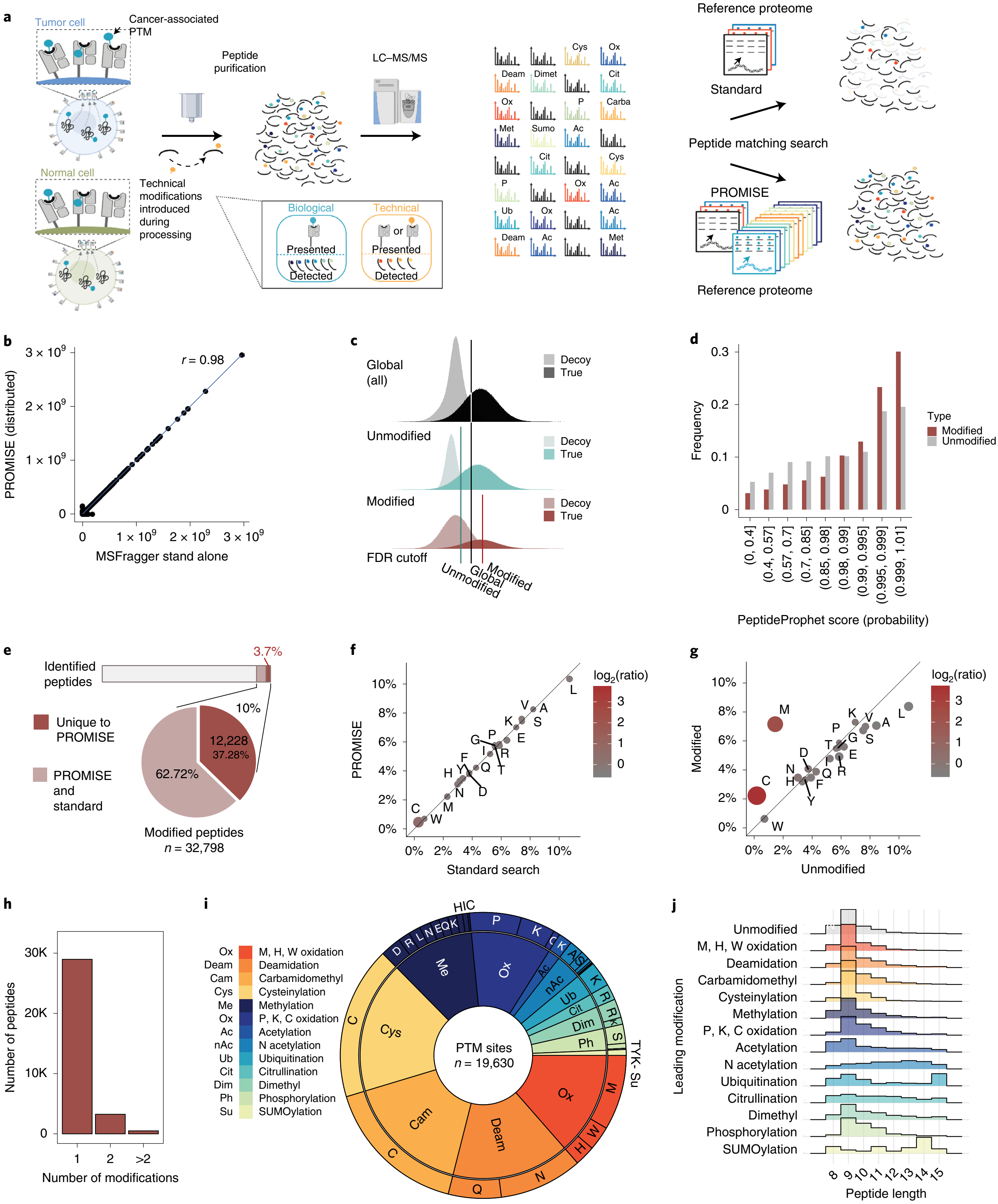


Table 1 | Description of the different combinations of PTM types and amino acids comprising the set searched against in our analysis

Modification name	Amino acid	Mass shift	UNIMOD accession no.	UNIMOD classification	Remark
Methionine oxidation	M	15.99490	35	Artifact	Technical—common chemical nonenzymatic modification; appears in most MS searches[76]
Protein N-terminus acetylation	[X]@N-terminus	42.01060	1	Multiple	Biological
Phosphorylation	YTS	79.96633	21	PTM	Biological
Acetylation	K	42.01060	1	Multiple	[77] Biological
Methylation	CHNQKRILDE	14.01565	34	PTM	Biological
Dimethylation	KR	28.0312	36	PTM	Biological
Oxidation	WHKPC	15.99490	35	KPC—PTM WH—artifact	KPC—biological, WH—technical[78]
Deamidation	NQ	0.98402	7	Artifact	Technical
Citrullination	R	0.98402	7	PTM	Biological—enzymatic modification
Ubiquitination	K	57.0215 (G)	1263	Other	Biological
		114.0429 (GG)	121	Other	
		270.144 (GGR)	-	-	
		383.228103 (GGRL)	535	Chemical derivative	
Sumoylation	K	215.0906 (GGT)	-	-	Biological—G and GG—cannot distinguish between ubiquitin, Sumo or FAT10
		343.149184 (GGTQ)	1293	Other	
FAT10	K	227.127 (GGI)	-	-	Biological
		330.136176 (GGIC)	1990	PTM	
Cysteinylation	C	119.004099	312	Multiple	Technical
Carbamidomethyl	C	57.021464	4	Chemical derivative	Technical—used as fix modification in trypsin digestion

Global identification of PTM motifs in HLA I peptides. We next examined the enriched peptide repertoire to understand the impact that PTMs have in shaping the cancer immunopeptidome. First, we asked whether a given PTM has the tendency to be in certain positions within the peptide. A broad view across different types of modifications (Supplementary Data 2) revealed that some modifications have a distinct site preference (Fig. 2a). For example, as previously shown[6,7], serine phosphorylation predominantly falls in the fourth position of the peptide, while methylation is distributed evenly across the peptide (Fig. 2a, light green). Further, we found that oxidation and cysteinylation are enriched at the end of the peptide (toward the C terminus) and carbamidomethyl is enriched in the third position, while cysteinylation is under-represented at the second position.

Next, we explored whether the distribution of these PTMs is distinct from the underlying distributions of the amino acid residues that they modify. In addition, we also examined an unbiased and broader background distribution by collectively defining all of the reported epitopes in the Immune Epitope Database (IEDB)[45]. As expected, when examining a modification that is widely generated by sample processing or handling, such as methionine oxidation, the correlation between the oxidized methionine position distribution and the unmodified methionine distribution was very high (Pearson correlation, 0.96; $P = 1.05 \times 10^{-6}$) (Fig. 2b). This suggests that the modification occurred randomly across the peptide during sample preparation (two-sided *F*-test, $P = 0.3543$). As this was not the case for all the PTMs, we ordered all of the PTMs based on the correlation of their distribution to the background (Fig. 2c). This metric highlights PTM motifs that may alter HLA binding preference or T cell receptor (TCR) recognition. Peptide binding to HLA I molecules depends on the biochemical properties of both the peptide and the HLA I structure. The most critical residues for HLA I binding are the ones that fit into the anchor pockets in the HLA I groove, typically the second and C-terminal positions[46]. By contrast, the TCR recognition motif is determined by the HLA I peptide complex and, therefore, most strongly influenced by the residues in positions 3–7 of the HLA I-bound peptide[47,48]. In the presented matrix, for example, known motifs, such as the tendency of serine phosphorylation modification at position 4 (refs. [6,7]), were also emphasized as having low correlation in this analysis (Pearson correlation, 0.41; $P = 0.21$) as there was a strong deviation between the phosphorylation and underlying serine distributions (Fig. 2d; two-sided *F*-test, $P < 2.2 \times 10^{-16}$). This was identified despite any experimental or computational enrichments for specific modifications, as we used a broad search that was not modification-specific. Beyond confirming known motifs, we also identified new ones. For example, lysine residues are generally under-represented in the HLA I binding pocket at the second position of the peptide. However, modified lysine residue distributions (for example, acetylated and methylated lysine) do not produce the same pattern (Fig. 2e). This suggests that unmodified lysine residues in the anchoring position are unfavorable for HLA I binding and that the modified state of a lysine residue may be preferred. In contrast, modified arginine residues, such as dimethylated and citrullinated arginine, are over-represented in positions 3–7 and, therefore, may impact TCR recognition[47] (Fig. 2f), as was previously shown for

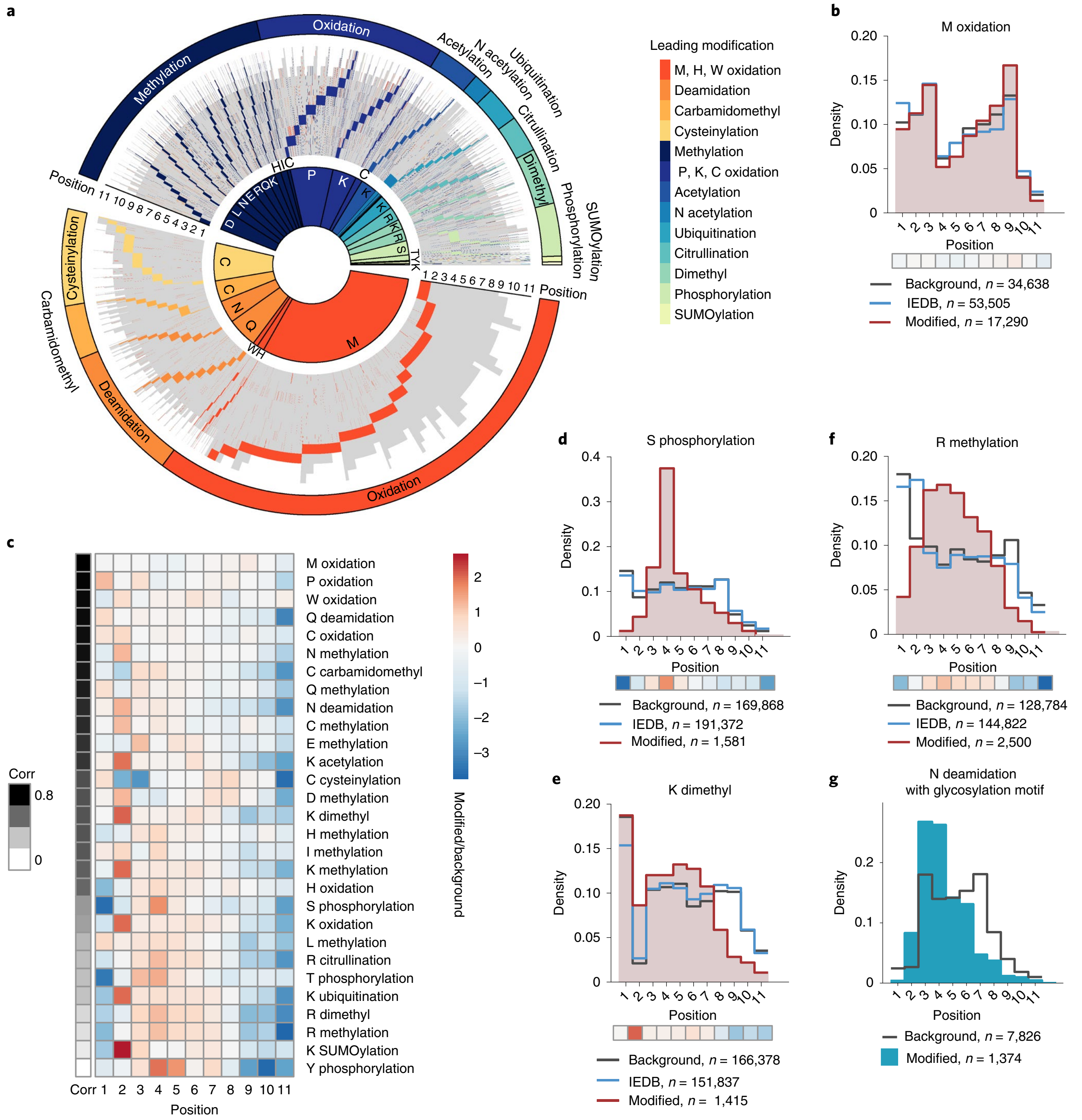


Fig. 2 | PTM-driven binding preference highlighted through an unbiased search of 29 modifications. a, All peptides with modifications identified in the reanalysis of the dataset from ref. [39] by PROMISE ($n = 12{,}268$ peptides) sorted by the modification type and position in the peptide. Each line represents a distinct peptide in gray with the modification site(s) colored. **b**, Correlation between oxidized methionine position distribution and the unmodified methionine distribution is very high (Pearson correlation, 0.96; $P = 1.05 \times 10^{-6}$), and as expected from a technical artifact, the distributions are not significantly different (two-sided *F*-test, $P = 0.3543$). **c**, Modification distributions are sorted by the correlation between the modified amino acid and the unmodified background. A low correlation means that the PTM distribution is distinct from the unmodified background, suggesting a PTM-driven motif. **d**–**f**, We compared the modified amino acid position distribution ('modified', red) to the distribution of the unmodified amino acid that carries this modification ('background', gray) in the analyzed datasets or identified in the IEDB[45] ('IEDB', blue). Major differences between those distributions suggest that the modified amino acid has position preferences not solely determined by the properties of the unmodified amino acid. Below each histogram, the fold change between the modified amino acid and unmodified amino acid distribution is presented as a heatmap bar (red indicates overrepresentation of the modified amino acid relative to the unmodified distribution). **d**, Distribution of serine showing that the phosphorylated form falls predominantly in the fourth position and is significantly different from the unmodified serine distribution (two-sided *F*-test, $P < 2.2 \times 10^{-16}$). **e**, Lysine residues are under-represented at the second position of the peptide; however, the distribution of the dimethylated form is enriched at the second position compared to the background (two-sided *F*-test, $P < 2.2 \times 10^{-16}$). **f**, Methylated arginine is enriched in positions 3–7 compared to background arginine (two-sided *F*-test, $P < 2.2 \times 10^{-16}$). **g**, Deamidated asparagine with a glycosylation motif is enriched in positions 3 and 4 compared to background asparagine (two-sided *F*-test, $P < 2.2 \times 10^{-16}$).

other modification types. Interestingly, while cysteine modifications on peptides in MS analyses are considered to be introduced by sample processing, in our analysis of the HLA I landscape they have a distinct distribution motif where cysteine carbamidomethyl is enriched in positions 3 and 4, and cysteinylation is enriched in positions 7 and 8 (Fig. 2c). The deamidation of asparagine residues occurs naturally at glycosylated sites on proteins[38], and these sites have a strong consensus sequence motif of N-X-S–T. Peptides with asparagine deamidation and the glycosylation motif, suggesting they they are biological in origin, show a distinct tendency for these to be located in the third and fourth positions of the peptides (Fig. 2g; two-sided *F*-test, $P < 2.2 \times 10^{-16}$).

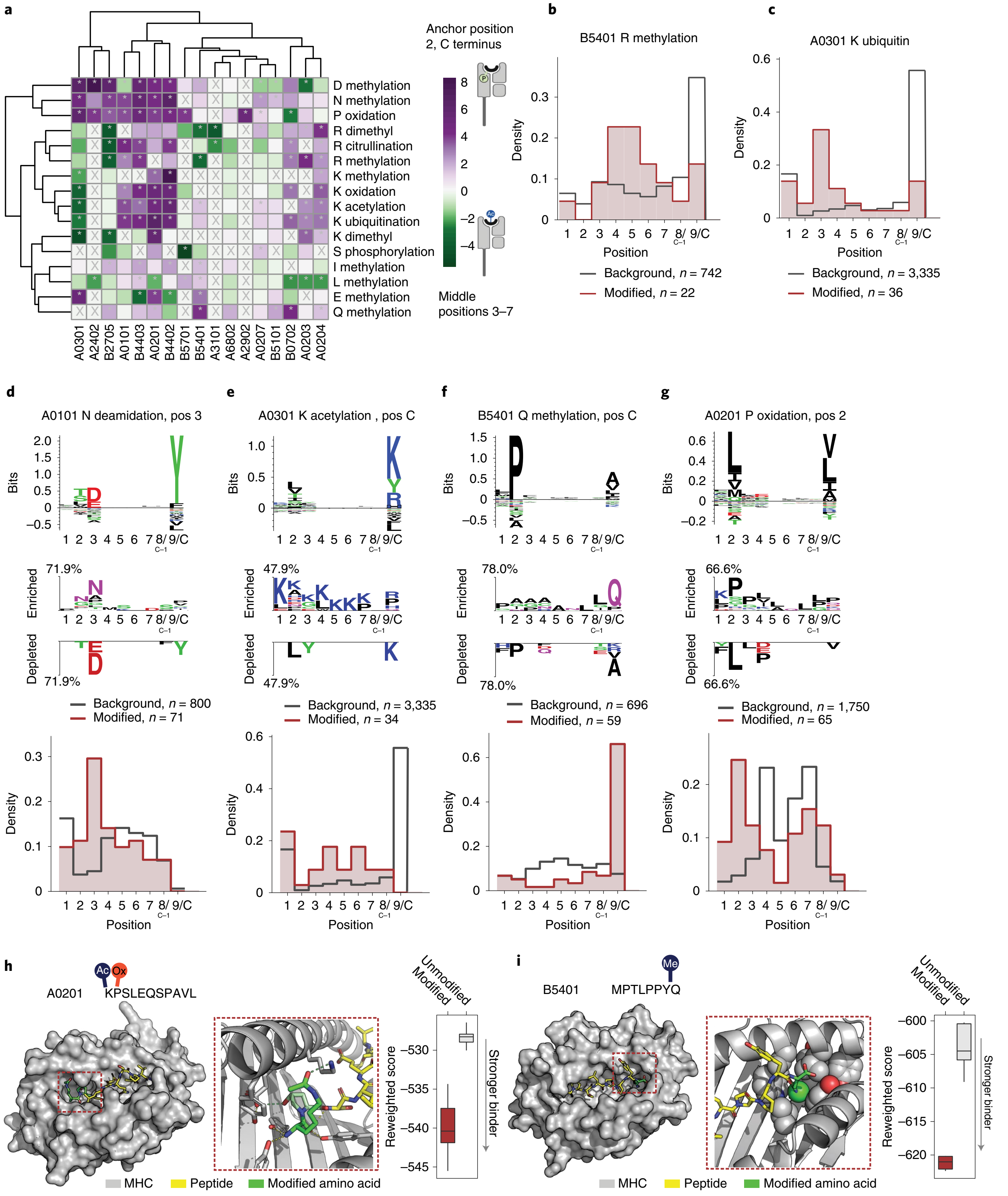
PTMs alter HLA I binding preferences and TCR recognition. The biochemical binding properties of specific HLA haplotypes are the strongest determinants of peptide motifs. To examine whether the PTM-driven motifs we have detected are associated with specific haplotypes, we re-analyzed monoallelic HLA immunopeptidomics data from refs. [18,49] (MassIVE: MSV000080527, MSV000084172—partial). We conducted the same multi-modification search as described (Table 1) on the spectra obtained in this study. Indeed, we could identify unique motifs that were haplotype-dependent, using the unmodified amino acid distribution as a background. To focus on the most prominent features, we defined a 'site score' such that enrichment in the anchor positions will result in a positive score, while enrichment in the middle of the peptide will result in a negative score. In the case where the PTM is present in many positions in the peptide, the score will be close to zero and we cannot classify the tendency of the modification to be in a specific region. We then clustered the biological PTMs and haplotypes contained in the dataset by their site score (Fig. 3a and Extended Data Fig. 2a). This analysis revealed that the same PTM might affect peptide–MHC–TCR interactions differently for different haplotypes. Intriguingly, among the specific HLA haplotypes that we analyzed, we found several HLA associations with human diseases. For example, HLA-A*03:01 was linked to increased risk for multiple sclerosis[50] and HLA-B*51:01 was linked to Behçet's disease[51]. Our analysis identified both haplotypes to be highly enriched for PTMs in the region that is predicted to affect TCR recognition. HLA-A*02:01 was previously reported to show a protective effect in patients with Epstein–Barr virus-related Hodgkin lymphoma[52] and was enriched for modifications on the anchoring position of the peptide in our analysis. While it remains to be examined whether certain PTMs have a role in disease-associated manifestations, it has been reported that low HLA binding of disease-associated epitopes can be increased by PTMs[53]. PTM enrichment in the middle of the peptide, potentially affecting TCR recognition, could be observed with methylated arginine on haplotype HLA-B*54:01 (Fig. 3b) or ubiquitin tail on lysine on haplotype HLA-A*03:01 (Fig. 3c).

PTM enrichment in an anchor position was classified into three groups: the first group comprises chemical mimics, where the modified amino acid is biochemically similar to a different amino acid that was known to be part of the motif. For example, we identified an enrichment of deamidated asparagine in position 3 of the haplotype HLA-A*01:01 motif. Deamidated asparagine is chemically similar to aspartic acid that appears in the HLA-A*01:01 binding motif at position 3 (Fig. 3d). As we could not find an unmodified peptide carrying asparagine bound to this haplotype, this result suggests that the modification occurred on the peptide before it bound to the HLA, possibly due to the removal of glycosylation[38], and the modified asparagine enables binding of the peptide to the HLA. Enrichments of deamidated asparagine and glutamine for HLA haplotypes HLA-A*68:02, HLA-B*44:02 and HLA-B*44:03 (Supplementary Data 3) are additional examples of chemical mimics. Across haplotypes, we were able to show that the chemical mimic PTMs behave more like the amino acid they are mimicking than the residue they modify (Extended Data Fig. 2b). The second group contains PTMs that cause binding interference. This group is defined by PTMs that sterically hinder the interaction of the peptide with the MHC haplotype, creating an unfavorable binder. For example, acetylated lysine is under-represented in the C terminus of haplotype HLA-A*03:01 (Fig. 3e) compared to the unmodified background. Notably, we found this observation to hold true for all of the modified lysines detected for this haplotype, suggesting that modification of the C terminus could be an immune evasion mechanism. Other examples of binding interference are methylated glutamic acid at anchor position 2 for haplotype HLA-B*44:02/3 and dimethylated arginine at the C-terminal position for haplotype HLA-A*31:01 (Supplementary Data 3). The third group is new motifs where the modified amino acid creates a favorable binder peptide that is different from the known unmodified motif. For example, the tendency of phosphoserine to be located in the fourth position (Fig. 2c) was also observed for HLA-B*07:02 and HLA-B*27:05 with the added haplotype resolution from the monoallelic data, revealing the RRXpS and K/RPXpS motifs, as previously described[54] (Extended Data Fig. 2c,d). We also detect methylated glutamine at the peptide C terminus for haplotype HLA-B*54:01 (Fig. 3f), and oxidized proline was observed at the anchor position 2 of haplotype HLA-A*02:01 (Fig. 3g). The latter observation is common to the whole haplotype superfamily HLA-A*02 (Supplementary Data 3).

Next, we evaluated the possibility of a new PTM binding motif using structural modeling. To this end, we chose two modified representative epitopes identified as binders of haplotype HLA-A*02:01 and one representative epitope identified as a binder of haplotype

Fig. 3 | PTM-driven HLA motifs. a, A recognition area score was calculated (Methods) to determine the tendency of a given modification to be located in the MHC anchor position (purple) or center of the peptide (green) for a given HLA haplotype. 'X' marks groups with fewer than 15 peptides, and an asterisk indicates significance in a Benjamini–Hochberg multiple-comparison-corrected χ^2 test ($P \leq 0.05$). **b**,**c**, Histograms representing the modified amino acid frequency in each position (red) compared to the unmodified amino acid background (gray). The C terminus and C-1 are presented at positions 8 and 9 (Methods). **b**, Methylated arginine in haplotype HLA-B*54:01 is enriched in positions 4–6. **c**, A ubiquitin tail on lysine is enriched in position 3 for haplotype HLA-A*03:01. **d**–**g**, Motif of the reported unmodified epitopes in the IEDB for the indicated haplotype (top). The canonical modified motif was then compared to the amino acid motif for a given modification (middle). The histogram represents the modified amino acid frequency in each position (red) compared to the unmodified amino acid background (gray). **d**, Chemical mimics motif: aspartic acid is favored in the HLA-A*01:01 binding motif at position 3. **e**, Binding interference: acetylated lysine is under-represented in the C terminus for haplotype HLA-A*03:01. **f**,**g**, New motif: methylated glutamine at the peptide C terminus in haplotype HLA-B*54:01 and oxidized proline at anchor position two for haplotype HLA-A*02:01 create favorable binder peptides. **h**,**i**, Rosetta FlexPepDock structural models of the interactions between the modified peptide (yellow sticks), including the modified amino acid (green), and the MHC molecule (gray surface\cartoon). The effect of the modified amino acid is shown in detail in the zoom-in picture. FlexPepDock reweighted score was calculated for the interaction between the MHC and modified or unmodified peptide ($n = 5$ simulations; box and whiskers indicate mean and quartiles, respectively). **h**, Interaction between K(ac)P(ox)SLEQSPAVL and haplotype HLA-A*02:01; hydrogen bonds introduced by the modification shown as dashed green lines, others as yellow. **i**, Interaction between MPTLPPYQ(me) and haplotype HLA-B*54:01; the glutamine methyl group is shown as a green sphere, and MHC-interacting residues are shown as gray spheres.

HLA-B*54:01. All of these are shared across cancer cell lines and patient tumor samples. We used Rosetta FlexPepDock[55] to model the structure of the interactions of these new MHC-binding PTM motifs, K(ac)P(ox)SLEQSPAVL, KP(ox)LKVIFV and MPTLPPYQ(me). For each such motif, we modeled both the modified and unmodified peptides and compared their calculated binding energies and structures ('reweighted score'). In all cases, the interactions between the MHC and the modified peptide were predicted to be considerably stronger, suggesting that the complex is more stable than the unmodified counterpart (Fig. 3h,i and Extended Data Fig. 2e), in



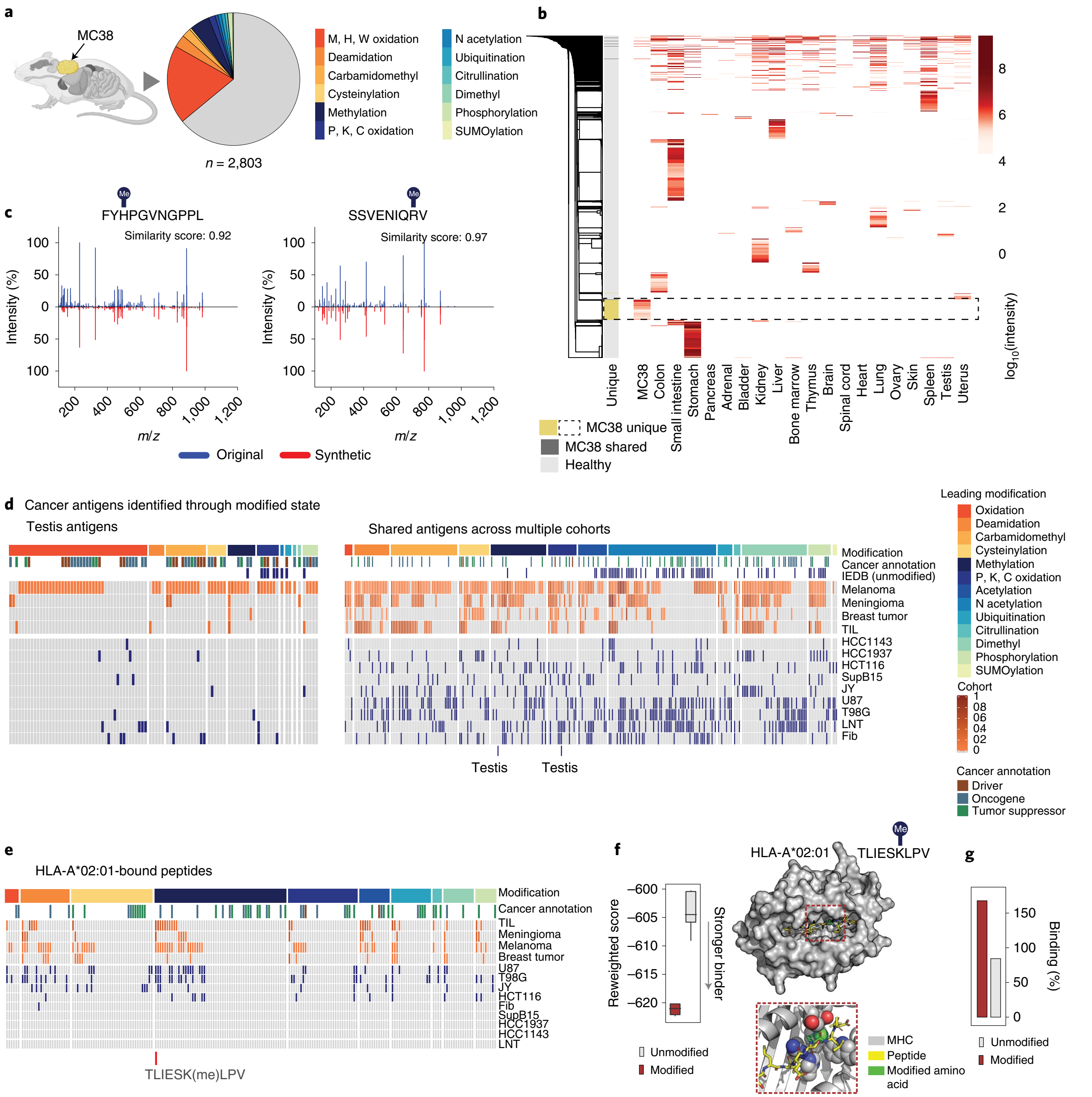


Fig. 4 | The modified HLA I landscape uncovers hundreds of tumor-associated modified antigens, including antigens unique to tumor tissue. **a**, PROMISE analysis of the MC38 immunopeptidome identified 2,803 peptides. **b**, Peptides with biological modifications from PROMISE multi-modification search analysis of 19 different tissues from healthy mice, as well as the MC38 cancer cell line (total number of peptides = 4,535). The peptides are presented by their MS intensity in the relevant column specifying the tissue of origin in which they were identified. Tissues are clustered by similarity, revealing shared modified peptides between MC38 cells and healthy tissue (dark gray panel on the left) and peptides unique to MC38 celss (yellow; dashed square). **c**, For specifically modified peptides, a similarity score was calculated between the synthesized spectrum (red) and the original spectrum in the dataset (blue). **d**,**e**, Each list of antigens is sorted by the modification of the peptide. For each peptide, we mark the cancer annotation as documented in CancerMine[59], whether the peptide was reported in IEDB[45] in its unmodified state and whether it is a cancer-testis antigen. For patient samples (orange), the color indicates the percentage of the patients in whom the peptide was identified. For cancer cell lines (blue), color indicates that the peptide was detected. **d**, Modified cancer-testis antigen list ($n = 98$, left) and a list of shared antigens ($n = 300$, right) identified through the modified state. **e**, A list of HLA-A*02:01 bound modified peptides that were not reported in the IEDB. **f**, Rosetta FlexPepDock structural model of the interactions between TLIESK(me)LPV (yellow sticks) and the HLA-A*02:01 molecule (gray surface/cartoon). The methylated lysine (green) is packed against hydrophobic residues of the MHC molecule (gray spheres). Three-dimensional model reweighted score shows that the modification created a more stable interaction with the MHC molecule ($n = 5$ simulations; box and whiskers indicate mean and quartiles, respectively). **g**, TLIESK(me)LPV in its modified and unmodified form was validated with a binding affinity test through ProImmune in vitro binding assay and matches the 3D model prediction. Binding scores are normalized to an internal standard, which is represented as 100%.

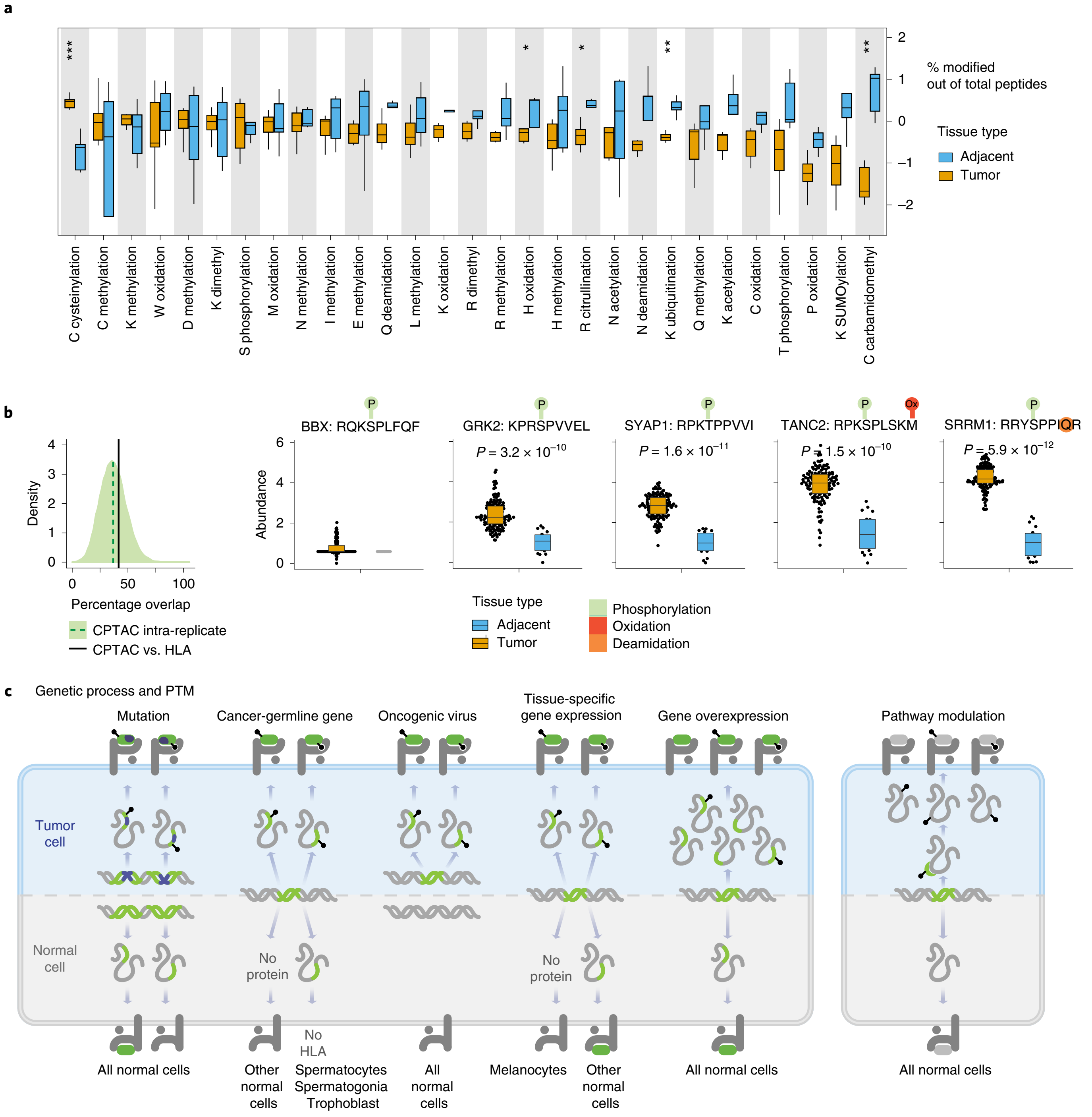


Fig. 5 | Modified HLA-bound peptides create cancer-specific signatures. a, The percentage of immunopeptides identified with each of the indicated modifications was calculated for a cohort of TNBC tumors and adjacent tissue (ref. [42]; $n = 6$ tumors and 5 adjacent tissue samples). The modifications are sorted from the most enriched in tumor tissue at the left to the most enriched in adjacent tissue at the right. A two-sided Student's *t*-test was used to determine significance of the observed change in percentage: cysteine cysteinylation is significantly enriched in tumors (***$P = 0.00045$), while histidine oxidation (*$P = 0.044$), arginine citrullination (*$P = 0.013$), lysine ubiquitination (**$P = 0.0031$) and cysteine carbamidomethylation (**$P = 0.0078$) are significantly enriched in normal tissue (box and whiskers indicate mean frequency and quartiles, respectively). **b**, The percentage of overlapping sites between a randomly chosen subset of the cohort (30 peptides from six samples) and the remaining samples. This was repeated 10,000 times to generate the intrareplicate distribution (light green; the mean is depicted as a dark green dashed line). The overlap between the identified phosphosites in the immunopeptidomics data and the CPTAC data is shown as a black line (top left). The abundance in the CPTAC cohort for the five overlapping phosphosites is shown in the tumor and adjacent tissue samples (two-sided Wilcox *P* values for abudance in tumor versus adjacent tissue are indicated). **c**, Typically, antigenic peptides are classified by their genetic origin, including mutations, cancer-germline genes expressed outside of their biological context, oncogenic virus genes, genes with highly tissue-specific expression patterns or overexpression of genes with low endogenous expression (left block). In all these cases, PTMs can increase both the identification and therapeutic potential of these antigenic peptides. Further, PTMs can themselves be a source of antigenicity when pathways are activated in a disease-specific manner (right block), creating a PTM-driven antigenic epitope.

agreement with the predictions from PROMISE immunopeptidomics analysis. In the case of peptide K(ac)P(ox)SLEQSPAVL binding to HLA-A*02:01, our model suggests that the hydroxyl group of P(ox) at position 2 forms a stabilizing hydrogen bond with E87 in the receptor (Fig. 3h). Overall, our models recapitulate an interaction similar to that in a solved structure of HLA-A*2, in which T2 forms hydrogen bonds with residues K90 and E87 in the receptor (1TVB; ref. [56]). As for K(ac) at position-1, in some of our models it interacts with the aliphatic part of receptor residue K90, while in others it further stabilizes the peptide. In the case of peptide MPTLPPYQ(me) binding to HLA-B*54:01, Q-8 is positioned in the highly hydrophobic pocket that binds the canonical aliphatic C-terminal peptide position. Methylation allows the otherwise polar (negative) side chain of glutamine to approach ('fill') the pocket and thereby stabilize the complex (Fig. 3i).

To calculate the prevalence of new motifs, we used netMHC[17] to calculate a predicted binding score for the unmodified versions of the peptides we identified. Aside from technical modifications, an average of 50% of the modified peptides across the haplotypes we examined were not binders in their unmodified state (Extended Data Fig. 2f). For some PTMs, such as citrullination, above 70% of the peptides were not binders without citrullination (Extended Data Fig. 2f). Together, our findings show that modified peptides are distinct from their unmodified forms in haplotype preference, binding motifs and structural interactions. We, therefore, wished to examine whether PTMs may also alter the antigenic repertoire.

PROMISE uncovers hundreds of tumor-associated modified antigens. To test whether PROMISE-identified peptides can be unique to cancerous tissue, we performed immunopeptidomics on MC38 mouse colon cancer cells and compared the results to healthy mouse tissue immunopeptidomics data from ref. [57]. PROMISE analysis on the MC38 immunopeptidomics data revealed 2,803 peptides, 36% of which had at least one PTM (Fig. 4a and Supplementary Data 4). When comparing modified peptides from the MC38 immunopeptidome to those in healthy tissue, we could identify a subset that were unique to cancer (Fig. 4b). Of note, the differences between the immunopeptidomes for MC38 colon cancer cells and healthy mouse tissue may also arise from differences in sample processing, culture conditions or instrumentation. Thus, we focused on biological PTMs that may represent potential tumor-specific modified antigens. We chose 20 modified peptides that did not appear in healthy tissues and were not reported in IEDB and proceeded to synthesize and validate these peptides in their modified state. All the synthesized peptides were confirmed to match the original identification through manual annotation and scoring of spectrum similarity (Fig. 4c, Extended Data Fig. 3 and Supplementary Information). These included peptides with N-terminal acetylation, citrullination, dimethylation, methylation, phosphorylation and SUMOylation remnants (GGT).

Given the growing interest in identifying antigenic targets for immunotherapy, we examined whether we identified modified peptides originating from cancer-associated or testis antigens across the human cell lines and clinical samples we analyzed. We identified 98 peptides that originated from a protein annotated as a testis antigen (CT Antigens Database[58]; Fig. 4d, left). For these, we examined their mRNA expression in The Cancer Genome Atlas (TCGA) data of the matching cancer types (Extended Data Fig. 4a) and found a subset to be overexpressed in tumor tissue when compared to the adjacent controls (Extended Data Fig. 4b). We also identified 300 peptides that are highly shared between patients and across cancer cohorts (Fig. 4d, right). Many of these proteins are also annotated as oncogenes, cancer drivers or tumor suppressors[59], highlighting the importance of studying the state of these proteins in tumor immunogenicity. None of these cancer-associated target peptides would have been identified without including PTMs in the protein search space. To show that PROMISE can accurately identify modified epitopes through retrospective analysis, we synthesized 20 peptides that were shared across multiple patients. These epitopes were not reported previously in the IEDB in their modified state. All the peptides were confirmed to match the original identification through manual annotation and scoring of spectrum similarity (Extended Data Fig. 5 and Supplementary information).

To focus on the modified peptides we identified with PROMISE associated with a specific haplotype, we filtered for modified peptides that were identified in immunopeptidomics data from an HLA-A*02:01 cell line and that were not identified in IEDB in their unmodified form (Fig. 4e). We then examined whether the difference in detection of the modified peptides and their unmodified counterparts was due to their relative ability to bind HLA-A*02:01. Using structural modeling, we were able to show that methylation on the lysine in position 6 of TLIESKLPV is located between three other positively charged residues (H98, R121 and H138; Fig. 4f). Methylation of K6 removes its positive charge and thereby alleviates electrostatic repulsion. In addition, the methyl group is nicely packed into the hydrophobic MHC groove. This then causes a more stable peptide–MHC interaction as reflected in a lower reweighted score (Fig. 4f). To validate our model, we synthesized the peptide in both modified and unmodified forms and examined its binding against a standard using a binding assay (ProImmune; Methods). Both peptide forms were confirmed as HLA I binders, while the modified form was a stronger binder as the model predicted (Fig. 4g).

Intracellular cancer-associated PTMs are presented on HLA I. To determine whether these signatures are also specific to the cancer state in clinical settings, we analyzed immunopeptidomics data from a cohort of triple-negative breast cancer (TNBC) and adjacent tissue[42]. Within this cohort, we found 2,771 modified peptides. We assessed whether there were classes of PTMs that were more frequent in the immunopeptidome of the tumor samples versus their adjacent controls. We found several modifications that were significantly reduced in frequency in the tumor immunopeptidome, including carbamidomethyl and citrullination modifications (Fig. 5a). Further, we found that the frequency of cysteinylated peptides was significantly increased in the tumor immunopeptidome. The tumor and adjacent tissues were processed and analyzed together and, therefore, are not expected to exhibit differential effects on modifications that were generated merely by the processing procedures. As such, the results likely signify changes in modifications elicited by the biological system. These changes may reflect alterations in metabolic pathways or peptide processing. For example, it is known that TNBC is addicted to cysteine[60,61], potentially explaining the increase in cysteinylated immunopeptides.

Although the frequency of phosphorylation did not differ significantly between the tumor tissue and the adjacent control, we found 27 phosphorylated peptides, that only appeared in the tumor tissue and not in the adjacent control. We hypothesized that these tumor-specific phosphopeptides might originate from proteins that are phosphorylated more in breast tumor tissue. To examine this, we compared the immunopeptidomics data to clinical phosphoproteomics data. Surprisingly, of the sites identified in both immunopeptidomics and phosphoproteomics, 42% were phosphorylated in both (Fig. 5b). This is despite the fact that, on an average, when comparing between different samples in phosphoproteomics, there is only a 37% overlap in phosphosites (Fig. 5b). Furthermore, of the phosphosites identified in both cohorts, all had increased frequency in the tumor compared to adjacent tissue, both on the phosphoprotein and HLA I-bound peptide levels (Fig. 5b). This suggests that tumor-induced alterations of modifications on cellular proteins can propagate to changes in the presented landscape.

Discussion

By developing PROMISE, we systematically analyzed the PTM landscape in the immunopeptidome and identified thousands of modified peptides across different cancers. Although numerous studies have examined HLA I presentation of modified peptides in the context of tumor antigenicity and autoimmune disease[6–10,62], such analysis relied on experimental enrichment of the modification of interest. The capability to search a large number of PTMs allowed us to identify types of modifications that were not examined before in the context of antigen presentation. For example, recent studies have suggested that the proximal ubiquitin may undergo proteasome degradation with its substrate[63,64]. Indeed, we could detect some remnants of ubiquitin-like modifications in our analyses. However, while we were able to validate their spectra, indicating that these are true identifications, it is not yet clear that they are loaded and presented on MHC I.

Modified peptide analysis, coupled with structural modeling and binding assays, strongly suggests that modifications may generate new HLA I binding motifs that could not be identified merely by the amino acid sequence. For example, cysteine is under-represented in HLA I ligand datasets[44], hampering accurate binding predictions of cysteine-containing peptides[65]. However, by including several cysteine modification types in our search space, we could identify presented peptides containing cysteine with distinct motifs. Another example is the under-representation of unmodified lysine residues in the second position anchoring site in the reported epitopes in the IEDB compared to the presence of modified lysine at this position. Notably, some of the peptides that we have identified do not match the consensus binding motif (8-11 mers). This may be driven by the PTM or reflect previous observations of binding of longer peptides[66–72]. As binding motifs are the dominant selection criteria for antigen prediction algorithms[49,65], PTM-driven motifs should prove invaluable to the next generation of binding prediction software[62]. It will be intriguing to examine PROMISE in the context of additional modifications, mutations and the MHC II repertoire. Due to current software limitations, analyses of cryptic, spliced, noncanonical and modified peptides[22,30–34,73] are usually carried out in isolation, preventing an examination of how these search spaces interact. It will be interesting to use PROMISE for an extended analysis that combines multiple new reference spaces and examines potential improvements in peptide identification and illumination of the dark matter of the immunopeptidome.

Beyond expanding the HLA landscape, modified peptides may also signal changes in metabolic and signaling states of the cells under physiological or pathological circumstances. Notably, by comparing cancerous tissues and adjacent controls in TNBC, we found changes in the frequency of different modification types in the immunopeptidome. Further, we could confirm that 40% of the phosphorylation sites identified on HLA I-bound peptides also exhibited increased abundance in phosphoproteomics analysis of breast cancer (CPTAC data). We note that detection of a peptide may be due to both a higher abundance of the protein and increased phosphorylation. Nevertheless, these results suggest that intracellular changes in the phospho-state of proteins, before degradation, may be kept and loaded onto HLA for presentation to create unique HLA signatures in breast tumors (Fig. 5c). Although beyond the scope of this study, this also raises the intriguing possibility that drug-induced alterations in the activation of specific pathways may, in turn, alter the HLA repertoire. In the future, this feature may be used to direct immune responses against specific antigenic peptides in combination with targeted therapies[1–4].

Previous studies[6–12,74,75], together with our analyses, highlight the potential of modified antigens to have an immunomodulatory role in tumor–host interactions and may drive either immune suppression or immune evasion. While this class of modified antigens may offer new therapeutic opportunities, there are important questions that remain to be addressed before these may be used for cancer therapy. PTMs are more labile than mutations and, therefore, changes in cell source, processing, laboratory settings or instrumentation can all affect peptide modification status. As PTMs become a more routine addition to immunopeptidomics analyses, efforts will need to be taken to standardize practices for their robust detection and assignment. Further, in determining the use of PTMs as cancer-specific targets, comparing potential immunogenic modifications between healthy and cancerous tissue will be needed to determine specificity, heterogeneity and stability in the context of T cell recognition. Nevertheless, our analyses identified hundreds of modified testis antigens and tumor-associated peptides, which may serve as a new source of modified neoantigens in the context of immunotherapy. Coupled with patient-specific modifications, which occur sporadically and can be targeted for individualized therapy, we foresee that a broad range of potentially therapeutic antigens may be detected when analyzing peptide modification states. Beyond cancer, our approach may be used to expand our understanding of the PTM-driven HLA repertoire across different human pathologies, ranging from infectious diseases to autoimmunity and neurodegeneration.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01464-2.

Received: 2 November 2021; Accepted: 9 August 2022;
Published online: 6 October 2022

References

1. Obara, W. et al. Present status and future perspective of peptide-based vaccine therapy for urological cancer. *Cancer Sci.* **109**, 550–559 (2018).
2. Jiang, D., Niwa, M., Koong, A. C. & Diego, S. Cancer immunotherapy: moving forward with peptide T cell vaccines. *Eur. J. Vasc. Endovasc. Surg.* **49**, 48–56 (2016).
3. Xia, A.-L., Wang, X.-C., Lu, Y.-J., Lu, X.-J. & Sun, B. Oncotarget chimeric-antigen receptor T (CAR-T) cell therapy for solid tumors: challenges and opportunities. *Oncotarget* **8**, 90521–90531 (2017).
4. Finn, O. J. & Rammensee, H. G. Is it possible to develop cancer vaccines to neoantigens, what are the major challenges, and how can these be overcome? Neoantigens: nothing new in spite of the name. *Cold Spring Harb. Perspect. Biol.* **10**, a028829 (2018).
5. Hsiue, E. H. C. et al. Targeting a neoantigen derived from a common *TP53* mutation. *Science* **371**, eabc8697 (2021).
6. Alpízar, A. et al. A molecular basis for the presentation of phosphorylated peptides by HLA-B antigens. *Mol. Cell. Proteomics* **16**, 181–193 (2017).
7. 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).
8. Mohammed, F. et al. The antigenic identity of human class I MHC phosphopeptides is critically dependent upon phosphorylation status. *Oncotarget* **8**, 54160–54172 (2017).
9. Marcilla, M. et al. Increased diversity of the HLA-B40 ligandome by the presentation of peptides phosphorylated at their main anchor residue. *Mol. Cell. Proteomics* **13**, 462–474 (2014).
10. Marino, F. et al. Arginine (di)methylated human leukocyte antigen class I peptides are favorably presented by HLA-B*07. *J. Proteome Res.* **16**, 34–44 (2017).
11. Malaker, S. A. et al. Identification of glycopeptides as posttranslationally modified neoantigens in leukemia. *Cancer Immunol. Res.* **5**, 376–384 (2017).
12. Petersen, J., Purcell, A. W. & Rossjohn, J. Post-translationally modified T cell epitopes: immune recognition and immunotherapy. *J. Mol. Med.* **87**, 1045–1051 (2009).
13. Ramarathinam, S.H., Croft, N.P., Illing, P.T., Faridi, P. & Purcell, A.W. Employing proteomics in the study of antigen presentation: an update. *Expert Rev. Proteomics* **15**, 637–645 (2018).
14. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

15. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225 (2020).
16. Karasaki, T. et al. Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. *Cancer Sci.* **108**, 170–177 (2017).
17. Jurtz, V. et al. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368 (2017).
18. Abelin, J. G. et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326 (2017).
19. Gfeller, D. et al. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol.* **201**, 3705–3716 (2018).
20. Bulik-Sullivan, B. et al. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* **37**, 55–71 (2019).
21. O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* **11**, 42–48 (2020).
22. Ouspenskaia, T. et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* **40**, 209–217 (2022).
23. Chong, C., Coukos, G. & Bassani-Sternberg, M. Identification of tumor antigens with immunopeptidomics. *Nat. Biotechnol.* **40**, 175–188 (2022).
24. Yu, F. et al. Identification of modified peptides using localization-aware open search. *Nat. Commun.* **11**, 4065 (2020).
25. Devabhaktuni, A. et al. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat. Biotechnol.* **37**, 469–479 (2019).
26. Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. Enhanced global post-translational modification discovery with metaMorpheus. *J. Proteome Res.* **17**, 1844–1851 (2018).
27. Zhang, J. et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).
28. Geiszler, D.J. et al. PTM-Shepherd: analysis and summarization of post-translational and chemical modifications from open search results. *Mol. Cell Proteomics* **20**, 100018 (2021).
29. Skinner, O. S. & Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **33**, 717–718 (2015).
30. Laumont, C. M. et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238 (2016).
31. Starck, S. R. & Shastri, N. Nowhere to hide: unconventional translation yields cryptic peptides for immune surveillance. *Immunol. Rev.* **272**, 8–16 (2016).
32. Erhard, F., Dölken, L., Schilling, B. & Schlosser, A. Identification of the cryptic HLA-I immunopeptidome. *Cancer Immunol. Res.* **8**, 1018–1026 (2020).
33. Liepe, J., Sidney, J., Lorenz, F. K. M., Sette, A. & Mishto, M. Mapping the MHC class I-spliced immunopeptidome of cancer cells. *Cancer Immunol. Res.* **7**, 62–76 (2019).
34. Faridi, P. et al. Comment on "A subset of HLA-I peptides are not genomically templated: evidence for *cis*- and *trans*-spliced peptide ligands". *Sci. Immunol.* **4**, eaaw1622 (2019).
35. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).
36. Gurd, F. R. N. et al. Overalkylation of a protein digest with iodoacetamide. *Proc. Natl Acad. Sci. U. S. A.* **25**, 3576–3582 (1991).
37. Du, Y., Wang, F., May, K., Xu, W. & Liu, H. Determination of deamidation artifacts introduced by sample preparation using ^{18}O-labeling and tandem mass spectrometry analysis. *Anal. Chem.* **84**, 6355–6360 (2012).
38. Mei, S. et al. Immunopeptidomic analysis reveals that deamidated HLA-bound peptides arise predominantly from deglycosylated precursors. *Mol. Cell. Proteomics* **19**, 1236–1247 (2020).
39. 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).
40. Chong, C. et al. High-throughput and sensitive immunopeptidomics platform reveals profound interferonγ-mediated remodeling of the human leukocyte antigen (HLA) ligandome. *Mol. Cell. Proteomics* **17**, 533–548 (2018).
41. 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. Proteom.* **15**, 3058–3070 (2016).
42. Ternette, N. et al. Immunopeptidomic profiling of HLA-A2-positive triple negative breast cancer identifies potential immunotherapy target antigens. *Proteomics* **18**, 1700465 (2018).
43. Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*(Suppl 16), S1 (2012).
44. Bassani-Sternberg, M. et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* **13**, e1005725 (2017).
45. Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
46. Deres, K., Beck, W., Faath, S., Jung, G. & Rammensee, H. G. MHC/peptide binding studies indicate hierarchy of anchor residues. *Cell. Immunol.* **151**, 158–167 (1993).
47. MacLachlan, B. J. et al. Using X-ray crystallography, biophysics, and functional assays to determine the mechanisms governing T-cell receptor recognition of cancer antigens. *J. Vis. Exp.* **120**, 54991 (2017).
48. Wang, Y. et al. How an alloreactive T-cell receptor achieves peptide and MHC specificity. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4792–E4801 (2017).
49. Sarkizova, S. et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
50. Fogdell-Hahn, A., Ligers, A., Gronning, M., Hillert, J. & Olerup, O. Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. *Tissue Antigens* **55**, 140–148 (2000).
51. Wallace, G. R. HLA-B*51 the primary risk in Behçet disease. *Proc. Natl. Acad. Sci.* **111**, 8706–8707 (2014).
52. Hjalgrim, H. et al. *HLA-A* alleles and infectious mononucleosis suggest a critical role for cytotoxic T-cell response in EBV-related Hodgkin lymphoma. *Proc. Natl Acad. Sci. U. S. A.* **107**, 6400–6405 (2010).
53. Sidney, J. et al. Low HLA binding of diabetes-associated CD8+ T-cell epitopes is increased by post translational modifications. *BMC Immunol.* **19**, 12 (2018).
54. Alpízar, A. et al. A molecular basis for the presentation of phosphorylated peptides by HLA-B antigens. *Mol. Cell. Proteomics* **16**, 181–193 (2016).
55. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* **78**, 2029–2040 (2010).
56. Borbulevych, O. Y., Baxter, T. K., Yu, Z., Restifo, N. P. & Baker, B. M. Increased immunogenicity of an anchor-modified tumor-associated antigen is due to the enhanced stability of the peptide/MHC complex: implications for vaccine design. *J. Immunol.* **174**, 4812–4820 (2005).
57. Schuster, H. et al. A tissue-based draft map of the murine MHC class I immunopeptidome. *Sci. Data* **5**, 180157 (2018).
58. Almeida, L. G. et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* **37**, D816–D819 (2009).
59. Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R. & Jones, S. J. M. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* **16**, 505–507 (2019).
60. Timmerman, L. A. et al. Glutamine sensitivity analysis identifies the xCT antiporter as a common triple-negative breast tumor therapeutic target. *Cancer Cell* **24**, 450–465 (2013).
61. Tang, X. et al. Cystine addiction of triple-negative breast cancer associated with EMT augmented death signaling. *Oncogene* **36**, 4235–4242 (2017).
62. Solleder, M. et al. Mass spectrometry based immunopeptidomics leads to robust predictions of phosphorylated HLA class I ligands. *Mol. Cell. Proteomics* **19**, 390–404 (2020).
63. Singh, S. K. et al. Synthetic uncleavable ubiquitinated proteins dissect proteasome deubiquitination and degradation, and highlight distinctive fate of tetraubiquitin. *J. Am. Chem. Soc.* **138**, 16004–16015 (2016).
64. Sun, H. et al. Diverse fate of ubiquitin chain moieties: the proximal is degraded with the target, and the distal protects the proximal from removal and recycles. *Proc. Natl Acad. Sci. USA* **116**, 7805–7812 (2019).
65. Nielsen, M., Andreatta, M., Peters, B. & Buus, S. Immunoinformatics: predicting peptide–MHC binding. *Annu. Rev. Biomed. Data Sci.* **3**, 191–215 (2020).
66. Hassan, C. et al. Naturally processed non-canonical HLA-A∗02:01 presented peptides. *J. Biol. Chem.* **290**, 2593–2603 (2015).
67. Bade-Döding, C. et al. The impact of human leukocyte antigen (HLA) micropolymorphism on ligand specificity within the HLA-B*41 allotypic family. *Haematologica* **96**, 110–118 (2011).
68. Burrows, S. R., Rossjohn, J. & McCluskey, J. Have we cut ourselves too short in mapping CTL epitopes? *Trends Immunol.* **27**, 11–16 (2006).
69. Ebert, L. M. et al. A long, naturally presented immunodominant epitope from NY-ESO-1 tumor antigen: implications for cancer vaccine design. *Cancer Res.* **69**, 1046–1054 (2009).
70. Hassan, C. et al. The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell. Proteomics* **12**, 1829–1843 (2013).
71. Mommen, G. P. M. et al. Expanding the detectable HLA peptide repertoire using electron-transfer/ higher-energy collision dissociation (EThcD). *Proc. Natl Acad. Sci. USA* **111**, 4507–4512 (2014).

72. Probst-Kepper, M. et al. An alternative open reading frame of the human macrophage colony-stimulating factor gene is independently translated and codes for an antigenic peptide of 14 amino acids recognized by tumor-infiltrating CD8 T lymphocytes. *J. Exp. Med.* **193**, 1189–1198 (2001).
73. Bartok, O. et al. Anti-tumour immunity induces aberrant peptide presentation in melanoma. *Nature* **590**, 332–337 (2020).
74. Cobbold, M. et al. MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia. *Sci. Transl. Med.* **5**, 203ra125 (2013).
75. Mohammed, F. et al. Phosphorylation-dependent interaction between antigenic peptides and MHC class I: a molecular basis for presentation of transformed self. *Nat. Immunol.* **9**, 1236–1243 (2009).
76. Kim, M., Zhong, J. & Pandey, A. Common errors in mass spectrometry-based analysis of posttranslational modifications. *Proteomics* **16**, 700–714 (2017).
77. Li, Y., Silva, J. C., Skinner, M. E. & Lombard, D. B. Mass spectrometry-based detection of protein acetylation. *Methods Mol. Biol.* **1077**, 81–104 (2013).
78. Verrastro, I., Pasha, S., Jensen, K. T., Pitt, A. R. & Spickett, C. M. Mass spectrometry-based methods for identifying oxidized proteins in disease: advances and challenges. *Biomolecules* **5**, 378–411 (2015).

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

Methods

PROMISE. Current proteomics software focuses on data from samples where an exogenous enzyme, such as trypsin, was used to digest the proteins into peptides. This reduces the potential search space to only peptides with either lysine or arginine terminal residues. By contrast, HLA I peptides are cleaved by the proteasome and a number of endopeptidases, generating peptides that are between 8 and 15 residues and may have any terminal residue. Computationally, this means that the search space for endogenously cleaved peptides with modifications must contain every potential protein fragment with multiple potential mass shifts, leading to an exponential growth of the search space and making the duration of the search challenging[79]. PROMISE optimizes search efficiency with the following two stages: (1) matching phase and (2) a prioritizing phase (Supplementary Information). The matching phase reduces the algorithm running time, using the ultrafast MSFragger[35] software and parallel computing on a CPU cluster. The prioritizing phase includes several computational steps to distinguish between true and false hits, validate PTM identifications and site position and rank predictions by their biological relevance and antigenic potential. To evaluate pipeline performance, we used the total human proteome from UniProtKB as reference data and searched for endogenous proteasome-cleaved peptides[80] (length between 6 and 40 amino acids) with five variable modifications, creating a search space of ~31 billion potential peptides. In cases where the PSMs conflicted between standard and multi-modification searches (1.22% of PSMs), PROMISE prioritized the highest scoring match. Although the scoring alone is not a guarantee of a true assignment, it does suggest that the inclusion of a modification in the predicted peptide better describes the spectrum.

For the analyses in the manuscript, we used a subgroup FDR, whereby we split the identifications into the following three groups: unmodified, standard search modification types (N-acetylation and methionine oxidation) and the other modification types. For MC38 immunopeptidomics, where the cohort was too small to successfully execute subgroup FDR (Fig. 4) and where additional enrichment analyses were being performed (Figs. 2, 3 and 5a), we used a global FDR. In both cases, the cutoff was set to 5%. In cases where subgroup FDR was used across multiple cohorts, we included any peptide that passed the subgroup FDR in at least one cohort. Detailed software architecture and performance can be found in the Supplementary Information.

Modification annotation and classification. To assess the effects of modifications in a holistic manner, we considered both modifications that may arise during sample processing or handling ('technical') and ones that reflect an altered cellular state ('biological'). This was done using the UNIMOD classification system (unimod.org) as indicated in Table 1. The inclusion of technical modifications in the search space allows the assignment of peptides that would otherwise be missed as they are captured as spectra only in their modified form. We explicitly note that the UNIMOD classification of the modification is not sufficient to determine whether the modified form was generated due to biological regulation or whether the peptide is presented in its modified form. Some modification types arise from multiple sources. As peptides may exist in the cell in either their modified or unmodified form, we chose peptides for validation that were most likely to contain biological modifications that differed between the cancerous and control conditions. When a peptide contained multiple modification types, we defined a leading modification, prioritizing biological modifications over some that may be considered technical (based on the UNIMOD classification).

In this work, there are some analyses which deconvolute the biological and technical origins of PTMs. These include the following:

a. Analysis of the asparagine deamidation motif, which can arise from either removal of glycosylation (biological) or processing of the sample (Fig. 2g).
b. Our analysis of chemical mimics (for example, glutamine or asparagine deamidation mimicking glutamic or aspartic acid, respectively). This suggests that some cases of deamidation occurred before HLA binding and are therefore biological in origin (Fig. 3d).
c. Paired analysis where sample processing was done together, such as the TNBC analysis (Fig. 5a).

Search mass boundary effect correction. The search space in the analysis is bounded by a peptide length of 15. This can result in incorrect assignments when a contaminant with a mass higher than 15 amino acids is assigned to a 15-mer peptide with a high mass shift modification. As we search for PTMs with large mass shifts (for example, a ubiquitin tail with the four-residue GGRL sequence of—383.228103 Da), this can lead to misassigned spectra. Because the longer peptide is not part of our search space, we cannot rule out the possibility that a better match exists or that there is a higher scoring match above 15 amino acids. Therefore, to avoid a bias, we filter out potential misassignments by limiting the total peptide mass to the average mass of peptides with 15 amino acids plus 100 Da when comparing peptide lengths (Fig. 1j).

HLA I motif. HLA I motif presentation was designed to capture both the main anchor position 2 and the C terminus and the TCR recognition area (positions 3–7). The presented motif was created by collecting all the epitopes reported for the specific HLA haplotype from the IEDB[45]. Epitopes with lengths of less than eight amino acids were discarded. To correct the discrepancies in length, the motif was constructed from positions 1–7, starting from the N terminus followed by the C terminus and its preceding position. For 9-mer epitopes, the motif is taken from all nine positions; for 8-mer epitopes, the seventh position is duplicated and presented as both positions 7 and 8/C-1. For epitopes longer than nine residues, the motif skips positions 8 untill the C -1 residue. Motif logos were plotted using Seq2Logo 2.0 (ref. [81]) with default parameters. The comparable motif was created using Two-Sample-Logo[82].

Site score. The score was designed to determine whether a PTM tends to fall within the peptide anchor positions or the center positions of the peptide. We manually determined anchoring or middle positions per haplotype based on the canonical binding motif (Supplementary Information). The percentage of modified residues or background unmodified residues was summed up for the anchor and middle position in each haplotype. Then, a site score odds ratio was calculated as follows:

$$\frac{Anchor_{mod}/Middle_{mod}}{Anchor_{bckgrnd}/Middle_{bckgrnd}}$$

An enrichment in the anchor positions will result in a high positive score, while an enrichment in the center of the peptide will result in a negative score. Each haplotype–PTM pair was also given a significance value, based on a χ^2 test comparing the percentage distribution between the anchor and the middle, in the modified and background residues. Benjamini–Hochberg correction was used to control for multiple hypothesis testing.

TCGA and CPTAC analysis. TCGA data were mined using the xenaPython package in Python 3.6. The results shown in this analysis are in whole or part based upon data generated by the TCGA Research Network available at http://cancergenome.nih.gov/. Colon adenocarcinoma (COAD), breast cancer (BRCA), skin cutaneous melanoma (SKCM) and glioblastoma (GBM) cohort data were used. Data used in this publication were generated by the CPTAC (NCI–NIH). The CPTAC breast cancer phosphoproteomics data[14] were compared to the TNBC immunopeptidomics data[42]. The CPTAC intrareplicate site overlap was calculated from the tumor samples in the cohort by randomly drawing 30 phosphosites from six samples in the same TMT experiment and comparing the identification to the remaining TMT experiments. This was done 10,000 times and is presented in Fig. 5b. The overlap between the CPTAC phosphoproteomics and immunopeptidomics data was defined as the number of phosphosites identified in both CPTAC and immunopeptidomics data ($n=5$) out of the sites that were covered by peptides in both datasets ($n=12$). The remaining 18 sites only had peptides covering them in the immunopeptidomics data and, therefore, could not be evaluated in the tryptic CPTAC cohort.

Modeling the peptide–receptor complex. *General modeling scheme.* The FlexPepBind scheme used in refs. [83,84] allows the structure-based evaluation of the relative binding affinities of different peptides for a given receptor, using a solved structure of a representative peptide–protein interaction as a template. Structures of peptide–MHC complexes were generated by 'threading' candidate peptide sequences onto this template, followed by refinement using Rosetta FlexPepDock[55]. The top-scoring models were selected to discriminate stronger from weaker binders and inspected for the structural details of an interaction.

Selection of templates for modeling. For each of the MHC alleles (receptors) and peptides, we evaluated different available Protein Data Bank (PDB) structures to serve as templates for the modeling of the structure and relative binding affinities of different peptides. Screening for relevant PDB templates was guided by the following three main requirements: (1) matching MHC allele, (2) matching peptide length and (3) similarity of peptide anchor residues. Specifically, for peptide K(ac)P(ox)SLEQSPAVL bound to HLA-A*02 (Fig. 3h), we used PDB code 5D9S (ref. [85]; HLA-A*02 bound to FVLELEPEWTV); for peptide KP(ox)LKVIFV bound to HLA-A*02 (Extended Data Fig. 2e), we used the peptide backbone from PDB code 4F7T (ref. [86]; HLA-A24 bound to RYGFVANF) and the same MHC receptor structure (PDB code 5D9S); and for peptide MPTLPPYQ(me) bound to HLA-B*54 (Fig. 3i), we used PDB code 3BWA (ref. [87]; HLA-B35 bound to FPTKDVAL). Residues that differed between the MHC alleles were 'mutated' using the fix backbone protocol (Rosetta fix_bb); for peptide TLIESK(me)LPV bound to HLA-A*02 (Fig. 4f), we used PDB code 3MRK (HLA-A*02 bound to PLFQVPEPV).

Modeling peptide onto MHC receptor using the selected template. Using the Rosetta fixbb protocol for fixed backbone design[88], we modeled the desired peptide sequence onto the template peptide, while keeping the side chains of the receptor fixed. We then used Rosetta FlexPepDock refinement in full-atom mode to optimize the structure of the complex with the threaded target peptide (all peptide atoms, as well as the receptor interface sidechains, were allowed to move). For each sequence, we generated 200 models. These were scored, and the top five models were selected to represent the MHC–peptide interaction of interest. Comparison of the top-scoring models of the modified peptides and corresponding unmodified peptides allowed inspection of the atomic details of their differential binding.

Scoring function. The standard Rosetta score function[89,90] was used and models were assessed according to their FlexPepDock reweighted score (sum of total score, interface score and peptide score; where total score is the overall Rosetta energy score for the complex, interface score is the energy of pair-wise interactions across the peptide–protein interface and peptide score is the sum of the Rosetta energy function over the peptide residues). This score was shown to discriminate well near-native structures in previous FlexPepDock modeling studies[91].

ProImmune binding assay. The ProImmune (https://www.proimmune.com) Module 2 REVEAL binding assay measures the yield of correctly conformed MHC–peptide complexes following incubation of the recombinant MHC allele with the peptide of interest, using a conformation-dependent antibody in an immunoassay. Each peptide is given a score relative to the positive control peptide, which is a known T cell epitope.

Reagents. A complete list of reagents, antibodies and chemicals can be found in the Supplementary Information.

Purification and analysis of the MC38 immunopeptidome. MC38 cells were kindly provided by A. Erez (Weizmann Institute). Cells were grown in DMEM with 10% FCS, glutamine and sodium-pyruvate at 37 °C with 5% CO_2. Cells were detached with trypsin and washed extensively before further processing. H2-Kb and H2-Db-bound peptides were isolated from three independent preparations of the MC38 cell line, each containing 5×10^8 cells, as in ref. [92]. Briefly, cells were lysed with lysis buffer comprising PBS supplemented with 0.25% sodium deoxycholate, 0.2 mM iodoacetamide, 1 mM EDTA, 1:200 protease inhibitor cocktail (Sigma), 1 mM PMSF and 1% octyl-β-D-glucopyranoside. The lysate was then shaken on a shaking table gently for 1 h at 4 °C and cleared by centrifugation at 4 °C and 47,580 *g* for 60 min (Sorval RC 6+ centrifuge, Thermo Fisher Scientific). After centrifugation, the supernatant was passed through a column containing the Y3 antibody (anti-H2-Kb) or 28-14-8 antibody (anti-H2-Db) covalently bound to Protein G Sepharose resin with dimethyl pimelimidate. Next, the columns were preconditioned with two column volumes of 0.1 N acetic acid followed by two column volumes of 20 mM Tris-HCl, pH 8.0. After passing the cleared cell extracts, the columns were washed with five column volumes of 400 mM NaCl and 20 mM Tris-HCl pH 8, followed by another wash with 20 mM Tris-HCl, pH 8. The MHC-bound peptides were eluted with 1% trifluoroacetic acid (TFA), desalted, concentrated and separated from the MHC molecules by reversed-phase fractionation using disposable Micro-Tip Columns C-18 (Harvard Apparatus). The peptides were eluted with 30% acetonitrile in 0.1% TFA, dried by vacuum centrifugation, and dissolved in 0.1% TFA for analysis by capillary chromatography combined with tandem MS (LC–MS/MS). Samples were resolved by capillary chromatography using an UltiMate 3,000 RSLC coupled by electrospray, to a Q-Exactive-Plus mass spectrometer (Thermo Fisher Scientific). Elution of the peptides was performed with a linear 2-h, 5–28% acetonitrile gradient in 0.1% formic acid, at a flow rate of 0.15 µl min^{-1}. The ten most intense ions in each full MS spectrum, with single- to triple-charged states, were selected for fragmentation by higher energy collision dissociation, at a relative collision energy of 25. Ion times were set to 100 ms. Automatic gain control target was set to 3×10^6 for the full MS and to 1×10^5 for ms^2. The intensity threshold was set at 1×10^4.

MS spectrum validation and visualization. Modified peptides were synthesized through the Peptide 2.0 company with a purification level above 95%, and then synthesized peptides were then analyzed in MS using target search mode. For asparagine deamidation, we synthesized the modification as aspartic acid. The spectrum comparison visualization and a similarity score between the original spectrum and the synthesized spectrum were created by R package OrgMassSpecR. To benchmark the similarity score, we used a synthetic modified tryptic peptide set created by ProteomeTools[93]. We compared its behavior under the following three conditions (Supplementary Fig. 3a):

(1) Comparing the differences between two PSM events from the same synthetic peptide in the same MS run.
(2) Comparing a native PSM from HeLa digested standard samples run in our lab to a matching synthetic peptide analyzed in a different instrument. This comparison is the most similar to what we have done in our manuscript.
(3) Decoy assessment by comparing two unmatched randomly selected PSMs.

By comparing the distribution of similarity scores from these three groups, we can see that scores above 0.3 are clearly different from decoy scores. We, therefore, define the similarity cutoff as 0.3.

Thermo Xcalibur version 4.0 Qual Browser was used to manually annotate spectrum. Spectra visualization was created using PDV 1.5.4 software[94] including a, y and b ions and all potential losses.

MSFragger search parameters. Search params were set to default for close search with the following changes: precursor true tolerance was set to 10 ppm and fragment mass tolerance was set to 20 ppm. Search enzyme was set to nonspecific enzyme with cleavage after ARNDCQEGHILKMFPSTWYV. Peptide lengths were set between 8 and 15. Num enzyme termini, 0; clip nTerm M, 1; allow multiple variable mods on residue, 0; max variable mods per mod, 3; max variable mods combinations, 65,000.

Bioinformatics and data analysis. Statistical analyses were performed in Prism 8 software (GraphPad) and R v3.6.1. Heatmaps were drawn with pheatmap 1.0.12 and ComplexHeatmap 2.2.0R package with Euclidean distances for clustering where relevant. Flow cytometry data were analyzed with FlowJo v10 from Becton, Dickinson, and Company. Experimental schematics in Figs. 1a and 4a were generated using BioRender.com.

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

Data availability

MC38 immunopeptidomics data were deposited in the PRIDE archive with ID PXD017448 and standard MaxQuant[95] analysis results. All public data references and accession IDs are listed in the deposited data table in the Supplementary Information.

Code availability

PROMISE is accessible at https://github.com/merbllab/PROMISE.

References

79. Na, S. & Paek, E. Software eyes for protein post-translational modifications. *Mass Spectrom. Rev.* **34**, 133–147 (2015).
80. Wolf-Levy, H. et al. Revealing the cellular degradome by mass spectrometry analysis of proteasome-cleaved peptides. *Nat. Biotechnol.* **36**, 1110–1116 (2018).
81. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40**, W281–W287 (2012).
82. Vacic, V., Iakoucheva, L. M. & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536–1537 (2006).
83. Alam, N. & Schueler-Furman, O. Modeling peptide-protein structure and binding using Monte Carlo sampling approaches: Rosetta FlexPepDock and FlexPepBind. *Methods Mol. Biol.* **1561**, 139–169 (2017).
84. London, N., Lamphear, C. L., Hougland, J. L., Fierke, C. A. & Schueler-Furman, O. Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity. *PLoS Comput. Biol.* **7**, e1002170 (2011).
85. McMurtrey, C. et al. *Toxoplasma gondii* peptide ligands open the gate of the HLA class I binding groove. *elife* **5**, e12556 (2016).
86. Liu, J. et al. Cross-allele cytotoxic T lymphocyte responses against 2009 pandemic H1N1 influenza A virus among HLA-A24 and HLA-A3 supertype-positive individuals. *J. Virol.* **86**, 13281–13294 (2012).
87. Wynn, K. K. et al. Impact of clonal competition for peptide-MHC complexes on the CD8 + T-cell repertoire selection in a persistent viral infection. *Blood* **111**, 4283–4292 (2008).
88. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1369 (2003).
89. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
90. Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* **6**, 132 (2005).
91. Alam, N. et al. High-resolution global peptide–protein docking using fragments-based PIPER—FlexPepDock. *PLoS Comput. Biol.* **13**, e1005905 (2017).
92. Milner, E. et al. The effect of proteasome inhibition on the generation of the human leukocyte antigen (HLA) peptidome. *Mol. Cell. Proteomics* **12**, 1853–1864 (2013).
93. Paul Zolg, D. et al. ProteomeTools: systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Mol. Cell. Proteomics* **17**, 1850–1863 (2018).
94. Li, K., Vaudel, M., Zhang, B., Ren, Y. & Wen, B. PDV: an integrative proteomics data viewer. *Bioinformatics* **35**, 1249–1251 (2019).
95. Cox, J., Michalski, A. & Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass. Spectrom.* **22**, 1373–1380 (2011).

Acknowledgements

We thank the members of the Merbl lab, as well as A. Solomon, O. Bartok, G. Cafri, C. Cohen and M. Besser for scientific discussion and I. Cohen, A. Eisenberg-Lerner, J. DeMartino, C. Putterman, E. Zisman and A. Erez for critical reading of the manuscript. We would like to thank S. Meril for technical help and A. Erez for MC38 cells. Y.M. is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 677748), the I-CORE Program of the Planning and Budgeting Committee, the Israel Science Foundation (grant nos. 1775/12 and 2109/18), the University of Michigan: UM/Israel Research Partnership Weizmann and the Moross Integrated Cancer Center (MICC). This research was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center and by a research grant from Madame Olga Klein—Astrachan. A.I.N. was supported by the US National Institutes of Health grants R01-GM-094231 and U24-CA210967. Y.M. is the incumbent of the Leonard and C. Berall Career Development Chair. A.K. was partially supported by the Israeli Council for Higher Education(CHE) via Weizmann Data Science Research Center and by the research grant from Madam Olga Klein Astrachan.

Author contributions

A.K. and A.J. led the study and performed all computational analyses unless otherwise mentioned. M.P.K. carried out sampling preparation and experiment design, T.S. performed 3D modeling, D.M. and Y.L. consulted regarding MS analyses and algorithm development. E.B. generated the HLA I peptidomics data and A.K., G.C.T., F.V.L. and F.Y. performed software development. Y.S. consulted regarding assay design, O.S.F., A.A., L.E. and A.I.N. supervised the work of respective group members, A.J., A.K. and Y.M. wrote the manuscript and Y.M. guided and supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

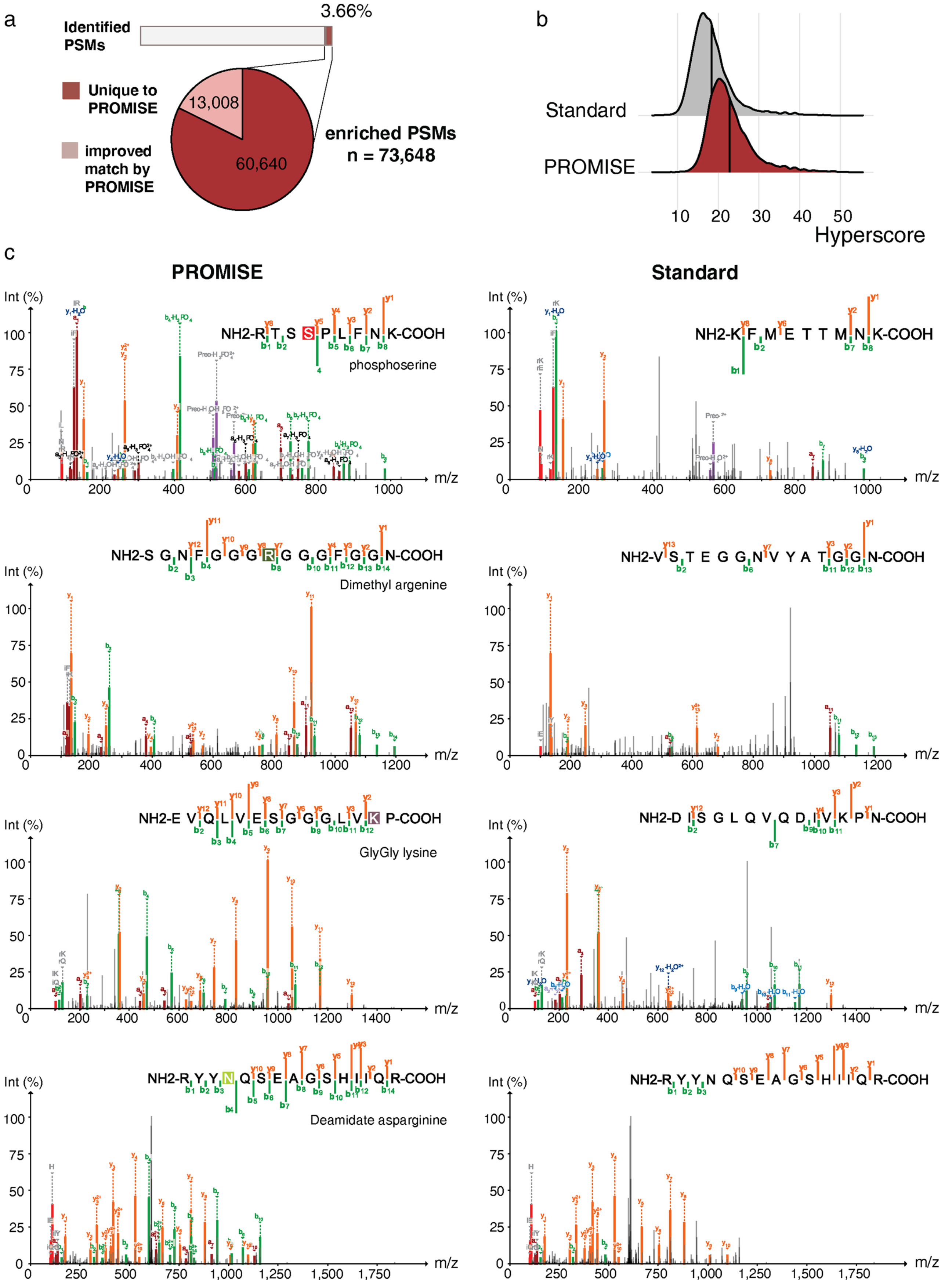
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01464-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01464-2.

Correspondence and requests for materials should be addressed to Yifat Merbl.

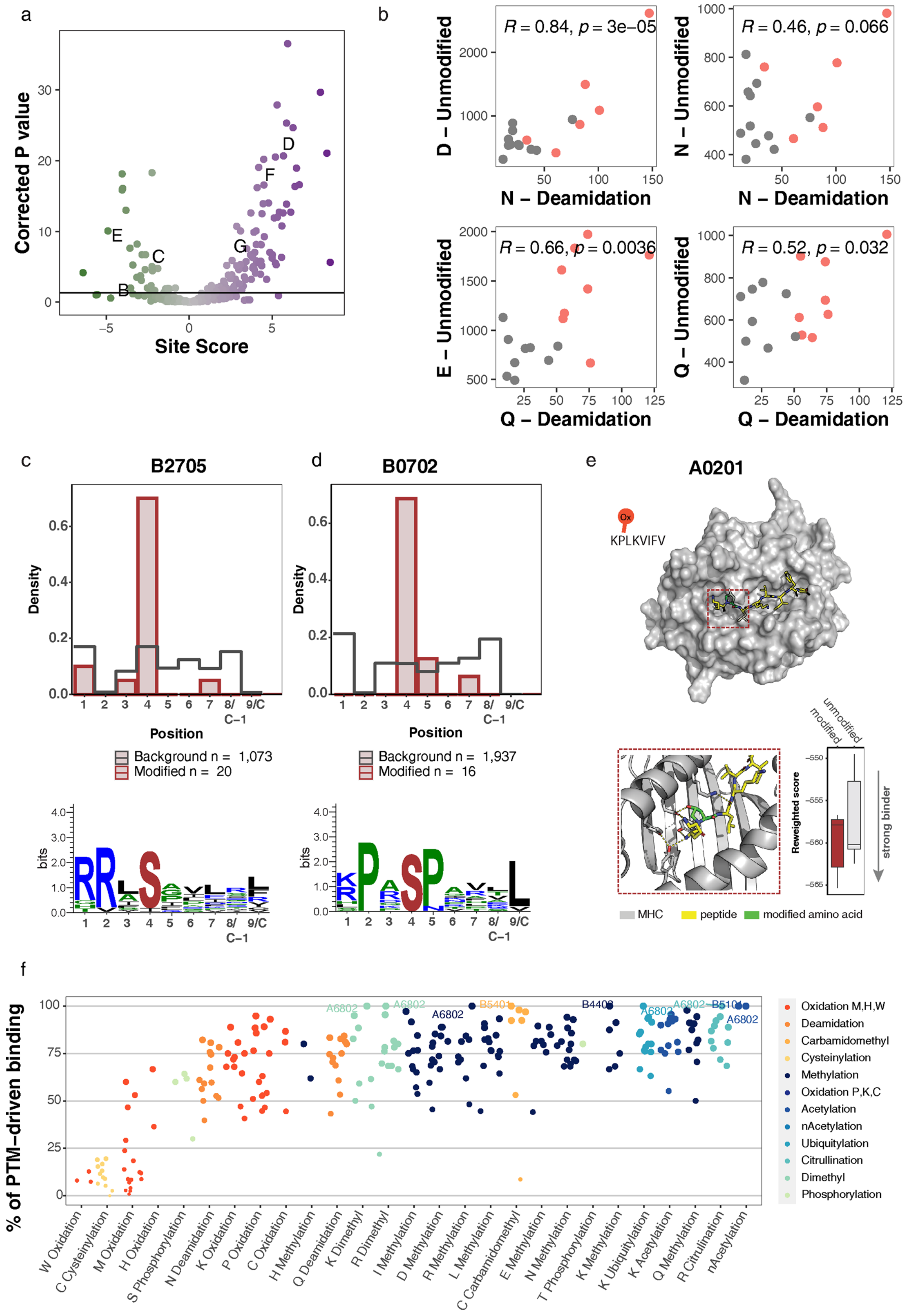
Peer review information *Nature Biotechnology* thanks Mark Cobbold, Michele Mishto and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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



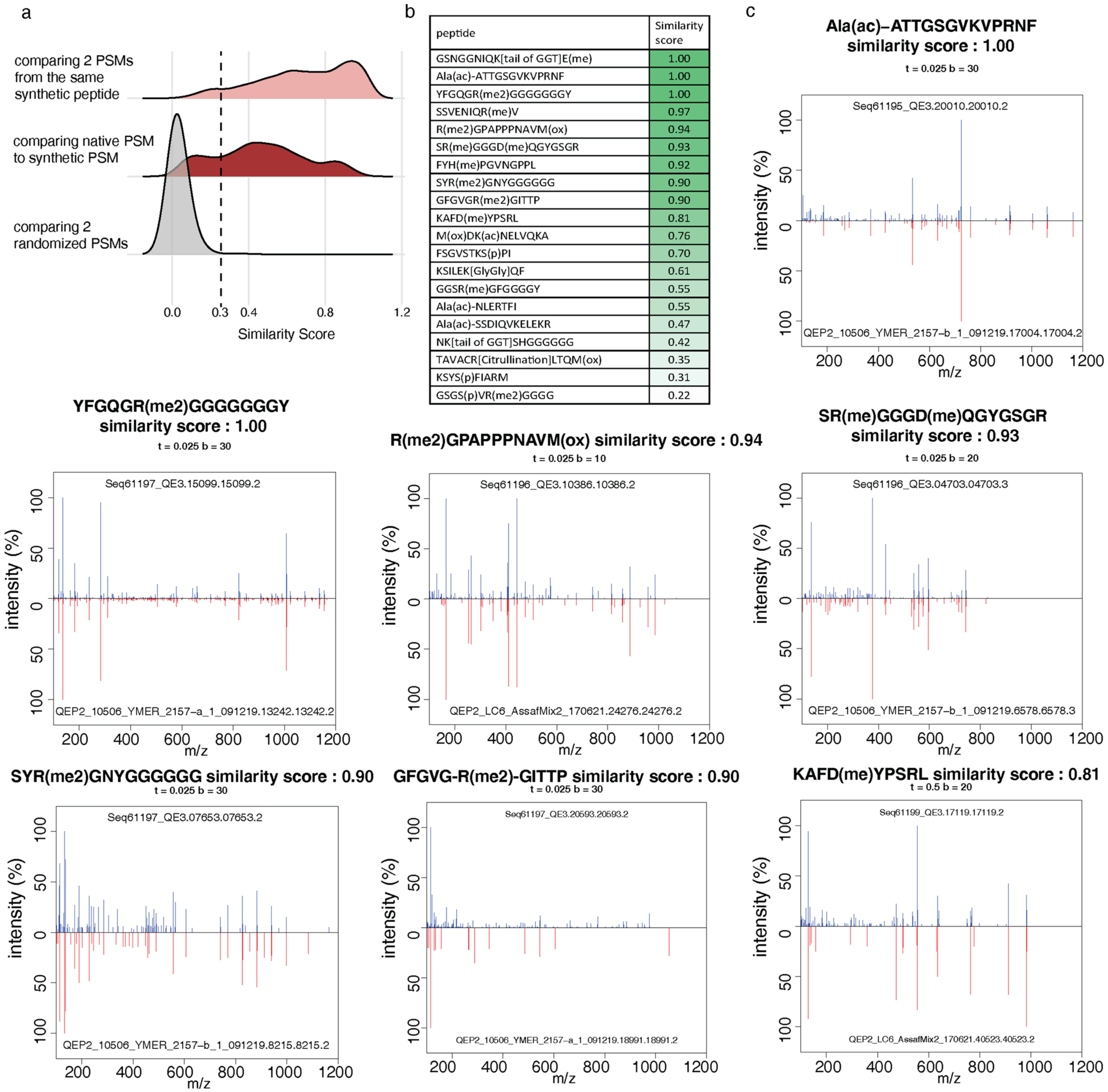
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | PROMISE enrichment in PSM level. (**a**) Percentage of novel PSMs with modifications that were identified through PROMISE (reds), on multiple immunopeptidomics datasets, out of the PSMs identified in standard search (gray). Bottom, pie chart of PSMs enriched by PROMISE search. Out of 73,648 modified PSMs identified in the analysis, 60,640 were IDs unique to PROMISE. (dark red) and 13,008 had improved matching score compared to the standard search (light red). (**b**) Distribution of hyperscores for PSMs which conflicted between Standard (gray) and PROMISE (dark red). Vertical lines mark the average score (**c**) Examples of 4 spectra that received a better peptide match in PROMSIE (left) compared to the standard search (right).



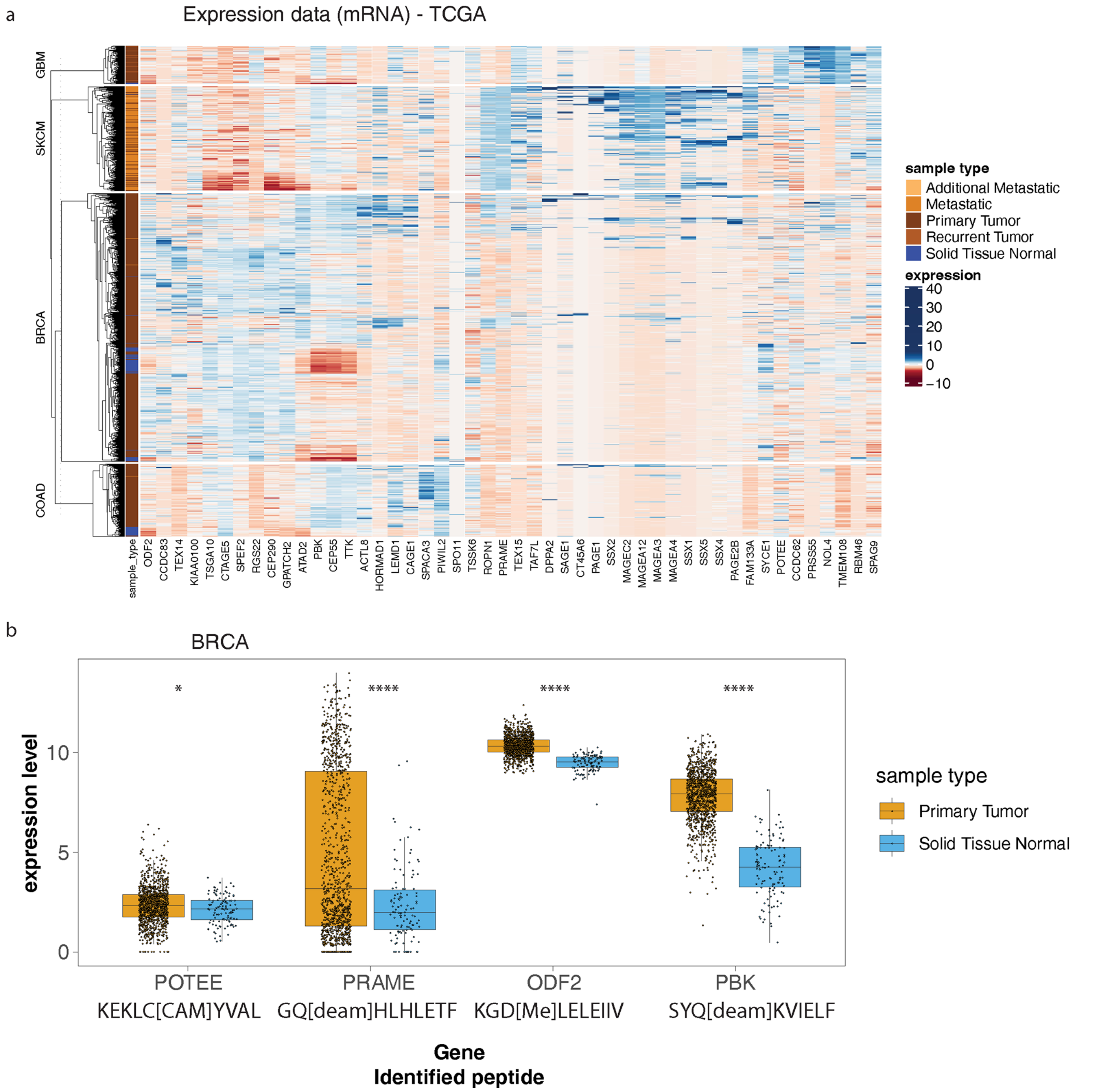
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Monoallelic binding preferences. (**a**) Volcano plot showing the site score plotted against the negative log10 transformed p value from the $\chi2$ test with Benjamini-Hochberg multiple comparison correction. Letters indicate the motifs in Fig. 3 labeled by their panels. (**b**) The counts of peptides containing the indicated modification per haplotype are plotted against the counts of peptides containing unmodified amino acids. The Pearson correlation and p value for the correlation are indicated on each graph. Counts of N - Deamidation are more correlated to its mimic D - unmodified (top) than its source amino acid (N - unmodified). Counts of Q - Deamidation are more correlated to its mimic E - unmodified (bottom) than its source amino acid (Q - unmodified). The haplotypes that have canonical binding motifs that contain an E or D are labeled in pink in the graphs. (**c**, **d**) Reanalysis of monoallelic HLA data recapitulates phosphoserine peptides features as described in Adán Alpízar et al. (**c**) HLA-B*27:05 Phosphoserine position density (top) and the sequence logo (weblogo3) of the peptides carrying phosphoserine in position 4 RRXpS motif (bottom). (**d**) HLA-B*07:02 Phosphoserine position density (top) and K/RPXpS motif (bottom). (**e**) Rosetta FlexPepDock structural model of the interaction between the modified peptide KP(ox) LKVIFV (yellow sticks) and the MHC molecule haplotype HLA-A0201 (gray surface \ cartoon). The modified amino acid (green) creates a more stable interaction with the MHC molecule as compared to the unmodified form. The effect of the modified amino acid is shown in detail in the zoom-in picture. The proline hydroxyl group at position 2 forms a stabilizing hydrogen bond with MHC receptor residue E-87 (shown as dashed yellow line, as well as other hydrogen bonds between peptide and receptor). FlexPepDock reweighted score was calculated for the interaction between the MHC and modified or unmodified peptide (n = 5 simulations, box and whiskers indicates mean and quartiles (**f**) The percentage of peptides in each haplotype with the indicated modification that were not considered binders in NetMHC in their unmodified forms. This indicates that the binding is due to the alteration caused by the PTM. Modifications are sorted by their average percentage of PTM-driven binding and the haplotypes that had the highest percentages are labeled.

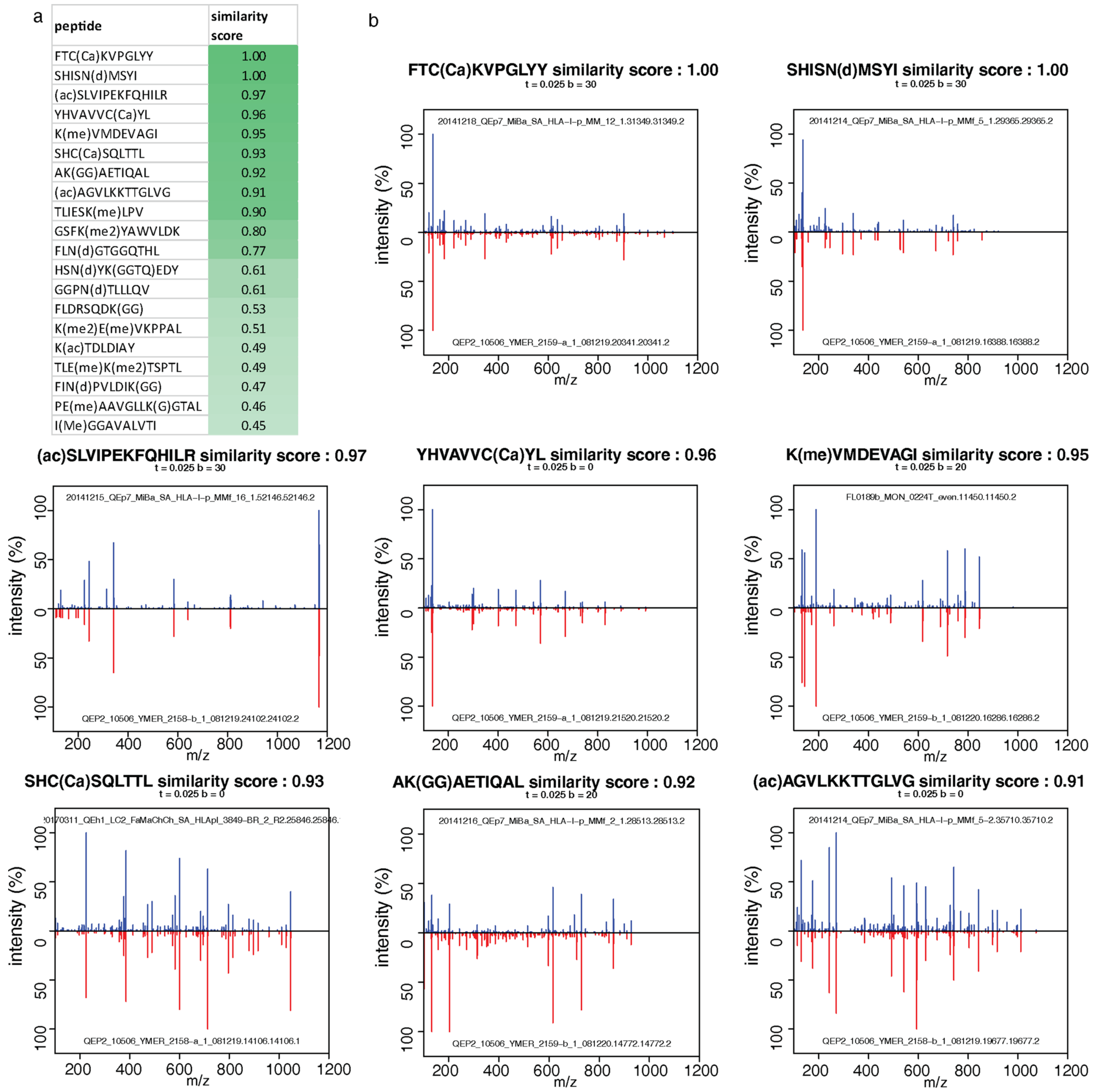


peptide	Similarity score
GSNGGNIQK[tail of GGT]E(me)	1.00
Ala(ac)-ATTGSGVKVPRNF	1.00
YFGQGR(me2)GGGGGGGY	1.00
SSVENIQR(me)V	0.97
R(me2)GPAPPPNAVM(ox)	0.94
SR(me)GGGD(me)QGYGSGR	0.93
FYH(me)PGVNGPPL	0.92
SYR(me2)GNYGGGGGG	0.90
GFGVGR(me2)GITTP	0.90
KAFD(me)YPSRL	0.81
M(ox)DK(ac)NELVQKA	0.76
FSGVSTKS(p)PI	0.70
KSILEK[GlyGly]QF	0.61
GGSR(me)GFGGGGY	0.55
Ala(ac)-NLERTFI	0.55
Ala(ac)-SSDIQVKELEKR	0.47
NK[tail of GGT]SHGGGGGG	0.42
TAVACR[Citrullination]LTQM(ox)	0.35
KSYS(p)FIARM	0.31
GSGS(p)VR(me2)GGGG	0.22

Extended Data Fig. 3 | Mouse spectra validation. (**a**) Similarity score distribution for three types of PSM pairs: (top) two PSMs event taken from the same synthetic peptide in the same sample run (light red, n = 300). We compared the PSM with the highest hyperscore to the PSM with the median hyperscore. (middle) A native PSM taken from HeLa digest standard proteomics compared to a matching synthetic spectrum (dark red, n = 261). (bottom) Similarity score between two randomly chosen PSMs (gray, n = 300). (**b**, **c**) Modified HLA peptides, identified in MC38 cell line and not in healthy mouse colon tissue or reported in the IEDB dataset, were synthesized (Peptide 2.0 Inc) and their spectra were captured using mass spectrometry. For each modified peptide, a similarity score was calculated between the synthetic spectrum and the original spectrum using R package OrgMassSpecR. For a similarity score below 80%, manual annotation was done to validate the spectra. (**b**) summary table (**c**) spectrum comparison visualization and a similarity score are created by R package OrgMassSpecR, synthesized spectrum (red) in a mirror image of the original spectrum in the dataset (blue). In case manual annotation was done, visualization is created using PDV software[94] including a,y,b ions and all potential losses. For the full spectra validation list see Supplementary Information.



Extended Data Fig. 4 | Expression of genes encoding for testis antigens identified in PROMISE. (**a**) TCGA mRNA expression data of testis genes in four different cancer type from which patient sample or cell lines immunopeptidomics data was analyzed by PROMISE: COAD – HCT116; BRCA – PXD009738, HCC1143 and HCC1937; SKCM – PXD004894; GBM – PXD003790. (**b**) The expression of the parent gene from 4 modified HLA I-bound peptide identified in PROMISE is shown for TCGA expression data from BRCA primary tumor and normal tissue (Tumor n = 1097, Normal n = 114, box and whiskers indicate mean and quartiles). The parent testis gene is significantly overexpressed in the tumor tissue vs. the normal (Wilcox p values for tumor vs. adjacent abundance indicated in figures).



peptide	similarity score
FTC(Ca)KVPGLYY	1.00
SHISN(d)MSYI	1.00
(ac)SLVIPEKFQHILR	0.97
YHVAVVC(Ca)YL	0.96
K(me)VMDEVAGI	0.95
SHC(Ca)SQLTTL	0.93
AK(GG)AETIQAL	0.92
(ac)AGVLKKTTGLVG	0.91
TLIESK(me)LPV	0.90
GSFK(me2)YAWVLDK	0.80
FLN(d)GTGGQTHL	0.77
HSN(d)YK(GGTQ)EDY	0.61
GGPN(d)TLLLQV	0.61
FLDRSQDK(GG)	0.53
K(me2)E(me)VKPPAL	0.51
K(ac)TDLDIAY	0.49
TLE(me)K(me2)TSPTL	0.49
FIN(d)PVLDIK(GG)	0.47
PE(me)AAVGLLK(G)GTAL	0.46
I(Me)GGAVALVTI	0.45

a
b
FTC(Ca)KVPGLYY similarity score : 1.00
t = 0.025 b = 30
20141218_QEp7_MiBa_SA_HLA−I−p_MM_12_1.31349.31349.2
QEP2_10506_YMER_2159−a_1_081219.20341.20341.2
SHISN(d)MSYI similarity score : 1.00
t = 0.025 b = 30
20141214_QEp7_MiBa_SA_HLA−I−p_MMf_5_1.29365.29365.2
QEP2_10506_YMER_2159−a_1_081219.16388.16388.2
(ac)SLVIPEKFQHILR similarity score : 0.97
t = 0.025 b = 30
20141215_QEp7_MiBa_SA_HLA−I−p_MMf_16_1.52146.52146.2
QEP2_10506_YMER_2158−b_1_081219.24102.24102.2
YHVAVVC(Ca)YL similarity score : 0.96
t = 0.025 b = 0
QEP2_10506_YMER_2159−a_1_081219.21520.21520.2
K(me)VMDEVAGI similarity score : 0.95
t = 0.025 b = 20
FL0189b_MON_0224T_even.11450.11450.2
QEP2_10506_YMER_2159−b_1_081220.16286.16286.2
SHC(Ca)SQLTTL similarity score : 0.93
t = 0.025 b = 0
20170311_QEh1_LC2_FaMaChCh_SA_HLApI_3849−BR_2_R2.25846.25846.
QEP2_10506_YMER_2158−a_1_081219.14106.14106.1
AK(GG)AETIQAL similarity score : 0.92
t = 0.025 b = 20
20141216_QEp7_MiBa_SA_HLA−I−p_MMf_2_1.28513.28513.2
QEP2_10506_YMER_2159−b_1_081220.14772.14772.2
(ac)AGVLKKTTGLVG similarity score : 0.91
t = 0.025 b = 0
20141214_QEp7_MiBa_SA_HLA−I−p_MMf_5−2.35710.35710.2
QEP2_10506_YMER_2158−b_1_081219.19677.19677.2
intensity (%)
100
50
0
50
100
m/z
200
400
600
800
1000
1200

Extended Data Fig. 5 | Human spectra validation. (**a**, **b**) Modified HLA peptides, that were shared across multiple patients, were synthesized (Peptide 2.0 Inc) and their spectra were captured using mass spectrometry. For each modified peptide, a similarity score was calculated between the synthetic spectrum and the original spectrum using R package OrgMassSpecR. For a similarity score below 80%, manual annotation was done to validate the spectra. (**a**) summary table (**b**) spectrum comparison visualization and a similarity score are created by R package OrgMassSpecR, synthesized spectrum (red) in a mirror image of the original spectrum in the dataset (blue). In case manual annotation was done, visualization is created using PDV software[94] including a,y,b ions and all potential losses. For the full spectra validation list see Supplementary Information.

nature research

Corresponding author(s): Yifat Merbl

Last updated by author(s): Aug 9, 2022

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 our Editorial Policies 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	Commercial Software: Xcalibur version 4.0
Data analysis	The pipeline code is accessible at https://github.com/merbllab/PROMISE Data analysis and visualization was done using: R v 3.6.1, BioRender, Python 2.7, Python 3.7, Philosopher 4.0 MSFragger-3.1.1, CRUX 3.1, MaxQuant 1.6.0.16 , Rosetta macromolecular modeling suite v2019.14

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 and 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

Data accession ID , Paper , DOI, Description , Figures
PXD004894, Bassani-Sternberg, M. et al., 10.1038/ncomms13404, Melanoma tissue, 1E,1D,1F,1G,1H,1I,1J,S1,2B,2C,2D,2E,2F,2G,4D,4E,4F,4G,S4,S5
PXD000394, Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M, 10.1074/mcp.M114.042812, Cancer cell lines, 1B,1D,1E,1F,1G,1H,1I,1J,S1,2A,2B,2C,2D,2E,2F,2G,4D,4E,4F,4G,S4,S5
PXD006939, Chong, C. et al., 10.1074/mcp.TIR117.000383, TILs and cell lines, 1D,1E,1F,1G,1H,1I,1J,S1,2B,2C,2D,2E,2F,2G,4D,4E,4F,4G,S4,S5
PXD003790 Shraibman, B., Kadosh, D. M., Barnea, E. & Admon, 10.1074/mcp.M116.060350, Glioblastoma cell lines,

1D,1E,1F,1G,1H,1I,1J,S1,2B,2C,2D,2E,2F,2G,4D,4E,4F,4G,S4,S5
PXD009738, Ternette, N. et al., 10.1002/pmic.201700465, Triple Negative Breast Cancer, 1D,1E,1F,1G,1H,1I,1J,S1,2B,2C,2D,2E,2F,2G,4D,4E,4F,4G,S4,S5,5A,5B
MSV000080527, Abelin, J. G. et al., 10.1016/j.immuni.2017.02.007, 16 monoallelic cell lines, 3,S2
MSV000084172, Sarkizova, S. et al., 10.1038/s41587-019-0322-9, 95 monoallelic cell lines (only HLA B2705 and B0702 were used in this study), 3,S2
IEDB, , Update from 2/2020, 2B,2D,2E,2F
PXD008733, Schuster, H. et al., 10.1038/sdata.2018.157, , 4B
PXD017448, , , MC38 cell line (this study), 4A,4B,4C,S3,S5
CPTAC Breast Cancer S039, Mertins, P. et al., , Breast cancer proteomics , 5B
TCGA, http://cancergenome.nih.gov/, BRCA, S4
PXD009449, Daniel Paul Zolg et al., 10.1074/mcp.TIR118.000783, Synthetic modified peptides, S3A, Supplementary pipeline description

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	Sample size for each experiment is stated in figure captions and data descriptions. Sample sizes were not predetermined.
Data exclusions	no exclusions
Replication	All new experiments were carried out in at least triplicate in an independent manner with a high degree of correlation.
Randomization	Experimental groups were divided based on the treatment conditions in the experiments. No allocation was done.
Blinding	We did not utilize blinding in our experimental design however, data collection and analysis was not performed manually or in a subjective manner.

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 and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

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

Antibodies

Antibodies used	All antibodies are listed in supplementary information in the Reagents table
Validation	antibodies were validated by manufacturers

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	MC-38 cells were provided by Ayelet Erez (Weizmann Institute)
Authentication	MC-38 were not authenticated;

Mycoplasma contamination	tested and results were negative;
Commonly misidentified lines (See ICLAC register)	n/a