

Proteogenomic Analysis Unveils the HLA Class I-Presented Immunopeptidome in Melanoma and EGFR-Mutant Lung Adenocarcinoma

Authors

Yue A. Qi, Tapan K. Maity, Constance M. Cultraro, Vikram Misra, Xu Zhang, Catherine Ade, Shaojian Gao, David Milewski, Khoa D. Nguyen, Mohammad H. Ebrahimabadi, Ken-ichi Hanada, Javed Khan, Cenk Sahinalp, James C. Yang, and Udayan Guha

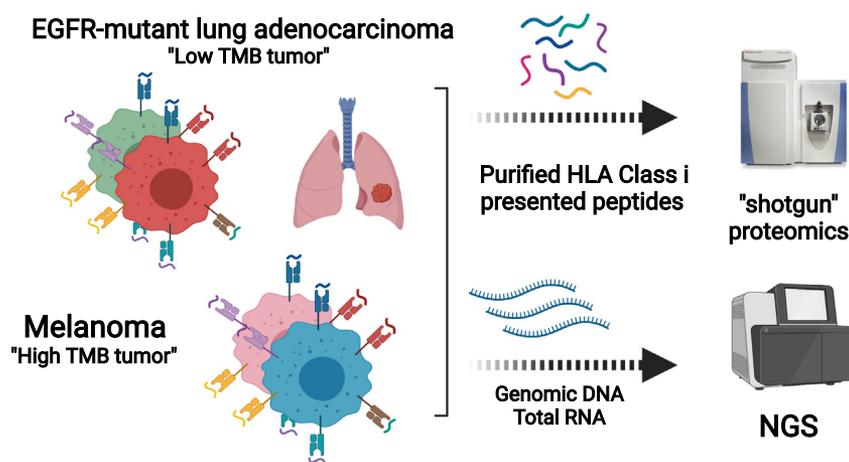
Correspondence

andy.qi@nih.gov; udayan.guha@nih.gov

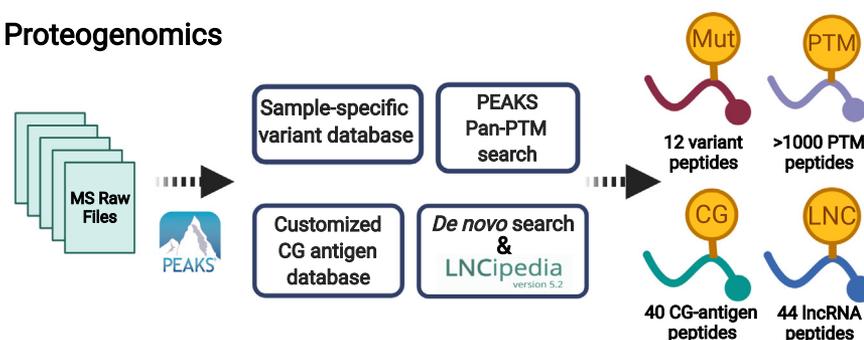
In Brief

Cancer immunotherapy is ineffective in low TMB EGFR-mutant lung adenocarcinoma. Qi *et al.* performed a comprehensive proteogenomic profiling of HLA class I-presented immunopeptides in high TMB melanoma and low TMB EGFR-mutant lung cancer. Similar numbers of immunopeptides were identified from both. Variant, CG antigen, PTM, and lncRNA-derived peptides were identified. A novel strategy to identify lncRNA-derived peptides was developed. The direct identification of class I-presented immunopeptides will potentially accelerate precision immunotherapy for low TMB tumors.

Graphical Abstract



Proteogenomics



Highlights

- Proteogenomics identified ~35,000 class I-presented peptides.
- CG antigen and PTM peptides identified in melanoma and lung cancer.
- *De novo* search identified variant and lncRNA-derived peptides.
- A new strategy to identify class I-presented lncRNA-derived peptides developed.



Proteogenomic Analysis Unveils the HLA Class I-Presented Immunopeptidome in Melanoma and EGFR-Mutant Lung Adenocarcinoma

Yue A. Qi^{1,*}, Tapan K. Maity¹, Constance M. Cultraro¹, Vikram Misra¹, Xu Zhang¹, Catherine Ade², Shaojian Gao¹, David Milewski³, Khoa D. Nguyen¹, Mohammad H. Ebrahimabadi^{4,5}, Ken-ichi Hanada², Javed Khan³, Cenk Sahinalp⁴, James C. Yang², and Udayan Guha^{1,6,*}

Immune checkpoint inhibitors and adoptive lymphocyte transfer-based therapies have shown great therapeutic potential in cancers with high tumor mutational burden (TMB), such as melanoma, but not in cancers with low TMB, such as mutant epidermal growth factor receptor (EGFR)-driven lung adenocarcinoma. Precision immunotherapy is an unmet need for most cancers, particularly for cancers that respond inadequately to immune checkpoint inhibitors. Here, we employed large-scale MS-based proteogenomic profiling to identify potential immunogenic human leukocyte antigen (HLA) class I-presented peptides in melanoma and EGFR-mutant lung adenocarcinoma. Similar numbers of peptides were identified from both tumor types. Cell line and patient-specific databases (DBs) were constructed using variants identified from whole-exome sequencing. A *de novo* search algorithm was used to interrogate the HLA class I immunopeptidome MS data. We identified 12 variant peptides and several classes of tumor-associated antigen-derived peptides. We constructed a cancer germ line (CG) antigen DB with 285 antigens. This allowed us to identify 40 class I-presented CG antigen-derived peptides. The class I immunopeptidome comprised more than 1000 post-translationally modified (PTM) peptides representing 58 different PTMs, underscoring the critical role PTMs may play in HLA binding. Finally, leveraging *de novo* search algorithm and an annotated long noncoding RNA (lncRNA) DB, we developed a novel lncRNA-encoded peptide discovery pipeline to identify 44 lncRNA-derived peptides that are presented by class I. We validated tandem MS spectra of select variant, CG antigen, and lncRNA-derived peptides using synthetic peptides and performed HLA class I-binding assays to demonstrate binding to class I proteins. In summary, we provide direct evidence of HLA

class I presentation of a large number of variant and tumor-associated peptides in both low and high TMB cancer. These results can potentially be useful for precision immunotherapies, such as vaccine or adoptive cell therapies in melanoma and EGFR-mutant lung cancers.

Cancer immunotherapy has become an essential component of therapy for diverse cancers. The treatment outcome and patient survival rate are positively correlated with their tumor mutational burden (TMB) (1–3). Epidermal growth factor receptor (EGFR)-mutant lung adenocarcinoma occurs predominantly in never smokers or oligosmokers and exhibits a relatively low TMB (4). Immunotherapy has been less successful in EGFR-mutant lung cancer, in part, because of its low TMB (5, 6). In contrast, melanoma, a cancer with high TMB because of UV exposure, responds well to current immune checkpoint blockade immunotherapy (7). Consequently, the use of immunotherapy to treat low TMB cancers has been an unmet need. While classic immune checkpoint inhibition activates the natural immune response against cancer, more recently, there has been some success with adoptive T-cell therapy (ACT) that creates a repertoire against “nonself” neoantigens or tumor-associated antigens (e.g., cancer germ line [CG] antigens) (8, 9). Thus, the identification of cancer-specific or cancer-associated antigen-derived peptides is important for the development of immunotherapeutic strategies for the treatment of low TMB tumors.

Recently, MS-based proteomics has become a powerful approach for large-scale profiling of the human leukocyte antigen (HLA) class I-presented peptidome (10, 11). Unlike traditional HLA-epitope prediction algorithms, MS sequencing

From the ¹Thoracic and GI Malignancies Branch, ²Surgery Branch, ³Genetics Branch, and ⁴Cancer Data Science Laboratory, Center for Cancer Research, NCI, NIH, Bethesda, Maryland, USA; ⁵Department of Computer Science, Indiana University, Bloomington, Indiana, USA; and ⁶Bristol-Myers Squibb, Lawrenceville, New Jersey, USA

* For correspondence: Yue A. Qi, andy.qi@nih.gov; Udayan Guha, udayan.guha@nih.gov.

Present address for Yue A. Qi: Center for Alzheimer's and Related Dementias, National Institute on Aging, NIH, Bethesda, MD, USA.

Present address for Udayan Guha: Bristol Myers Squibb, Lawrenceville, NJ 08901, USA and Thoracic and GI Malignancies Branch, Center for Cancer Research, NCI, NIH, Bethesda, MD, USA.

provides direct experimental evidence of the presented peptides, and it allows for the relative quantification of cell surface peptide presentation. This high-throughput method can be used to profile thousands of *in vivo* HLA-associated immunopeptides (12, 13). When combined with next-generation sequencing (NGS), to reveal somatic mutations, this approach is capable of detecting mutant peptides (14) and noncanonical peptides derived from noncoding regions (15). However, despite the fact that MS-based cancer antigen discovery has been widely employed for directly assessing antigen presentation, many previous studies only focused on high TMB tumors. Here, our goal was to develop a comprehensive proteogenomic platform to identify potentially targetable class I-presented peptides in both melanoma and lung adenocarcinoma. We hypothesized that both low TMB-associated EGFR-mutant lung tumors and high TMB-associated melanoma present a repertoire of tumor-specific or tumor-associated antigen-derived peptides on HLA class I.

To develop this method, we enriched the cell surface-presented HLA class I-bound peptides in two primary melanoma cell lines, two EGFR-mutant lung adenocarcinoma cell lines, and one primary tumor from an EGFR-mutant patient who had undergone EGFR tyrosine kinase inhibitor therapy and identified those peptides using high-resolution tandem MS (MS/MS). Our datasets contain five major categories of class I-presented peptides derived from (1) common driver oncogenes; (2) mutated peptides derived from variants, herein, referred to as variant peptides; (3) CG antigen-derived peptides; (4) post-translationally modified (PTM) peptides; and (5) long noncoding RNA (lncRNA)-derived peptides. We also validated select class I-presented peptides utilizing synthetic standard peptides and an HLA stability-binding assay.

EXPERIMENTAL PROCEDURES

Human subject and cell lines

Lung tumor specimen was obtained at rapid autopsy from left lower lobe of the lung of osimertinib-treated patient NCI-RA007, a 70-year-old male, with a primary EGFR^{L858R} mutation. The tumor is a mixture of tumor cells and surrounding immune cells and fibroblasts that constitute the tumor microenvironment. The tumor specimen was snap frozen in liquid nitrogen upon collection. The genomic alterations in the tumor of this patient were described before (16). In accordance with the Declaration of Helsinki, the patient was consented with National Cancer Institute (NCI) Institutional Review Board-approved protocol 13-C-0131 (NCT01851395) entitled "A pilot study of inpatient hospice with procurement of tissue on expiration in thoracic malignancies." The patient was offered hospice treatment with life expectancy less than 3 months at Center for Cancer Research (CCR), NCI. The rapid autopsies were initiated within 3 h upon patient death that resulted in good quality genomics and proteomics data from tumor tissues as reported before (17). The two primary melanoma cell lines, NCI-3784Mel and NCI-3795Mel, were obtained from patients with melanoma treated at the NCI, under the protocol, NCT00068003. NCI-3784Mel has been reported previously (18). The melanoma cells were cultured in high glucose Dulbecco's modified Eagle's medium supplemented with 20% fetal bovine serum (FBS). EGFR-mutant lung

adenocarcinoma cell line H1975, was purchased from American Type Culture Collection (ATCC) and PC9 cell line was obtained from the Varmus Laboratory (MSKCC). The lung adenocarcinoma cells were maintained in RPMI1640 cell growth medium supplemented with 10% FBS. The T2 (174 × CEM.T2) cell line, purchased from ATCC, was maintained in ATCC-formulated Iscove's modified Dulbecco's medium supplemented with 10% FBS.

HLA class I-presented peptide enrichment and purification

For HLA class I-presented peptide enrichment of the melanoma and lung adenocarcinoma cell lines, 2.0×10^8 cells per biological replicate were harvested in 4 ml ice-cold lysis buffer (20 mM Tris-HCl, pH = 8.5, 100 mM NaCl, 1 mM EDTA, and 1% Triton X-100 supplemented with Halt 1:100 protease inhibitor cocktail [catalog no. 78430; Thermo Scientific]). After 30 min on ice, lysates were subjected to needle sonication for 30 s. For the human tumor tissue from rapid autopsy, 30 mg snap-frozen lung tumor tissue per replicate was homogenized in 4 ml ice-cold lysis buffer for 30 s at 4 °C using the Qiagen TissueLyser II. Cell/tissue lysates were centrifuged at 20,000g for 2 h at 4 °C, and the supernatant was used in subsequent experiments. HLA-peptide complexes were isolated by immunoprecipitating with 0.5 mg pan-HLA class I antibody clone W6/32 (BioXcell) precoupled with 200 μ l slurry of protein A/G PLUS agarose resin (Santa Cruz Biotechnology) overnight at 4 °C with constant rotation. Agarose beads were then washed three times with ice-cold lysis buffer (without Triton and protease inhibitors), followed by two washes in ice-cold 20 mM Tris-HCl (pH = 8.5), and then one wash in ice-cold HPLC grade water. Complexes were eluted four times with 0.15% TFA in water at room temperature and combined. To purify immunopeptides, 50 mg C₁₈ desalting columns (Sigma Millipore) were activated by two washes with 100% acetonitrile (ACN) and equilibrated by two washes with 0.1% TFA in water. HLA-peptide complexes were loaded on the preconditioned 50 mg C₁₈ columns followed by three 0.1% TFA in water washes. HLA peptides were eluted with 40% ACN in 0.1% TFA. Purified peptides were lyophilized at -80 °C for 2 h followed by desalting step using pierce C₁₈ Spin tips (catalog no. 84850; Thermo Scientific). The spin tips were activated using 20 μ l of 80% ACN in 0.1% TFA and equilibrated using 20 μ l of 0.1% TFA by centrifuging at 1000g for 1 min. We loaded the acidified HLA peptides (reconstituted in 40 μ l of 0.1% TFA) on to prepared spin tips, washed the peptides twice with 0.1% TFA, and eluted the peptides in 20 μ l of 80% ACN in 0.1% TFA. The resulting peptides were lyophilized and reconstituted in 0.1% TFA and 2% ACN loading buffer for LC-MS/MS analysis.

Sample preparation of whole-cell proteomic profiling

The sample preparation protocol of H1975 whole-cell proteome profiling has been described previously (19). Briefly, three of 10-cm cell culture plates of H1975 (n = 3) were harvested using the same lysis buffer as for HLA peptide purification. About 1 mg protein was reduced by 5 mM Tris (2-carboxyethyl) phosphine hydrochloride at 50 °C for 30 min and was alkylated by 10 mM iodoacetamide in dark for 30 min followed by MS grade trypsin/lysC digestion (Promega) for 16 h at 37 °C. The resulting tryptic peptides were fractionated by an off-line high-pH (pH = 8.5) reverse-phase LC into a 96-well deep plate and were further pooled into 12 fractions. The LC mobile phase A was 10 mM triethylammonium bicarbonate in water, and the LC mobile phase B was 10 mM triethylammonium bicarbonate in ACN. The lyophilized tryptic peptides were reconstituted in phase A and separated with a 30 min LC gradient (10–35% solvent B) on an XBridge C18, 100 × 2.1 mm analytical column (Waters) using a flow rate at 0.25 ml/min. A total of 96 fractions were collected and acidified to 0.5% TFA and pooled into 12 final fractions immediately. The fractionated peptides were lyophilized and desalted by Pierce C₁₈ Spin

tips according to the described protocol for HLA peptides desalting. Purified peptides were reconstituted in 0.1% TFA and 2% ACN buffer, and peptide concentration was evaluated by NanoDrop at an absorbance at 215 nm. About 500 ng per fraction was subjected for LC-MS/MS analysis.

MS/MS analyses

For the HLA immunopeptidome profiling, the purified and desalted HLA peptides were loaded to a 2 cm nano Acclaim trap column (catalog no. 164535) followed by separation on a 25 cm EASY-spray reverse phase column (catalog no. ES802A) for 90 min effective gradient with 4 to 35% 0.1% formic acid in ACN on an Ultimate 3000 Nano LC instrument (Thermo Scientific). The separated peptides were analyzed on an Orbitrap Q-Exactive HF mass spectrometer (Thermo Scientific) with discovery mode for data acquisition. The MS1 full scan (375–1650 *m/z*) was set to 120,000 resolution, and top 15 most abundant peptides per cycle were subsequently fragmented by high-energy collision dissociation. To identify only nontryptic-digested neutral HLA class I-presented peptides (e.g., without either lysine or arginine), typically containing 8 to 14 amino acid residues, we also included single charged ions resulting in charge state 1 to 4 for MS2 peak picking. The MS2 sequencing scans acquired the peptide fragments at 30,000 resolution and 200 ms maximum injection time window. Since enriched HLA peptidome is commonly less complex than whole-cell proteome, we used dynamic exclusion at 20 s to collect more MS1 data points, allowing better peak area-based quantitation. For the whole-cell proteome, H1975 tryptic peptides from each fraction were separated by the same LC method described previously for the HLA peptidome profiling. The MS1 analysis was set at 60,000 resolution, and top 20 most abundant peaks were selected for MS/MS analysis in which we used 15,000 resolution and 32 ms maximum injection window. We included charge state 2 to 6 for MS2 fragmentation, and dynamic exclusion was set at 30 s. In addition, we validated select variant peptides, CG antigen and lncRNA-derived peptides. Select variant peptides and lncRNA-derived peptides were *in vitro* synthesized by GenScript. Select melanoma CG antigen-derived peptides were *in vitro* synthesized by the Peptide Synthesis and Antigen Discovery Core, Surgery Branch, NCI. Synthetic peptides were serially diluted to 1 pmol/ μ l in 0.1% TFA and 2% ACN in water. The diluted synthetic peptides were pooled, and 5 μ l of the pool was subjected to MS/MS analysis using the same LC-MS/MS instrument method for HLA peptidome profiling. To correct potential retention time (RT) shifting using different batches of the nanoLC columns, Pierce Peptide Retention Time Calibration Mixture (Thermo Fisher; catalog no. 88321) containing 15 heavy-labeled peptides (1 pmol/ μ l) was used for the RT correction.

Database and de novo search of MS raw files

Patient- and cell line-specific protein sequence databases (DBs) were first generated. The aligned whole-exome sequencing (WES) BAM files from patient blood (germ line) or tumor tissue/cell lines were used to retrieve the variant call format (VCF) files using HaplotypeCaller (Broad Institute) (20), and the intermediate VCF files were further annotated by SnpEff (Microsoft Genomics) (21), which filtered out only non-synonymous variants on exome regions including single nucleotide variants (SNVs) and small insertions and deletions (INDELs). Similarly, BBduk (v35.2) (DOE Joint Genome Institute) was used to remove adapter sequences and low-quality reads from paired-end FASTQ files, which were then used as input for STAR-Fusion (version 1.6.0) (<https://github.com/STAR-Fusion/STAR-Fusion/>) (22). The final VCF files were *in silico* translated to sample-specific protein sequence libraries (FASTA files) using QUILTS (<http://openslice.fenyolab.org/cgi-bin/pyquits.cgi.pl>) (23), which were merged with refseq hg38 converted human proteome DB (v20130727). Herein, we generated specific FASTA file for each

patient and cell line. To avoid using hypercomplex DB, which may increase the false-positive rate, DB search for variant peptide discovery was carried out for individual tumor or cell line separately. For the identification of nonvariant peptides (i.e., CG antigen and PTMs), DB search was conducted with all MS raw files using UniProt human proteome DB (v20170207). The DB search of MS raw files was carried out by PEAKS studio (version 8.5) (24) (Bioinformatics Solutions) using the patient-/cell line-specific DBs or standard human proteome DB described previously. In the PEAKS search engine, no enzyme digestion was selected because HLA peptides are natural peptides without artificial enzyme digestion. Importantly, the unique PEAKS built-in functions, pan-PTMs including 650 different variable modifications and *de novo* search algorithm, were used. The precursor mass tolerance was set to 15 ppm, and the fragment ion tolerance was set to 0.05 Da. For the DB search, the false discovery rate (FDR) of peptide identification, estimated by decoy-fusion DB, was chosen at 0.01. For the *de novo* search, the average local confidence (ALC%) score of each peptide was chosen to be >50%. MS1 peak area-based label-free quantitation method was used for peptide quantification. The direct output peptide intensity from PEAKS was log₂ transformed for further statistical analysis. For the whole-cell proteome, 12 fractions from each sample were pooled, and DB was searched by MaxQuant (version 1.5.7.4) (Max Planck Institute) using UniProt human proteome DB (v20170207) that contains 70,948 entries including isoforms. The mass tolerance for precursor ions was set to 4.5 ppm, and mass tolerance for fragment ions was set to 20 ppm. Trypsin and lysC were selected as digestion enzymes, maximum missed cleavage was set to 4, and methionine oxidation and N-terminal acetylation were selected as variable modifications. Both FDRs at peptide and protein levels were set to 0.01. The proteins were quantified by label-free quantification method.

HLA class I typing

First, HLA typing (six digits) of the patient donors of the two primary melanoma cell lines was performed using sequence-specific primer (SSP) and Sanger sequencing technologies by the Department of Transfusion Medicine of Clinical Center at National Institutes of Health (NIH). In addition, our laboratory conducted four-digit HLA class I calling from the WES data of these two melanoma cells using Seq2HLA package (25) and demonstrated consistent results with the SSP-Seq (supplemental Fig. S1G). Therefore, we performed Seq2HLA-based informatic HLA typing from the WES results of all samples.

T-cell epitopes/HLA-binding prediction, motif analysis, and hydrophobicity index prediction

All peptides except PTM peptides in each cell/tissue sample were used for T-cell epitope and HLA-binding prediction using Immune Epitope Database and Analysis Resource (IEDB) (26) and/or NetMHCpan-4.0 (<http://www.cbs.dtu.dk/services/NetMHCpan-4.0/>) (27). To generate reviewed monoallelic epitope datasheets using IEDB (supplemental Fig. S1F), we applied stringent filters for those epitopes: (1) “*Homo sapiens*” as antigen organism; (2) “Humans” as host; and (3) specific HLA class I subtype (e.g., HLA-A*02). The resulting known epitopes from IEDB DB were subject to motif analysis using iceLogo (<https://iomics.ugent.be/icelogsolver/>) (28), where we used *H. sapiens* SwissProt composition as the reference set, and start position was set to 1. The percent difference in frequency of the amino acid at given locations was chosen as the readout of motif analysis. To compare the motifs from those reported epitopes and MS-detected HLA immunopeptides, we performed similar motif analyses of our peptide datasets. The hydrophobicity index (HI) prediction of select *de novo* peptides was performed in SSRCalc (version Q) where 100 Å C18 column, 0.1% formic acid (2015), and HI_(Best) were selected (29).

Whole exome and total RNA-Seq

WES and total RNA-Seq were performed as described previously (17). Briefly, the genomic DNA and total RNA of the cell lines and tumor were extracted and sent to the NGS core facility at NCI Frederick National Laboratory. The samples were sequenced as 2 × 126 nt paired end reads with Illumina HiSeq2500 sequencers with >100 million reads per sample. The raw FASTQ files were aligned to hg38 by TopHat (version 2.0.13) (<https://github.com/infp/htophat>) (30); the aligned BAM files were used for downstream variant calling. For the samples with a corresponding germ line specimen, Strelka (version 1.0.10) was used for somatic variant calling (31). The total RNA-Seq of H1975 (n = 3) was normalized and quantified using DESeq2 (version 1.30.0) (32).

Generation of CG antigen DB

We compiled a CG antigen DB using existing antigens reported in CT DB (<http://www.cta.lncc.br/>), Human Protein Atlas (<https://www.proteinatlas.org/>), and Cancer Antigenic Peptide DB (<https://caped.icp.ucl.ac.be/>).

Immunoblotting of HLA class I antigens

One million cells from each cell line were lysed in ice-cold modified radioimmunoprecipitation buffer for 30 min. The cell lysate was spun down at 20,000g for 15 min at 4 °C, and supernatant protein concentration was determined by the bicinchoninic acid protein assay. About 10 µg of protein from each cell line underwent SDS-PAGE. Subsequently, separated proteins were transferred to polyvinylidene fluoride membrane and incubated with primary anti-HLA class I mouse horseradish peroxidase monoclonal antibody at 1:5000 dilution (EMR8-5; Funakoshi) overnight at 4 °C, and then briefly incubated with SuperSignal horseradish peroxidase substrates (Thermo Scientific) before imaging. The membranes were exposed for 5 s, and images were acquired by Odyssey Fc imager (LI-COR Biosciences).

Flow cytometry analyses for HLA class I expression

One million cells were collected in 100 µl fluorescence-activated cell sorting buffer (PBS + 5% FBS). After 30 min of blocking, half of the cells were incubated with FITC anti-HLA class I, W6/32 (catalog no. 311404; BioLegend) for 30 min at 4 °C, and washed twice with fluorescence-activated cell sorting buffer. Flow cytometry was carried out on CytoFLEX platform (Beckman), and 20,000 events were collected per sample. The postanalyses and statistics were conducted by FlowJo (version 10.6.2) (BD Biosciences).

HLA-binding affinity T2 cell assay

The antigen peptide transporter (transporter associated with antigen processing [TAP])–deficient HLA-A2 only expressing T2 cells presents lower affinity peptides because of TAP1 deficiency and allow for an efficient exchange of high-affinity peptides. We suspended T2 cells at 1.0 million/ml in growth medium and plated them in 6-well tissue culture plates (2 ml/well). The synthetic peptides for validation and NY-ESO-1 peptide (positive control) were reconstituted in dimethyl sulfoxide (DMSO) and diluted to a final concentration of 10 µM with growth medium and incubated with T2 cells at 37 °C with 5% CO₂ for 12 h. Each testing peptide and positive control was performed in triplicate and negative control (DMSO) in duplicate. The cells were incubated with FITC anti-HLA class I, W6/32 (catalog no. 311404; BioLegend) for 30 min at 4 °C followed by two washes. Flow cytometry was performed to detect the cell surface total HLA class I protein expression. Data analysis was conducted by FlowJo (version 10.6.2), and geometric mean fluorescence intensity was used for quantification.

Identification of HLA-presented lncRNA-derived peptides

To identify lncRNA-coded peptides, we generated an lncRNA-translated protein sequence DB, which was *in silico* 6 frame translated from a high-confident lncRNA DB, LNCipedia (www.lncipedia.org), containing 107,039 lncRNA transcripts from 49,372 annotated distinct lncRNAs (version 5.2) (33). This resulted in 642,234 lncRNA-encoded protein sequence entries. The *de novo*–only identified 8- to 14-mer peptides that did not match the normal Uniprot human DB (70,948 entries including isoforms) from all five samples were queried against this lncRNA DB-derived six-frame translated protein sequence DB. Furthermore, we only kept the peptides predicted to be an HLA binder (%Rank <2.0) to at least one HLA allele in respective samples. To ensure those matched lncRNA-derived peptides were truly expressed at RNA level, we generated intersect BED files between each sample's total transcriptome and LNCipedia. Briefly, for the genomic data processing, quality control was done on total RNA-Seq FASTQ files by first assessing read quality using FASTQC (version 0.11.8) followed by removing adapter sequences and low-quality reads using BBduk, part of the BBTools package (version 38.42). The resulting FASTQ files were aligned against the RefSeq human genome version hg38 using STAR (version 1.3.4) (34) and sorted using the Samtools mappings sorter (version 1.1.1) (35). Duplicate reads were then removed using Picard MarkDuplicates (version 2.1.1). These alignments, sorting, and deduplication steps were run on the DNA-Nexus platform. The resulting BAM file was converted to BED format using the bamtoBED tool from the BEDTools suite (version 2.29.0) (36). The BEDTools intersect tool was used to find the intersection between the tumor or cell-line BED file and the LNCipedia BED file (version 5.2). The resulting BED files contained matched expressed lncRNA genes in the sample transcriptome. Because of the possible misannotated lncRNAs in LNCipedia DB, we further visualized each individual matched lncRNA using the intersect BED files in Integrative Genomics Viewer (IGV; version 2.5.3) (<https://software.broadinstitute.org/software/igv/>) (37). Furthermore, each peptide-coding region was identified using BLAST-like Alignment Tool (BLAT) in IGV. Any peptide whose corresponding matched lncRNA also matched to the non-coding region or intron region, but not the coding exon region of the human genome, was kept as potential lncRNA-derived peptides. For the ribosome sequencing, we utilized a well-recognized ribo-seq genome browser (<https://gwips.ucc.ie/index.html>), which compiles 46 ribosome profiles (38). The lncRNA containing ORFs were manually searched on the Genome-Wide Information on Protein Synthesis (GWIPS) to confirm the lncRNA-derived peptide-coding regions RNAs were bound to ribosome.

For the empirical *p* value evaluation, 50,000 nonoverlapping genomic segments, each of length 2000 nucleotides (the mean length of lncRNAs in the LNCipedia DB), were sampled uniformly at random from the human reference genome sequence. These random genomic segments, similar in size to LNCipedia, were used as a “mock”/decoy lncRNA DB. Subsequently, an *in silico* translation of these mock/decoy lncRNA sequences was performed *in silico*, using each of the six possible reading frames, to generate a mock/decoy protein sequence DB. Of the entire set of peptides identified by the PEAKS *de novo* algorithm within class I immunopeptidome, 66 had a match in this mock/decoy protein DB. For obtaining a match, the translation of each of the six potential reading frames for every mock/decoy lncRNA sequence was considered. For obtaining a rather pessimistic estimate (*i.e.*, larger than the correct value) on the *p* value for the actual number of matches between the peptides identified by the PEAKS *de novo* algorithm and the LNCipedia DB, we set the probability of a chance match of a peptide (which, on average, has 11 amino acids, coded by 33 nucleotides) to $p = 66 \text{ matches} / (50,000 \text{ transcripts} \times 6 \text{ reading}$

frames \times 2000 nucleotides per transcript/33 nucleotides per peptide) = $3.63e^{-6}$ (this forms the null assumption). LNCipedia DB includes 107,039 lncRNA transcripts from 49,371 distinct lncRNAs. The number of *de novo* peptides identified by the PEAKS *de novo* algorithm within class I immunopeptidome, which had a match in the LNCipedia DB, was 195. The probability of obtaining exactly k matches among 107,039 lncRNA transcripts in the LNCipedia DB, each giving rise to $(6 \times 2000)/33$ potential peptides can be calculated using the aforementioned *null assumption* as

$$q_k = \binom{n}{k} p^k (1-p)^{(n-k)} \quad (1)$$

where $n = (107,038 \times 120,000)/33$. The empirical p value for having 195 matches for class I immunopeptidome among the LNCipedia DB entries can thus be calculated as:

$$p \text{ value} = \sum_{k=195}^{k=n} q_k \quad (2)$$

The lncRNA-derived immunopeptidome bioinformatics analysis was performed on the Biowulf Linux cluster at the NIH (<https://hpc.nih.gov/docs/userguide.html>). The “mock/decoy” DBs, empirical p value calculation method, and the original Python scripts of the identification of HLA-presented lncRNA-derived peptides can be found at Github (https://github.com/YueAndyQi/lncRNA_immunopeptidome_Scripts).

Functional annotation of source proteins of HLA immunopeptidome

We utilized ingenuity pathway analysis (IPA) to determine the subcellular localization/molecular function and cell signaling pathway analysis of the source protein of identified HLA peptides (39). Furthermore, we identified the upstream transcriptional modulators of those source proteins using the unique upstream regulator analysis feature in IPA.

Experimental design and statistical rationale

Two EGFR-mutant lung adenocarcinoma cell lines, PC9 and H1975, harboring EGFR^{Del E746-A750} and EGFR^{L858R/T790M}, respectively, were used for this study representing low TMB tumors. To apply the comprehensive HLA immunopeptidome discovery pipeline in human tumor tissue, a lung tumor harboring EGFR^{L858R} that was procured at rapid autopsy from a patient (NCI-RA007) treated with the third-generation EGFR tyrosine kinase inhibitor, osimertinib, was selected. In addition, two melanoma cell lines, NCI-3784Mel and NCI-3795Mel, generated from tumors of patients with melanoma treated at the NIH Clinical Center, which represent tumors with high TMB. We identified somatic mutations by WES of tumor or tumor-derived cell lines and corresponding germ line DNA from patients NCI-RA007, NCI-3784Mel, and NCI-3795Mel. Expressed somatic mutations were also identified by RNA-Seq of the cell lines and tumor. PC9 and H1975 lung adenocarcinoma cell lines did not have corresponding germ line DNA; hence, all mutations identified by WES and RNA-Seq were included for our data analyses. The identified mutations included SNVs, small INDELs, and fusions. The pan-HLA class I antibody was used to immunoprecipitate class I proteins along with their presented peptides that were sequenced by high-resolution MS/MS. The MS data were searched against cell line- or tumor-specific peptides created by adding all corresponding mutant peptides to the normal human DB. The MS data were also searched using the *de novo* search algorithm in PEAKS studio. To assess experimental reproducibility and perform statistical tests, three biological replicates were performed for PC9, H1975, and NCI-3795Mel and four biological replicates for NCI-

3784Mel and NCI-RA007. Each biological cell line replicate was initiated from the same cell passage number but cultured in separated dishes. Four biological replicates of the tumor specimen NCI-RA007 were obtained by sampling four regions of the tumor obtained at rapid autopsy. The two-way t test and ANOVA test were applied to two-group comparison or more than two-group comparisons, respectively. Finally, since conventional decoy protein sequence-based FDR assessment is not applicable for the PEAKS *de novo* search results, we calculated empirical p value by creating a mock lncRNA DB to control the random matches and false positives for our lncRNA-derived peptide identification.

RESULTS

Identification and characteristics of the HLA class I-presented immunopeptidome and HLA class I expression in melanoma and lung adenocarcinoma

We identified 35,233 HLA class I-presented peptides containing 8 to 14 amino acid residues employing various experimental and informatic tools including pan-HLA class I immunoprecipitation, elution of the class I-presented peptides, high-resolution MS-based peptide sequencing, NGS of genomic DNA/RNA to create patient- and cell line-specific DBs incorporating variant peptides, and computational algorithms (Fig. 1A). These include 14,876 DB-searched peptides and 20,357 *de novo* sequencing algorithm-searched peptides (supplemental Fig. S1A and supplemental Table S1). We performed three to four biological replicates of HLA class I pull-down experiments and MS analyses of associated peptides from each cell line/tumor. Pairwise correlation coefficients of peptide intensities from three representative biological replicates from PC9 cells show relatively high correlation between replicates (supplemental Fig. S1B). We identified 2385 to 4401 DB-matched peptides in each sample. Interestingly, the number of HLA class I-presented peptides identified in melanoma and EGFR-mutant lung cancer was roughly similar (Fig. 1B). More peptides were identified from the lung adenocarcinoma cell lines, PC9 and H1975, compared with the EGFR-mutant tumor, NCI-RA007, and the melanoma patient-derived cell lines. A majority of enriched immunopeptides were 9-mer (Fig. 1C), consistent with the length of HLA class I-bound immunopeptidomes reported previously (13, 40). We further analyzed our dataset using NetMHCpan by which we determined weak and strong binding using %Rank <2.0 and <0.5, respectively. A majority of the enriched peptides were predicted to be binders; nonetheless, 9-mer and 10-mer peptides had lower scores and hence predicted stronger binding (Fig. 1D). The predicted HLA binders were assigned to the expressed HLA alleles in each sample (Fig. 1E). Somatic mutations were identified using WES of tumor and germ line DNA; as expected, EGFR-mutant lung adenocarcinoma patient NCI-RA007 had much fewer somatic mutations (289) compared with the melanoma patient-derived cell lines, NCI-3784Mel and NCI-3795Mel (2678 and 2031, respectively) (supplemental Fig. S1C and supplemental Table S2). The differences in abundance of class I-presented peptides may be a result of the expression level of HLA class I proteins. The total HLA class I protein expression detected

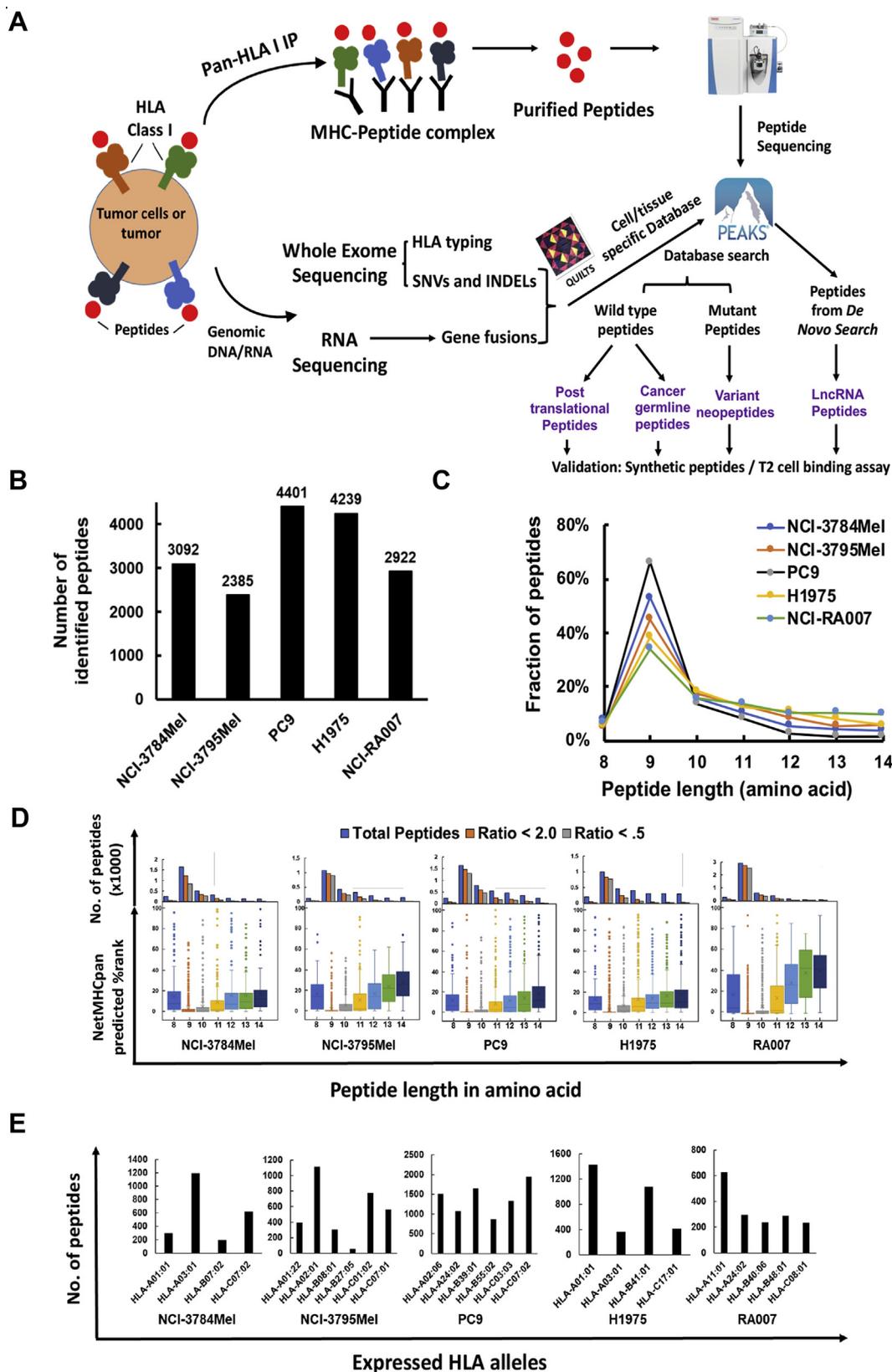


FIG. 1. Overview of HLA class I-presented immunopeptides identified from EGFR-mutant lung cancer and melanoma patient-derived cell lines by MS. A, strategic plan of proteogenomic analysis pipeline for HLA class I immunopeptidome using MS-based proteomics and next-

TABLE 1
HLA typing of cancer cell lines and lung adenocarcinoma tumor

Cell line/tumor	Origin	HLA-A		HLA-B		HLA-C	
NCI-3784Mel	Melanoma	A*01:01	A*03:01	B*07:02		C*07:02	
NCI-3795Mel	Melanoma	A*01:22	A*02:01	B*08:01	B*27:05	C*01:02	C*07:01
PC9	Lung tumor	A*02:06	A*24:02	B*39:01	B*55:02	C*07:02	C*03:03
H1975	Lung tumor	A*01:01	A*03:01	B*41:01		C*17:01	
NCI-RA007	Lung tumor	A*11:01	A*24:02	B*40:06	B*48:01	C*08:01	

by immunoblotting (supplemental Fig. S1D) was consistent with cell surface HLA class I presentation analyzed by flow cytometry (supplemental Fig. S1E) and was approximately similar between the melanoma and the lung adenocarcinoma cell lines. Since HLA proteins are highly polymorphic and the peptides presented are HLA allele restrictive (41, 42), HLA typing is important to further characterize the HLA class I-presented peptides identified. Tumors and cell lines were HLA typed using seq2HLA analysis (25) of the WES data (Table 1), and our data suggest that seq2HLA produced consistent results with conventional HLA typing using SSP sequencing in the two melanoma cell lines (supplemental Fig. S1F). It is interesting to note that NCI-3784, H1975, and NCI-RA007 had loss of heterozygosity of *HLA-B* and/or *HLA-C* alleles. HLA loss of heterozygosity, often caused by genetic alterations of chromosome 6p (43), has been suggested as a mechanism of immune escape (44). Furthermore, we validated our peptidome using previous reported and/or experimentally validated binding peptidome from the IEDB. Since 9-mer is the most common length for peptides in the HLA class I immuno-peptidome, we visualized the 9-mer peptide-binding motifs of our enriched HLA class I peptidome from tumor cell lines/tissue and their corresponding monoallelic datasheets from IEDB. For example, HLA-peptide binding motifs of four major HLA-typed alleles in PC9 cells are the main components of endogenous enriched PC9 immuno-peptide by overlaying the four-individual monoallelic binding motifs (supplemental Fig. S1G). Computational HLA-binding epitope prediction algorithms have been widely recognized as a powerful tool to estimate HLA-ligand binding affinity and tumor neoantigen prediction (45). Notably, although A*01:01 and A*03:01 were both present in NCI-3784mel and H1975, they barely share any binding peptides, suggesting the possibility of different HLA ligand processing and presentation machinery and/or different source protein expression levels in melanoma and lung cancer (supplemental Fig. S2, A and B). We observed a similar phenomenon in HLA-A*24:02 (typed in PC9 and NCI-RA007) and Cw*07 (typed in PC9, NCI-3784Mel, and NCI-3795Mel) (supplemental Fig. S2, C and D).

To annotate the source proteins of all class I-presented immuno-peptides, we classified them based on their subcellular localization and molecular function (supplemental Fig. S3, A and B). A large majority of peptides identified were from cytoplasmic proteins and enzymes. Pathway analysis of identified parent proteins identified key pathway proteins contributing to class I immuno-peptidome, such as eukaryotic initiation factor 2 (EIF2) signaling, protein ubiquitination pathway, EIF4 and p70S6K signaling, and others (supplemental Fig. S3C). IPA upstream regulator analysis showed tumor suppressor TP53 and proto-oncogenes *MYC*, *KRAS*, *ESR1*, *ERBB2*, *EGFR*, and *MTOR* to be among the top upstream potential regulators of the parent proteins identified (supplemental Fig. S3D). Network analysis confirmed that tumor suppressors (e.g., TP53, BRCA1) and oncogenes (e.g., EGFR, KRAS, MYC) were components of the network of parent proteins represented by the identified class I-presented peptides (supplemental Fig. S3E). To further supplement the bioinformatics analyses, we identified HLA class I-presented peptides from common proto-oncogenes, such as *KRAS*, *EGFR*, *MYC*, *JUN*, and tumor suppressors, such as *TP53*, *RB1*, and *BRCA2*. We identified 17 common cancer driver gene-derived wildtype peptides presented by HLA class I in lung adenocarcinoma cell lines/tumor and two peptides in primary tumor-derived melanoma cells; six of which are novel peptides not previously reported. In addition, we used NetMHCpan to predict the HLA class I restriction of the identified peptides from cancer drivers. We found that, though some peptides have been predicted to be binders for multiple respective HLA alleles (e.g., KQFEGTVEI derived from BRCA2), their MS intensity is not significantly higher than that of the peptides predicted to be monoallelic binders (e.g., KLISEEDLLRK derived from MYC) (supplemental Fig. S3F and supplemental Table S3). Taken together, our analyses identified 19 high-confidence oncogene/tumor suppressor-derived peptides that were presented by HLA class I.

generation sequencing. *B*, total number of peptides identified in the HLA class I immuno-peptidomes from patient-derived melanoma and EGFR-mutant lung cancer cell lines and tumor. *C*, the peptide length distribution within the class I immuno-peptidome from all samples. *D*, NetMHCpan 4.0 prediction algorithm-based scoring of each MS-identified peptide and distribution of binding scores among 8- to 14-mer peptides. *Upper panel* shows the distribution of total identified peptides, binders (%Rank <2.0), and strong binders (%Rank <0.5) for each peptide length. *Lower panel* shows box plots of the lowest NetMHCpan predicted %Rank for binding among the corresponding HLA class I alleles for each peptide identified for various peptide lengths. *E*, number of predicted binders (%Rank <2.0) assigned to different HLA alleles for each sample based on its corresponding HLA class I alleles. EGFR, epidermal growth factor receptor; HLA, human leukocyte antigen.

Discovery of variant class I-presented peptides in lung adenocarcinoma

Variant peptides derived from somatic mutations in tumors, if presented by MHC class I, have the potential to engage cytotoxic T cells to promote tumor immunity. Hence, the identification of variant peptides presented by the cognate class I proteins is of paramount importance. To identify mutant

peptides, we first constructed cell line- and patient-specific DBs by adding all somatic variants identified by NGS (RNA-Seq and WES) to normal human DB used for searching MS data when germ line DNA was available for sequencing (NCI-3784Mel, NCI-3795Mel, and NCI-RA007) (Fig. 2A). For PC9 and H1975 cells, which do not have available germ line DNA, all variants identified by exome sequencing were used. We identified 12 peptides harboring SNVs in the two lung

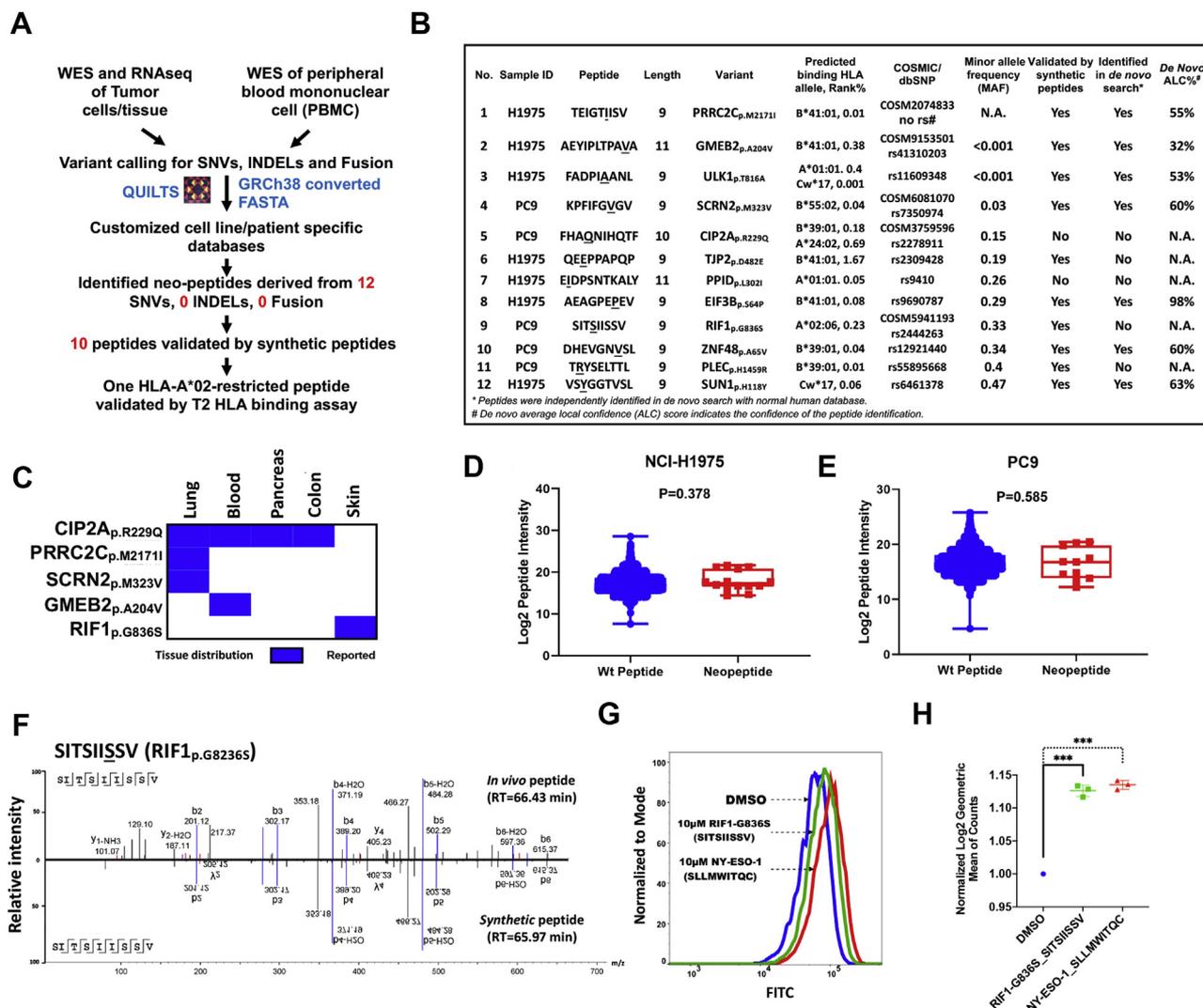


FIG. 2. Identification of class I-presented variant peptides in lung adenocarcinoma. A, workflow of integrated proteogenomic analysis using germ line (peripheral blood mononuclear cells) and tumor cell line/tissue WES and RNA-Seq datasets to identify SNVs, INDELS, and fusions and construction of tumor cell line-/tumor tissue-specific databases to interrogate the MS data of class I-associated peptides. B, list of 12 variant peptides, with the mutation underlined, predicted HLA restriction, dbSNP ID, synthetic peptide validation, and de novo sequencing search ALC score. C, tumor tissue distribution of five COSMIC-reported somatic mutations in the identified variant peptides. D and E, box plots show peptide intensity of wildtype and mutant peptides in all biological replicates from C, H1975 and D, PC9 cells. F, matched MS2 spectra of endogenous and its synthetic counterpart for RIF1-G836S-derived peptide SITSI^{ISSV}. G, T2 cell-based HLA stability assay showing RIF1_p.G836S-derived peptide (SITSI^{ISSV}) bounding and stabilizing HLA-A*02. H, box plot shows statistically significant increase of HLA expression (log₂ geometric mean of counts) in T2 cells incubated with RIF1_p.G836S-derived peptide (SITSI^{ISSV}) and positive control NY-ESO-1-derived peptide compared with those incubated with DMSO (p < 0.005). ALC, average local confidence; COSMIC, Catalogue of Somatic Mutations in Cancer; DMSO, dimethyl sulfoxide; HLA, human leukocyte antigen; INDEL, insertion and deletion; SNV, single nucleotide variant; WES, whole-exome sequencing.

adenocarcinoma cell lines but no INDELS and fusions. Indeed, further analyses of these variant peptides using NetMHCpan predicted all to be binders for at least one HLA allele in their corresponding cell line (Fig. 2B). We leveraged dbSNP and Catalogue Of Somatic Mutations in Cancer (COSMIC) DBs to classify whether the peptides were derived from normal polymorphism or somatic mutations. We identified five mutations that were reported in COSMIC DB, and three of them have a minor allele frequency <0.05, confirming they are rare mutations. Next, we interrogated the tumor types in which the five variants have been identified. Indeed, proline-rich coiled-coil 2C (PRRC2C)_{p.M2171I} and secernin 2 (SCRN2)_{p.M323V} were exclusively discovered in lung cancer; importantly, cellular inhibitor of PP2A (CIP2A)_{p.R229Q} (legacy identifier: COSM3759596) has been confirmed as a somatic mutation in multiple cancers (*i.e.*, lung, blood, pancreas, and colon) according to COSMIC DB (Fig. 2C). Therefore, peptide FHAQ-NIHQTF derived from somatic mutation CIP2A_{p.R229Q} might be a potential CD8 T-cell target to a variety of cancer patients carrying this somatic mutation and corresponding HLA alleles (*e.g.*, A*24:02 and/or B*39:01).

We compared the MS intensities of the 12 variant peptides and nonmutated class I-presented peptides and found no significant difference in median peptide intensities between the two groups for both the H1975 and PC9 cell lines (Fig. 2, D and E), suggesting that individual variant peptides are presented to similar extent as wildtype peptides. In addition, to validate the peptide sequences identified in our class I pull-down experiments, we synthesized a subset of the identified peptides and utilized LC MS/MS to compare the MS2 spectra identified for the synthesized and endogenous peptides. MS2 spectra and RT of ten of 12 synthetic peptides and their corresponding endogenous peptides were matched (Fig. 2F and supplemental Fig. S4, A–J). Next, we confirmed the binding of the peptides to specific predicted HLA alleles. Peptide SITSISSV, derived from RIF1_{p.G836S}, binds strongly to HLA-A*02:06 with %Rank at 0.23. To assay cell surface HLA stabilization, we pulsed this peptide overnight to a TAP-deficient T2 cell line that expresses only HLA-A*02 (46) and used HLA-A*02 binding peptide SLLMWITQC from NY-ESO-1 as a positive control in parallel. Both peptides significantly stabilized HLA to the cell surface in comparison to the DMSO control (Fig. 2, G and H). Taken together, we identified 12 variant peptides in lung adenocarcinoma cell lines; notably, five of these were derived from reported somatic mutations. With spectra and HLA-binding validations, our results suggest that these variant peptides are potential targets for cancer immunotherapy.

CG antigen-derived peptides presented by HLA class I in melanoma and lung adenocarcinoma

CG or tumor testis antigens hold great potential for generating tumor-specific antigens for T-cell-based therapy (47). CG antigens are exclusively expressed or overexpressed in tumor and germ cells. These peptide antigens are rarely

presented to immune cells because of the relatively low HLA expression in testis and germ line tissues (48, 49). In order to search for these CG antigens in our MS data for class I-presented peptides, we established a customized CG antigen library by compiling 285 CG antigens from various cancer testis antigen and tumor-associated antigen DBs (Fig. 3A and supplemental Table S4). We identified a total of 40 CG antigen peptides derived from 14 CG antigen proteins. Of the 40 CG antigen-derived peptides, 27 are from melanoma and 13 from lung adenocarcinoma, and of these, seven were novel and, to our knowledge, have not been reported previously (Fig. 3B and supplemental Table S5). We selected 16 synthetic peptides for further validation and confirmed seven of these peptides from the melanoma patient-derived cell lines and one from the lung adenocarcinoma patient tumor (Fig. 3C and supplemental Fig. S5, A–H). We identified 15 melanocyte protein PMEL (GP100)-derived epitopes in the 3784Mel cell line derived from the tumor of a patient whose tumor-infiltrating CD8⁺ lymphocytes have been previously shown to recognize the GP100 antigen (18). Although CG antigens have been extensively studied in melanoma, we identified one novel peptide, VTPVEVHIGT, derived from sperm-associated antigen 17. In contrast, and of particular interest, we identified four novel peptides in the H1975 lung adenocarcinoma cell line. These include peptides mapping to testis-expressed protein 15 (50) and lactate dehydrogenase C (LDHC) (51) (supplemental Fig. S5f). Next, we verified whether the genes for these proteins were expressed at the transcript (mRNA) and protein levels in H1975 cells where the total RNA and whole-cell proteome were profiled separately (supplemental Table S6). Interestingly, expression of testis-expressed protein 15 and LDHC RNA was lower, and protein was not detected, underscoring the possibility that class I presentation can occur for genes expressed at low levels (Fig. 3D). We also ranked CG antigen gene and protein expression in H1975 cells and found representation of class I-associated peptides for genes with protein levels undetectable by MS but detectable at the transcript level (Fig. 3E). Therefore, our data support the phenomenon that immunopeptidome lacked association with gene/protein expression, which has been demonstrated previously using dynamic stable isotope labeling by/with amino acids in cell culture approach (52, 53).

In vivo PTM peptides are presented by HLA class I and are potential neoantigens

PTMs may alter the binding affinity of class I-presented peptides. Prior studies have identified PTMs in class I-presented immunopeptides (12, 14). To identify the *in vivo* PTM peptides, we used PEAKS studio with the pan-PTMs selected (over 650 variable modifications) to detect all possible PTMs. Peptide modification artifacts induced by sample preparation (*e.g.*, urea, reducing, and alkylating reagents) and electrospray ionization are the major concerns for *in vivo* PTM identification (54). Our HLA class I-presented peptidome enrichment protocol

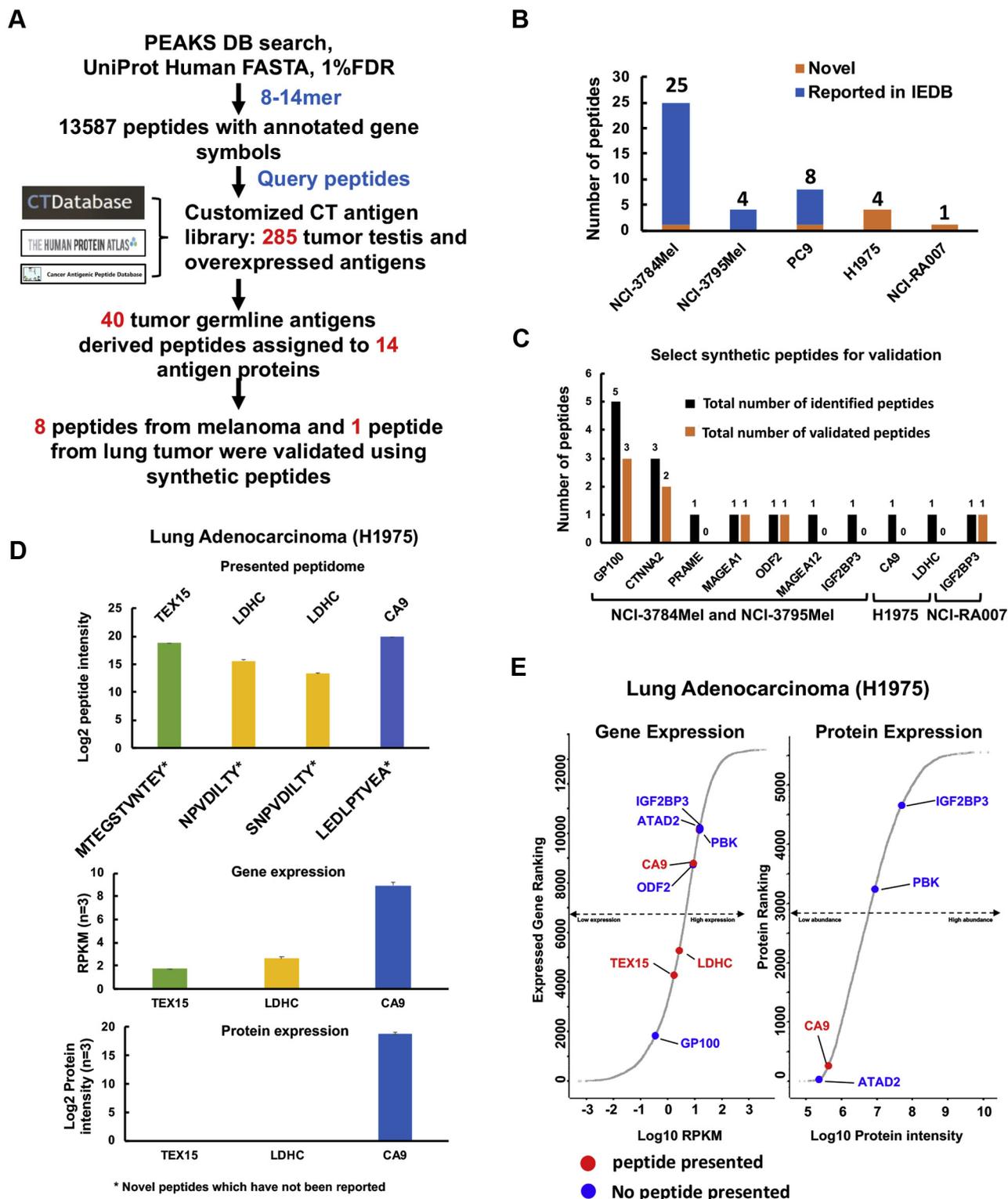


FIG. 3. Identification of cancer germ line antigen-derived peptides. *A*, overall strategy of CG antigen identification using customized library integrating previously reported immunogenic CG antigens; a total of 285 CG antigens were included in this custom database. *B*, the number of novel and reported CG antigen-derived peptides. *C*, summary chart of select peptides validated in NCI-3784Mel, NCI-3795Mel, H1975, and NCI-RA007 using synthetic peptide spectra match. *D*, log₂ median peptide intensity of four immunopeptides identified in H1975 (*top panel*); gene expression (*middle panel*), and protein expression (*lower panel*) of three source antigen proteins. *E*, rank of transcript and protein expression of CG antigens identified in H1975 using RNA-Seq and MS, respectively. *Red dots* indicate CG antigens from which at least one class I-presented peptide was identified by MS. *Blue dots* indicate the CG antigens that were expressed in H1975 cells, but no class I-presented peptide was identified by MS. CG, cancer germ line.

did not involve the use of urea buffer, protein reduction, and alkylation or enzyme digestion. Electrospray ionization artifacts could be excluded by examining the RT of modified peptides and their unmodified counterparts. Since the artifact modifications are only added during the ionization process, these modified peptides must be coeluted with their unmodified counterparts; yet we did not observe coeluted peptides with and without modifications. We identified 1389 modified and 11,841 unmodified 8- to 14-mer peptides. We did not detect the corresponding unmodified form for 804 of the modified peptides. These groups of peptides are defined as “modified only.” On the other hand, for 411 of the unmodified peptides, we identified 527 modified counterparts this is a group of peptides defined as “modified and unmodified,” suggesting the existence of multiple PTMs for some peptides (Fig. 4A and supplemental Table S7). Approximately 10% of the total peptides identified had at least one PTM, and we identified 58 different PTMs making this the largest HLA PTM immunopeptidome identified to date. The heat map shows the median intensity of each PTM in each sample whereby methionine oxidation (639 peptides), deamidation (151 peptides), acetylation (144 peptides), and methylation (78 peptides) were the most abundant modifications seen (Fig. 4B) in agreement with previous reports (12, 13). Interestingly, among the 9-mer peptides identified, the N-terminal amino acid was most commonly modified; the first amino acid was modified in 102/293 9-mer PTM peptides (Fig. 4C). PTM HLA peptides were previously shown to be more abundant than their unmodified counterparts (12). However, in our dataset, pan-PTM peptides were significantly less abundant (as measured by peptide intensity) than unmodified peptides ($p = 5.4E-16$) (supplemental Fig. S6A). Similarly, MS intensity was lower for deamidated and methylated peptides than unmodified peptides (supplemental Fig. S6, B and C). However, the median intensity of glutamate- to pyroglutamate (pyroGlu)-modified peptides was similar to that of their unmodified counterparts (supplemental Fig. S6D). The pyroGlu modification occurred on glutamate at N-terminal position 1, which does not significantly affect the peptide conformation and binding affinity (55).

To validate the data quality and verify that the modifications were generated *in vivo* and not experimental artifacts, we selectively examined the tandem mass spectra of the modified peptides. A deamidated and a methylated form of peptide TNQELQEINR, derived from annexin A2, had an RT of 28.51 and 32.61 min, respectively, whereas the unmodified peptide had an RT of 28.20 min (Fig. 4D). We manually verified the RT of methionine-oxidized peptides and their unaffected counterparts and determined that they did not coelute during LC, suggesting that these PTM peptides likely were generated *in vivo* and were not ionization artifacts. Next, we determined whether the PTMs may alter the binding affinity to HLA class I. There are no HLA-binding prediction algorithms commercially available that accounts for PTM peptides. Using the unmodified forms of the peptides identified for NetMHCpan analysis of

tumor tissue from NCI-RA007, 637 of 782 PTM peptides were considered to be nonbinders (%Rank >2.0), suggesting that specific modifications of these peptides may have been critical for HLA binding. Notably, the percentage of predicted binders for peptides with both modified and unmodified counterparts was nearly twice (26.46%) that of solely modified (14.67%). Also, the percentage of modified-only peptides among predicted binders (53.1%) was much lower than for nonbinders (70.33%) (Fig. 4E). We plotted peptide length distribution of the binders and nonbinders and found that they had a similar pattern although more nonbinders were longer peptides (>10-mer) (Fig. 4F). This further suggests that PTM peptides may have nonconventional length distribution for HLA class I binding; importantly, the unmodified forms of these peptides are more likely to be predicted as “nonbinders.” Collectively, our findings suggest that PTMs may play a crucial role for generating a subset of HLA class I-binding peptides with unique binding motifs for antigen presentation.

De novo sequencing provides reliable immunopeptide identification

De novo search of MS spectra from large-scale MS data has been employed by various algorithms, including PEAKS studio. The impressive prediction accuracy of this approach has been extensively reported (24, 56, 57). We searched our entire class I-presented immunopeptidome MS data using the PEAKS studio *de novo* search algorithm. First, to evaluate the data quality of the peptides identified by *de novo* sequencing, we manually inspected the MS2 spectra. We then employed the NetMHCpan prediction to evaluate the binding capacity of our *de novo*-only peptides (not including DB-searched peptides). We determined that while an average of ~55% of the DB-searched peptides were predicted to be specific HLA allele binders for their corresponding cell line/tissue, an average of ~33% of *de novo* peptides were predicted to be a strong binder of at least one HLA allele in the respective sample (Fig. 5, A and B), the distribution of %Rank of DB and *de novo* search peptides in each sample is shown in supplemental Fig. S7A. The predicted HI and RT of 9-mer peptides identified by DB and *de novo* search showed significant correlations while, as predicted, the HI and RT of the tryptic peptides showed even higher correlation (supplemental Fig. S7, B and D). We employed a similar approach to evaluate the HLA-binding affinity of 8- to 14-mer peptides identified by DB search (Fig. 1D) by predicting the binding affinity and assigning each peptide to its highest predicted HLA allele and showing the distribution of 8- to 14-mer peptides according to their lowest %Rank. We observed the same trend with respect to binding score distribution of 8- to 14-mer peptides when comparing *de novo* to DB-identified peptides. The 9-mer peptides have the lowest binding scores and hence highest binding affinity to the best predicted HLA-binding allele (Fig. 5C). To further confirm the validity of the peptide identities from the *de novo* sequencing analysis, we

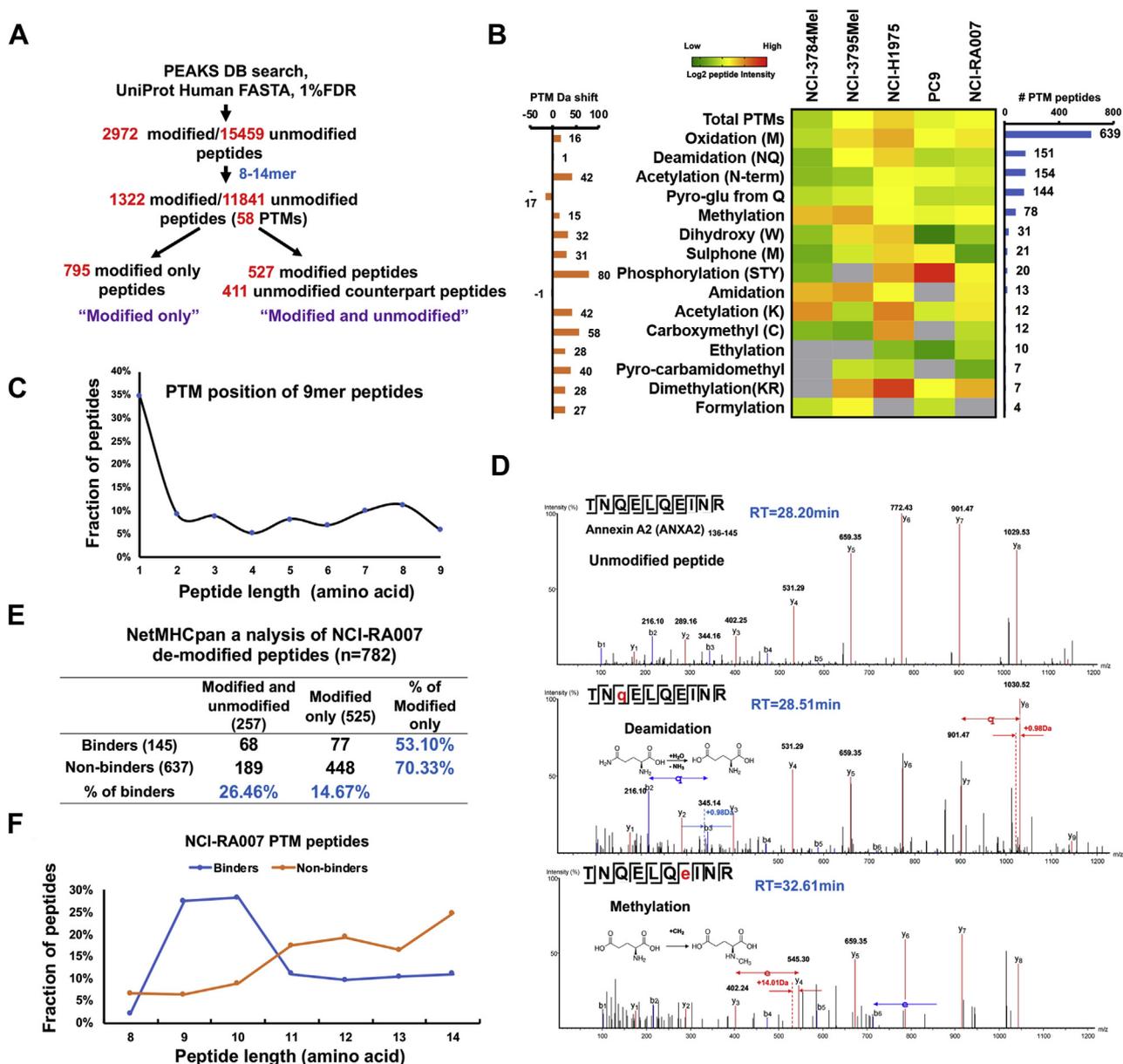


FIG. 4. Characterization of post-translationally modified (PTM) HLA class I-associated immunopeptides. *A*, schema of PTM immunopeptidome profiling pipeline. *B*, heat map shows the median log₂ intensity of all identified class I-presented PTM peptides; *left bar graph* displays the molecular weight shift of each PTM; *right bar graph* shows the total number of peptides identified in each PTM group. *C*, PTM amino acid position distribution of the 9-mer PTM peptides. *D*, MS/MS spectra of one representative peptide with multiple PTMs. ANXA2-derived peptide TNQELQEIINR, identified as unmodified (*top panel*), deamidated (*middle panel*), and methylated (*lower panel*). *E*, NetMHCpan analysis of 782 PTM peptides identified in sample NCI-RA007 where peptides with %Rank <2.0 are considered binders. The table shows the percentage of predicted “binders” in the “unmodified and modified” and “modified-only” peptides; it also shows the percentage of “modified-only” peptides in the “binders” and “nonbinders” peptides. *F*, peptide length distribution of predicted binder and nonbinder PTM peptides identified in NCI-RA007. HLA, human leukocyte antigen.

compared the predicted DB and *de novo* 9-mer HLA-binding peptide motifs. Interestingly, the peptide motifs were very similar between the two groups, reinforcing the validity of the *de novo*-identified peptides generated from all five samples (Fig. 5, D–H). We next aligned the *de novo* and DB search spectra for four representative endogenous variant peptides identified in H1975. The b and y ions of the spectra from the

endogenous variant EIF3B_{p.S64P} peptide (AEAGPEPEV) (Fig. 5I) as well as three additional variant peptides (supplemental Fig. S7, E–G) perfectly aligned with the spectra from corresponding synthetic peptides. Furthermore, the predicted HI and RT of these four *de novo* peptides identified in H1975 were well correlated with $r^2 = 0.99$ (supplemental Fig. S7H). We further validated our *de novo* sequencing

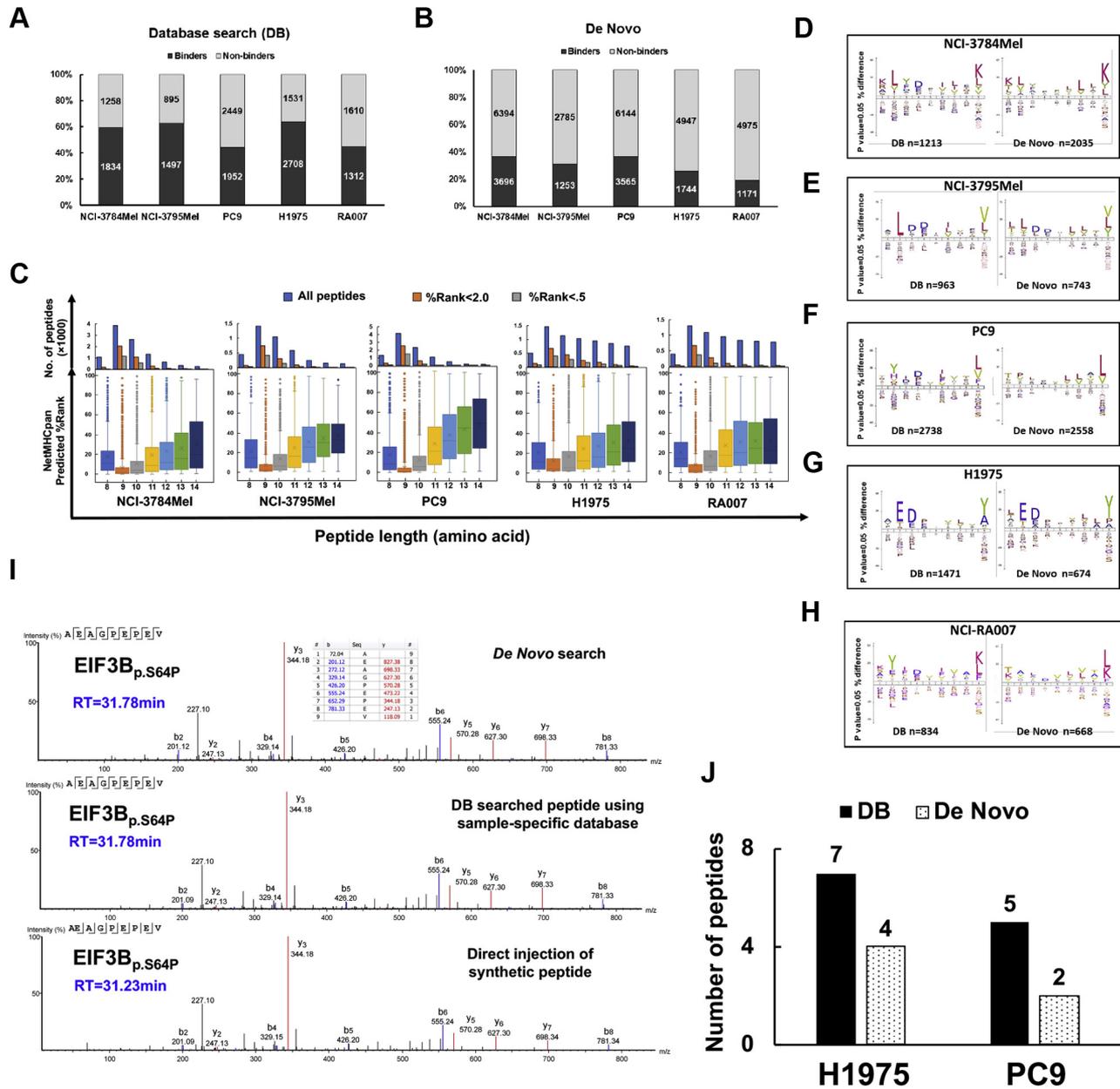


FIG. 5. *De novo* search of MS data to identify class I-associated immunopeptides. *A* and *B*, fraction of class I-associated peptides predicted to be HLA binders (NetMHCpan %Rank < 2.0) or nonbinders (%Rank > 2.0) by *A*, database (DB) search and *B*, *de novo* sequencing. *C*, NetMHCpan predicted %Rank distribution of total number of *de novo* sequencing-searched 8- to 14-mer peptides; HLA binders (%Rank < 2.0) and strong binders (%Rank < 0.5) (upper panel). Distribution of HLA-binding affinity (%Rank) of *de novo* sequencing-searched 8- to 14-mer peptides (lower panel). 9-mer peptides have the lowest predicted %Rank, suggesting the strongest binding. *D–H*, comparison of 9-mer peptide-binding motifs (binders only, NetMHCpan %Rank < 2.0) identified by DB search (left) versus *de novo* search (right) in *D*, NCI-3784Mel, *E*, NCI-3795Mel, *F*, PC9, *G*, H1975, and *H*, NCI-RA007. *I*, matched MS2 spectra and RT of one representative endogenous variant peptide, EIF3B_{p.S64P} (AEAGPEPEV), identified by *de novo* search (upper panel), proteogenomic DB search (middle panel), and direct injection of its synthetic peptide (lower panel). *J*, bar charts show that variant peptides identified by proteogenomic analysis were also retrieved in the *de novo*-only search. HLA, human leukocyte antigen; RT, retention time.

pipeline-identified variant peptides from the lung adenocarcinoma cell lines by comparison to those identified using cell line-specific DB search. Indeed, we found four of seven and two of five mutant peptides in H1975 and PC9 cells with

relatively high confident ALC%, respectively (Fig. 5J). In summary, we present strong evidence that this is a robust and reliable *de novo* sequencing pipeline for MS identification of the immunopeptidome.

Identification of lnc-RNA-derived peptides using de novo sequencing

Noncoding regions in the genome are the most unexplored; yet they are rich sources of neoantigens. Previous studies have profiled the noncoding immunopeptidome using the traditional proteogenomic approach of searching the MS raw files against sample-specific library generated from RNA-Seq data (58, 59), which always resulted in extremely massive customized noncoding sequence libraries because 99% of the human genome is noncoding (60), a majority of which are unannotated. We developed a pipeline to profile potential lncRNA-derived peptides by taking advantage of deep *de novo* analysis of MS data without using a predefined DB and then matching the MS-identified peptides with hypothetical peptides generated by six-frame translation of all lncRNAs from an available lncRNA DB, LNCipedia, containing 49,372 lncRNAs from the high-confidence genome assembly (33). The 8- to 14-mer *de novo* peptides were queried against all six potential reading frames of the translated LNCipedia-derived protein DB. We also confirmed the transcript expression of the lncRNAs coding the identified peptides in RNA-Seq gene expression data from the patient-derived cell lines. IGV was used to visualize peptide coding regions of these lncRNAs. We finally validated the endogenous peptides derived from the lncRNAs with synthetic peptide-based spectra matching and T2 cell-based HLA stability assay for binding to specific HLA alleles (Fig. 6A and supplemental Fig. S8A). A total of 195 distinct *de novo* sequencing-identified peptides matched to the six-frame translated lncRNAs in the LNCipedia DB, of which, 71 were predicted to be binders (%Rank <2.0) for at least one HLA allele in their corresponding cell line/tumor. We further analyzed the RNA-Seq data and found that the source RNAs of 53 peptides were transcribed. The feature counts of these transcribed lncRNAs were displayed in a heat map (supplemental Fig. S8B). We then confirmed 44 peptides had their specific coding regions transcribed and did not overlap with any protein-coding region using the BLAT in IGV (Fig. 6B and supplemental Table S8). Two representative examples of the data visualization using IGV for the peptides identified from lnc-JAM3-3:12 and lnc-LRP5-1:11 are shown (supplemental Fig. S8, C and D). Notably, we did not observe any peptide presented in more than one sample, implying that lncRNA-derived peptides may be tumor cell line specific and patient specific. To assess the significance of our findings, we generated a mock lncRNA pool from randomly sampled ~50,000 genomic sequences. We rejected the null hypothesis that our identified lncRNA-derived peptides randomly matched to a six-frame-translated LNCipedia-derived DB, as we obtained a significant empirical *p* value (1.1×10^{-5}) upon comparing to the random matches against the mock lncRNA pool-derived six-frame translated DB.

Next, we asked whether lncRNA-derived class I-presented peptides had low abundance. We found that lncRNA-derived

peptides were equally presented on class I as all other DB and *de novo* sequencing-derived peptides (Fig. 6C). Based on LNCipedia classification, the source lncRNAs matched to 20 antisense genes, 12 sense intronic genes, and 12 lncRNA genes (Fig. 6D). Furthermore, BLAT analysis revealed that 23 lncRNA-coding regions had no overlap with any coding region on hg38; five lncRNAs matched to the introns of coding genes, and 16 matched to novel ORFs because of the frameshift that intersected with exons of known protein-coding genes, but with different start codons (Fig. 6E). Frameshifted new ORFs have been suggested to be a rich source of neoantigens (61).

It remains unclear whether lncRNA can be translated to protein products (e.g., full-length/truncated proteins, peptides), and importantly, presented to cell surface by HLA class I. For in-depth illustration of our computational strategy, we analyzed further the lncRNA oncogene, *PVT1*-derived 9-mer peptide, FLLSSSLTL, identified in PC9 cells (Fig. 6F). RNA-Seq BAM files were converted to BED files for all five samples. The DNA coding sequence of the identified peptide, FLLSSSLTL, was retrieved from the *PVT1* nucleotide sequence in LNCipedia. The BLAT results of this 27-base pair sequence (i.e., TTC ... CTT) were *in silico* transcribed to RNA sequence (i.e., UUC ... CUU) and further translated to peptide sequence FLLSSSLTL. Interestingly, we found that the source lncRNA was transcribed in all five samples, but a truncated version was transcribed in NCI-3795Mel. However, the peptide was only identified in PC9 cells, suggesting that translation of lncRNA-derived peptides may be cell specific and context specific, or the presentation of lnc-RNA-derived peptides by class I is cell line specific or tumor specific. We reasoned that the specificity to PC9 cells could be a result of the lack of specific HLA allele to present this A*02 restricted peptide (%Rank = 0.11). Only PC9 and NCI-3795Mel express HLA-A*02, and NCI-3795Mel expresses a truncated version of FLLSSSLTL; this may explain why we only observed this peptide presented by class I in PC9 cells. We further confirmed that the ORF containing this peptide coding sequence has been reported in the ribosome profiling data based on deep sequencing of ribosome protected mRNA fragments that can be visualized using GWIPS-viz (38). We further confirmed our MS identification of these lncRNA-derived peptides that are presented by HLA class I. We used synthetic peptides and performed LC-MS/MS to compare the MS2 spectra of synthetic peptides with those of the endogenous peptides (Fig. 7, A and B). Finally, we used the T2 cell-based HLA stability assay to confirm that three lncRNA-derived peptides from the lncRNAs, *PVT1*, lnc-SYT2-4, and RC3H1-It1, were truly HLA-A*02 binders (Fig. 7C). Taken together, we report a novel MS-based class I-associated peptidome profiling platform for identification of lncRNA-derived peptides that are presented by HLA class I.

DISCUSSION

Neoantigens or cancer-associated antigens are attractive immunotherapeutic targets because they specifically engage

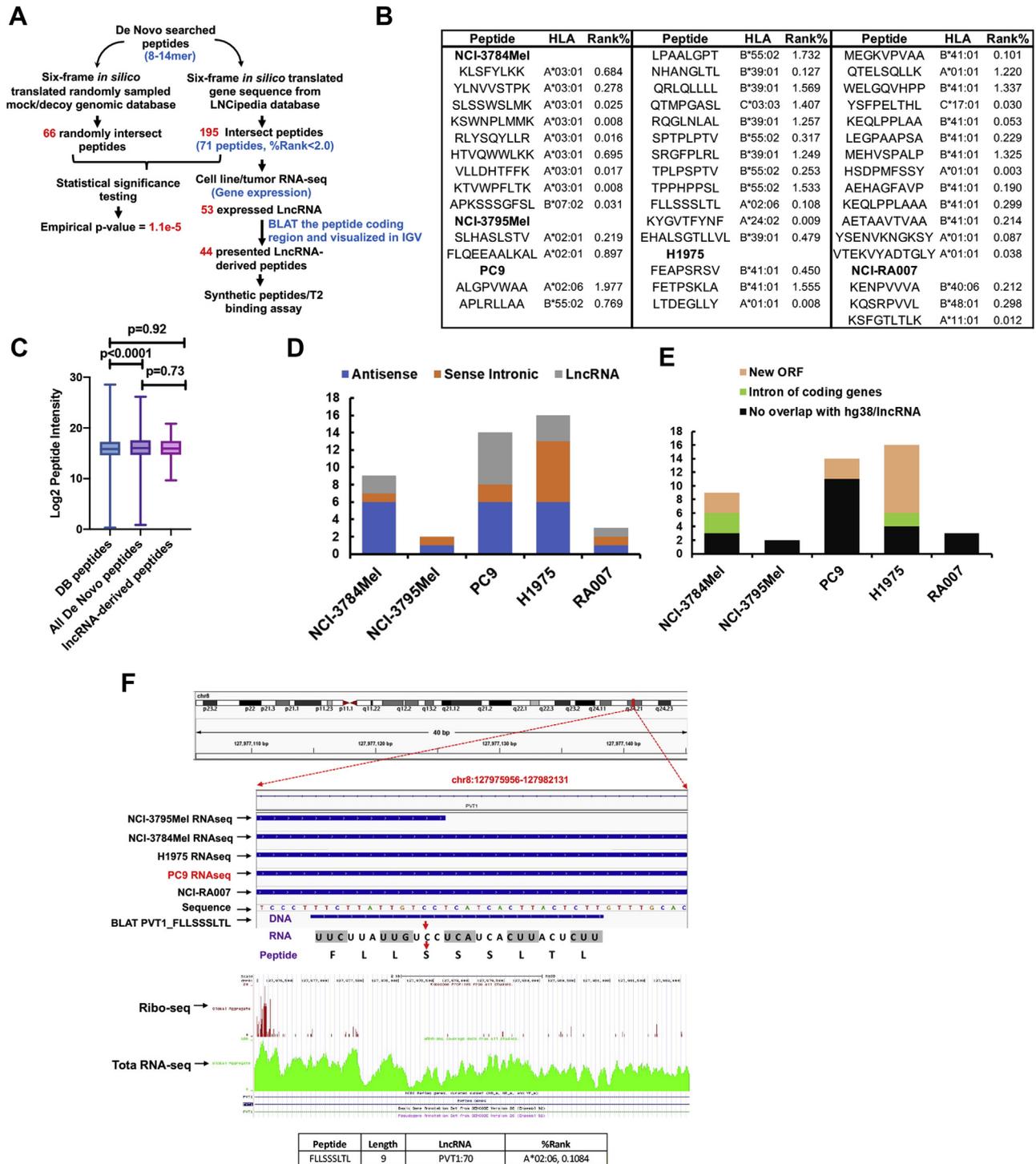


FIG. 6. Identification of lncRNA-derived peptides by proteogenomic and *de novo* sequencing analyses. A, schema showing the workflow used to identify lncRNA-derived peptides enriched from cancer cells and tumors. The *de novo*-only sequencing-searched class I-presented peptide pool was queried against a database (DB) generated using six-frame translated lncRNAs compiled in LNCipedia DB (*right workflow*). The statistical significance of our algorithm was determined by potential matching of the *de novo*-searched peptides against a “mock” DB created by the randomly picked gene blocks (~50,000 transcripts) from hg38, which resulted in an empirical *p* value <1.0e⁻⁵ (*left workflow*). B, 44 lncRNA-derived peptides identified using our algorithm with their predicted HLA alleles and binding affinity. C, comparison of log₂ peptide intensities of DB searched, *de novo* searched, and lncRNA-derived peptides. D, the classification of source lncRNAs for the identified lncRNA-derived peptides into antisense, sense intronic, and classic lncRNAs. E, lncRNA-derived peptides that match new ORF, introns of coding genes, and noncoding region. F, *top panel* displays a snapshot of IGV showing lncRNA PVT1-derived peptide FLLSSSLTL.

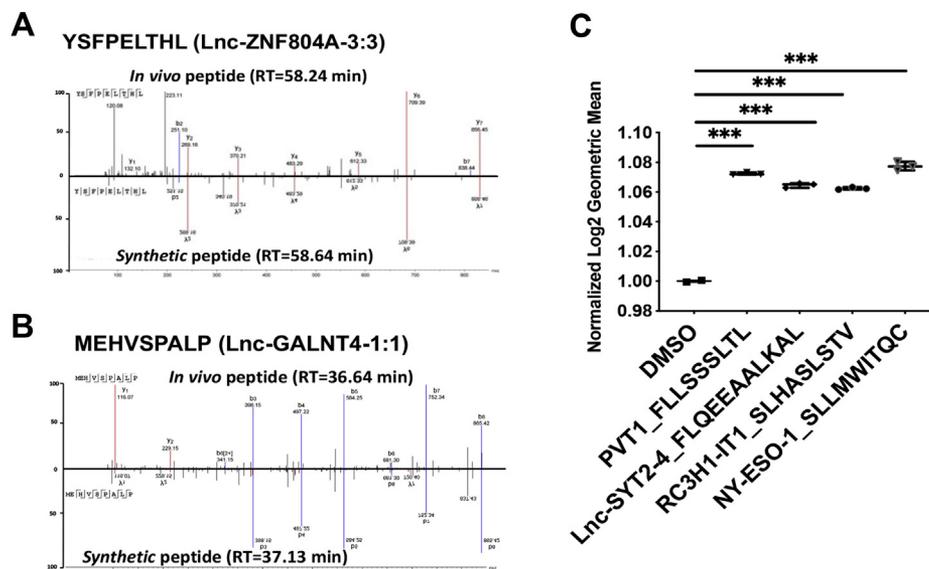


FIG. 7. Validation of lncRNA-derived peptides using synthetic peptides and T2 binding assay. A and B, MS2 spectra matching of two lncRNA-derived peptides and their synthetic counterparts, YSFPETHL and MEHVSPALP. C, T2 cell-based HLA stability assay showing the three lncRNA peptides predicted to be HLA-A*02 binders, FLLSSSLTL, QEEAALKAL, and SLHASLSTV, and the NY-ESO-1-derived positive control peptide binding and stabilizing the HLA allele. HLA, human leukocyte antigen; lncRNA, long noncoding RNA.

T-cell receptors in T cells that promote an immune response against the tumor tissue while sparing nearby healthy tissues. Emerging evidence obtained from breast cancer (62), bladder cancer (63), melanoma (64), and lung cancer (65) studies suggests that cancer neoantigens and cancer-associated antigens may be ideal targets for ACT and therapeutic cancer vaccines. MS-based peptide sequencing technology provides direct experimental evidence for a large number of HLA-presented peptides. As such, this approach has become a robust and quick method of neoantigen discovery (10, 13, 66). Nevertheless, a majority of studies have focused on tumor types with high TMB, such as melanoma (14, 67). EGFR-mutant lung cancer, with low TMB, and tumors with loss of neoantigen expression while on immunotherapy are relatively resistant to immune checkpoint therapy (68–70). In this study, it was our intent to leverage discovery proteomics and informatics to identify HLA class I-presented peptide antigens, including common driver oncogenes, variant peptides, CG antigen peptides, PTM peptides, and lncRNA-translated peptides for immunotherapy in EGFR-mutant lung adenocarcinoma, a tumor type that is historically less responsive to immune checkpoint inhibitor therapy for variety of reasons, including low TMB. Moreover, we utilized spectra matching between *in vivo* peptides and synthetic peptides as well as HLA-binding assays to further validate our identification of class I-presented peptides. We acknowledge that this study

contains relatively small sample size; nonetheless, it readily represented two cancer types, melanoma and lung cancer, and two types of specimen, cancer cell line and tumor tissue. To our knowledge, such a large-scale study has not been performed in EGFR-mutant lung cancer which does not respond to traditional checkpoint inhibitor therapy. Some of the identified peptides derived from tumor-specific antigens or tumor-associated antigens may be further validated and be candidates for developing precision immunotherapy.

To confirm the validity of the class I immunopeptidome identified in this study, we first verified the quality of our DB-searched HLA class I-presented peptides by confirming that a majority of the peptides have high binding scores against their cognate HLA alleles expressed in the source tumors and cell lines. This is consistent with previous large-scale mono-allelic HLA class I epitope profiling studies (13, 71). The NetMHCpan-predicted binding scores of the identified peptides favor 9- and 10-mer peptides that are of the optimal length for class I presentation (Fig. 1D). Combining SNVs, INDELS, and fusion variants, our customized search algorithm was able to identify five COSMIC-reported mutation-containing peptides, of which, FHAQNIHQTF, derived from somatic mutation CIP2A_{p.R229Q} has been associated with several solid cancers, including lung adenocarcinoma, colon cancer, and pancreatic cancer (Fig. 2C). No significant correlation was observed between total mutation burden and HLA

identified in PC9, with chromosomal location alignment of RNA-Seq of all five samples and peptide BLAT; the *middle panel* shows the ribo-seq searching results from GWIPS; the *lower panel* shows the predicted binding affinity of this peptide to HLA-A*02:06 that is expressed only in PC9 cells. BLAT, BLAST-like Alignment Tool; GWIPS, Genome-Wide Information on Protein Synthesis; HLA, human leukocyte antigen; IGV, Integrative Genomics Viewer; lncRNA, long noncoding RNA.

presentation in the low and high TMB cancers. Similar identification of modestly large HLA peptidome that was reported in a low TMB cancer, such as renal cell carcinoma, supports our findings (72). One possible interpretation is that total number of identified class I-presented immunopeptides more directly relates to the purified amounts of HLA proteins and its allotype diversity, and that is similar between our chosen patient-derived cell lines (Table 1 and supplemental Fig. S1D). Some of the mutated peptides may not be presented by HLA class I; rather, specific variant peptides are still presented in low TMB cancers, and identification of those by direct MS will be beneficial for designing precision immunotherapies. We acknowledge that class I- and class II-presented peptides harboring the truncal mutations in common oncogenes, such as *EGFR* and *KRAS*, may be the most attractive targets for ACT (8, 65). We have reported the identification of somatic mutated peptides by MS from the proteome of patients with lung cancer, including a novel somatic mutated CDK12_{p.G879V} peptide using similar methodology (19). Although we identified 19 peptide epitopes derived from common oncogenes, none of them contained somatic mutations (supplemental Fig. S3F). Possible explanation of the identification of relatively few mutant peptides from known oncogene mutations include the absence of the cognate class I allele and limitations of the data-dependent acquisition methodology for MS-based sequencing of peptides when the mutated peptides are just a minor fraction of the total wildtype peptides presented by HLA class I. However, interestingly, the median MS intensity of the few neoantigens we identified is very similar to that of wildtype peptides (p value > 0.05), indicating that select neoantigens are robustly presented by class I.

We have generated the most comprehensive CG antigen DB to date by leveraging the human proteome atlas and multiple peer-reviewed CG antigen DBs. This is a valuable resource that can be used to query CG antigens from other large cohort immunopeptidome studies. For instance, we identified LDHC (supplemental Fig. S5I) which, prior to this study, was almost exclusively observed in testis, and an association with lung adenocarcinoma had only been suggested (51). Given that CG antigens have been extensively investigated in melanoma (73, 74), we have now described the CG antigen landscape in EGFR-mutant lung cancer, unveiling novel class I-presented peptides reported in our study. As previously reported, our results also showed that CG antigen-derived peptide levels are not significantly correlated with mRNA and source protein expression (40, 75, 76).

One of the advantages of a PEAKS DB search is its pan-PTM search engine, which unveils all PTMs in one search, requiring no prior knowledge of the potential types of modifications expected. PeaksPTM uses the sequence-tag approach to identify PTMs without generating an extremely large sequence library containing variable modifications (77, 78). This allows identification of many PTMs in one search. To our knowledge, this study provides the deepest coverage of

the PTM HLA class I immunopeptidome, to date. Compared with previous profiling of PTM immunopeptides (12, 14), we, for the first time, systematically quantified the largest number of endogenous PTMs of class I-presented peptides in EGFR-mutant lung adenocarcinoma. We demonstrated that unmodified peptides were significantly more abundant than their deamidated or methylated counterparts. However, this does not apply to the conversion of glutamate to pyroGlu, where a modification could lead to protein misfolding (55). Generally, low peptide abundance or presentation in antigen-presenting cells may hinder T-cell recognition. It is possible that more abundant peptides are more likely represented as T-cell epitopes. Although phosphorylation is the most dominant PTM known to modulate cellular function, we identified only a minor fraction of class I-presented peptides phosphorylated; this could be due to the relatively large phosphate group, which may not easily fit into the HLA-binding groove. The binding of only around two thousand unique phosphopeptides to 72 HLA alleles has been reported (79). The transient reversible PTM phosphorylation generates less-stable peptides that are less likely to be good targets for HLA class I presentation. This is in contrast to irreversible PTMs such as deamidation, which are more likely to generate neoepitopes. We identified 20 deamidated peptides with a NX(S/T) motif, which supports the potential mechanism of asparagine deamidation that deamidated immunopeptides are derived from deglycosylation (80). We discovered that a single peptide could have multiple PTMs concurrently (Fig. 4D). Position 1 was the most frequently modified (Fig. 4C), which implies that small modifications on a nonanchor position may not dramatically affect binding affinity. We compared the predicted binding score of the modified-only peptides with the peptides that were identified as both modified and unmodified versions. The modified-only peptides have lower binding scores using NetMHCpan prediction (Fig. 4, E and F), indicating these peptides may possess a unique HLA-binding domain structure that is modification dependent and cannot be analyzed by available prediction algorithms. However, we acknowledge that a limited fraction of PTM peptides may occur spontaneously *in vitro*, although the sample preparation was conducted at 4 °C. For example, deamidated peptides with an NG motif have been reported as *in vitro* artifacts (81). Taken together, our results demonstrate that the PTM peptides that are not predicted by any bioinformatics algorithm for HLA binding can be identified using MS and these peptides may be a rich source of potential neoepitopes.

Deep learning-based *de novo* search algorithm embedded in PEAKS provided high accuracy to detect peptides that normally are missed in a DB search. Our pipeline evaluates in-depth the *de novo*-only peptides in class I immunopeptidome context. The validity of this approach for HLA peptidome profiling and variant peptide identification is underscored by several findings in this study. Although the identified *de novo*-only peptides may constitute many spuriously identified

peptides, which are not predicted to be binding to any HLA allele in respective samples, we believe that PEAKS *de novo* search unveiled reasonable fraction of true HLA class I-presented peptides considering no decoy DB is available to control the FDR (Fig. 5, A and B). We acknowledge that the correlation of SSRCalc-predicted HI and RT of tryptic peptides is higher than that of immunopeptides (supplemental Fig. S7, B–D). This indicated potential higher FDR in *de novo* search for HLA peptides. The lower correlation could also be a result of inaccurate prediction of HI by SSRCalc for nontryptic peptides, such as the HLA class I-presented 9-mer immunopeptides. SSRCalc was originally designed for tryptic peptides, which end with either lysine or arginine and always has been reduced and alkylated before LC–MS/MS analysis (82). Also, nontryptic HLA peptides usually have shorter length (8- to 12-mer) than tryptic peptides (8- to 25-mer). PEAKS overcomes this high FDR problem by using ALC scoring for peptide identification (83). The binding motif of the 9-mer *de novo*-only peptides was very similar to the DB-searched peptides (Fig. 5, D–H). MS/MS sequencing spectra of select variant peptides in H1975 identified from the *de novo* search and proteogenomic DB search were a perfect match to the corresponding synthetic peptides (Fig. 5I and supplemental Fig. S7, E–G). Half of the *de novo*-only searched variant peptides from H1975 and PC9 identified by the proteogenomic pipeline were also identified using our customized human proteome DB (Fig. 5J). We further validated the variant peptides identified by *de novo* search algorithm by matching the predicted HI and RT of the select four variant peptides from H1975 that were also validated by spectra matching with synthetic peptides. Further confirmation and validation of the variant and PTM peptides identified by *de novo* sequencing by the PEAKS search algorithm can be obtained by spiking synthetic heavy-labeled peptides and demonstrating coelution of endogenously identified peptides with the heavy-labeled synthetic peptides. Overall, our results suggest that *de novo* search by PEAKS identifies high-quality class I-presented peptides and is a powerful tool to identify noncanonical peptides/proteins for proteogenomic profiling.

Noncoding RNA is recognized as a rich resource for neoantigens (84), and MS-based proteogenomic platforms have been implemented to discover noncanonical peptides (15, 57). Laumont *et al.* (58) suggested that noncoding regions are the main source of neoantigens. In contrast to previous studies where MS spectra were mapped *in silico* to all potential reading frames of FASTA-derived RNA-Seq data, we leveraged deep learning methodology of *de novo* sequencing of peptides in PEAKS studio to query against the largest annotated lncRNA DB. Our approach extends the previous studies in many ways. One of the challenges in proteogenomic field is that the addition of genomic sequencing *in silico*-translated DB can result in a higher chance of false identifications (85). Since LNCipedia DB consolidates many datasets from a variety of sources and contains bidirectional

lncRNAs in which transcription can be initiated from both ends (33), we constructed a six-frame-translated DB derived from the lncRNA DB and interrogated potentially matched *de novo*-only searched peptides from the PEAKS search engine. As has been reviewed before, false-positive matched peptides increase upon interrogation of such large DBs (86). However, the highly significant *p* values obtained by our empirical *p* value calculation, similar to a permutation test, on the mock lncRNA DB, performed exactly the same way considering all six reading frames for a potential peptide match reduces the possibility that there are significant false positives in the identified lncRNA-derived immunopeptides. In addition, we visualized and manually verified that the lncRNA-derived peptide coding regions do not overlap with any protein-coding regions, introns of known protein-coding genes, or could result from a frameshift. We also selectively validated the presence of the source lncRNAs in the ribosome profiling data searching the GWIPS DB, suggesting that these lncRNAs are indeed translated on the ribosome machinery. To further ensure that the identified lncRNA-derived peptides were indeed not a part of a known expressed protein, we manually checked and visualized that the source lncRNA of each of these novel peptides was indeed present in the total RNA-Seq data from the same cell line and tumor and did not overlap with any coding region or was not because of a frameshift. Synthetic peptide validation and T2 cell HLA-binding assays further confirmed that the identified lncRNA-derived peptides were indeed presented and had high affinity to the HLA proteins. Taken together, our results show that lncRNA-derived immunopeptides are presented by HLA class I in tumors with both high and low TMB. The pipeline we developed in this study could be readily applied to any type of cancer to identify lncRNA-derived peptides presented by HLA class I.

Taken together, we report the largest characterization of potential cancer-associated class I immunopeptidome in EGFR-mutant lung adenocarcinoma to date. The combination of genomics, proteomics, and informatics allows us to develop this in-depth immunopeptidome-based cancer epitope profiling pipeline. Our results suggest that low TMB tumors possess as many potential immunotherapy-targetable epitopes as high TMB tumors, the identification of which needs in-depth analysis of MS data. We provide a valuable resource of EGFR-mutant lung cancer-specific neoepitopes as well as tumor-associated immunopeptides for possible design of precision immunotherapy and cancer vaccines.

DATA AVAILABILITY

The raw MS-based sequencing files of HLA class I immunopeptidome and whole-cell proteome and patient-/cell line-specific FASTA files have been deposited to ProteomeXchange *via* PRIDE. Data can be retrieved with the identifier PXD022949. All annotated MS/MS spectra of MaxQuant

searched results can be found on MS-viewer (<https://msviewer.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msviewer>) using search key “yqns2kpwjw.”

The genomic and transcriptomic data have been published by our group, and the original files can be accessed *via* our previous publications (16, 18).

Supplemental data—This article contains [supplemental data](#).

Acknowledgments—We thank Drs Paul F. Robbins and Maria R. Parkhurst (Surgery Branch, CCR, NCI, NIH) for their generous assistance in melanoma CG peptides synthesis and helpful discussions in melanoma CG peptide identification. We also thank Dr Jared Gartner (Surgery Branch, CCR, NCI, NIH) for providing assistance in analysis of RNA-Seq datasets of the melanoma cell lines. This study was supported by the Intramural Research Program of CCR, NCI of the United States National Institutes of Health.

Funding and additional information—The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions—Y. A. Q. and U. G. conceptualization; Y. A. Q. and U. G. methodology; Y. A. Q., V. M., M. H. E., and C. S. software; Y. A. Q., T. K. M., K. D. N., C. A., K.-i. H., J. C. Y., X. Z., V. M., and S. G. investigation; Y. A. Q., T. K. M., K. D. N., and C. A. resources; D. M. and J. K. data curation; Y. A. Q. writing-original draft; Y. A. Q., T. K. M., C. M. C., V. M., X. Z., C. A., S. G., D. M., K. D. N., M. H. E., K.-i. H., J. K., C. S., J. C. Y., and U. G. writing-review and editing; Y. A. Q., and U. G. supervision.

Conflict of interest—U. G. has a clinical trial agreement with AstraZeneca and had received research funding from AstraZeneca, Aurigene, and Esanex. U. G. is currently an employee of Bristol-Myers Squibb. The other authors declare no competing interests.

Abbreviations—The abbreviations used are: ACN, acetonitrile; ACT, adoptive T-cell therapy; ALC, average local confidence; ATCC, American Type Culture Collection; BLAT, BLAST-like Alignment Tool; CCR, Center for Cancer Research; CG, cancer germ line; CIP2A, cellular inhibitor of PP2A; COSMIC, Catalogue of Somatic Mutations in Cancer; DB, database; DMSO, dimethyl sulfoxide; EGFR, epidermal growth factor receptor; EIF, eukaryotic initiation factor; FBS, fetal bovine serum; FDR, false discovery rate; GWIPS, Genome-Wide Information on Protein Synthesis; HI, hydrophobicity index; HLA, human leukocyte antigen; IEDB, Immune Epitope Database and Analysis Resource; IGV, Integrative Genomics Viewer; INDEL, insertion and deletion; IPA, ingenuity pathway analysis; LDHC, lactate dehydrogenase C; lncRNA, long noncoding RNA; MS/MS, tandem MS;

NCI, National Cancer Institute; NGS, next-generation sequencing; PTM, post-translationally modified; pyroGlu, pyroglutamate; RT, retention time; SNV, single nucleotide variant; SSP, sequence-specific primer; TAP, transporter associated with antigen processing; TMB, tumor mutational burden; VCF, variant call format; WES, whole-exome sequencing.

Received December 16, 2020, and in revised form, August 3, 2021
Published, MCPRO Papers in Press, August 13, 2021, <https://doi.org/10.1016/j.mcpro.2021.100136>

REFERENCES

- Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., *et al.* (2019) Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206
- Goodman, A. M., Kato, S., Bazhenova, L., Patel, S. P., Frampton, G. M., Miller, V., Stephens, P. J., Daniels, G. A., and Kurzrock, R. (2017) Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* **16**, 2598–2608
- Sha, D., Jin, Z., Budczies, J., Kluck, K., Stenzinger, A., and Sinicrope, F. A. (2020) Tumor mutational burden as a predictive biomarker in solid tumors. *Cancer Discov.* **10**, 1808–1825
- Offin, M., Rizvi, H., Tenet, M., Ni, A., Sanchez-Vega, F., Li, B. T., Drilon, A., Kris, M. G., Rudin, C. M., Schultz, N., Arcila, M. E., Ladanyi, M., Riely, G. J., Yu, H., and Hellmann, M. D. (2019) Tumor mutation burden and efficacy of EGFR-tyrosine kinase inhibitors in patients with EGFR-mutant lung cancers. *Clin. Cancer Res.* **25**, 1063–1069
- Dong, Z. Y., Zhang, J. T., Liu, S. Y., Su, J., Zhang, C., Xie, Z., Zhou, Q., Tu, H. Y., Xu, C. R., Yan, L. X., Li, Y. F., Zhong, W. Z., and Wu, Y. L. (2017) EGFR mutation correlates with uninfamed phenotype and weak immunogenicity, causing impaired response to PD-1 blockade in non-small cell lung cancer. *Oncoimmunology* **6**, e1356145
- Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., *et al.* (2015) Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128
- Topalian, S. L., Sznol, M., McDermott, D. F., Kluger, H. M., Carvajal, R. D., Sharfman, W. H., Brahmer, J. R., Lawrence, D. P., Atkins, M. B., Powderly, J. D., Leming, P. D., Lipson, E. J., Puzanov, I., Smith, D. C., Taube, J. M., *et al.* (2014) Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J. Clin. Oncol.* **32**, 1020–1030
- Tran, E., Robbins, P. F., Lu, Y. C., Prickett, T. D., Gartner, J. J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E. M., Kriley, I. R., and Rosenberg, S. A. (2016) T-cell transfer therapy targeting mutant KRAS in cancer. *N. Engl. J. Med.* **375**, 2255–2262
- Kerker, S. P., Wang, Z. F., Lasota, J., Park, T., Patel, K., Groh, E., Rosenberg, S. A., and Miettinen, M. M. (2016) MAGE-A is more highly expressed than NY-ESO-1 in a systematic immunohistochemical analysis of 3668 cases. *J. Immunother.* **39**, 181–187
- Zhang, X., Qi, Y., Zhang, Q., and Liu, W. (2019) Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomed. Pharmacother.* **120**, 109542
- Caron, E., Kowalewski, D. J., Chiek Koh, C., Sturm, T., Schuster, H., and Aebbersold, R. (2015) Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol. Cell Proteomics* **14**, 3105–3117
- Mommen, G. P., Frese, C. K., Meiring, H. D., van Gaans-van den Brink, J., de Jong, A. P., van Els, C. A., and Heck, A. J. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (ETcD). *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4507–4512
- Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K.

- R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326
14. Bassani-Sternberg, M., Braunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., D. H. B., Peschel, C., *et al.* (2016) Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404
 15. Chong, C., Muller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B. J., Michaux, J., Bilic, I., Hirsekorn, A., Calviello, L., Simo-Riudalbas, L., *et al.* (2020) Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293
 16. Roper, N., Brown, A. L., Wei, J. S., Pack, S., Trindade, C., Kim, C., Restifo, O., Gao, S., Sindiri, S., Mehrabadi, F., El Meskini, R., Ohler, Z. W., Maity, T. K., Venugopalan, A., Cultraro, C. M., *et al.* (2020) Clonal evolution and heterogeneity of osimertinib acquired resistance mechanisms in EGFR mutant lung cancer. *Cell Rep. Med.* **1**, 100007
 17. Roper, N., Gao, S., Maity, T. K., Banday, A. R., Zhang, X., Venugopalan, A., Cultraro, C. M., Patidar, R., Sindiri, S., Brown, A. L., Goncarenco, A., Panchenko, A. R., Biswas, R., Thomas, A., Rajan, A., *et al.* (2019) APO-BEC mutagenesis and copy-number alterations are drivers of proteogenomic tumor evolution and heterogeneity in metastatic thoracic tumors. *Cell Rep* **26**, 2651–2666.e2656
 18. Gros, A., Parkhurst, M. R., Tran, E., Pasetto, A., Robbins, P. F., Ilyas, S., Prickett, T. D., Gartner, J. J., Crystal, J. S., Roberts, I. M., Trebska-McGowan, K., Wunderlich, J. R., Yang, J. C., and Rosenberg, S. A. (2016) Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat. Med.* **22**, 433–438
 19. Zhang, X., Nguyen, K. D., Rudnick, P. A., Roper, N., Kawaler, E., Maity, T. K., Awasthi, S., Gao, S., Biswas, R., Venugopalan, A., Cultraro, C. M., Fenyo, D., and Guha, U. (2019) Quantitative mass spectrometry to interrogate proteomic heterogeneity in metastatic lung adenocarcinoma and validate a novel somatic mutation CDK12-G879V. *Mol. Cell Proteomics* **18**, 622–641
 20. Ren, S., Ahmed, N., Bertels, K., and Al-Ars, Z. (2019) GPU accelerated sequence alignment with traceback for GATK HaplotypeCaller. *BMC Genomics* **20**, 184
 21. Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., and Lu, X. (2012) Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new Program, SnpSift. *Front Genet.* **3**, 35
 22. Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213
 23. Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M. D., Clauser, K. R., Tabb, D. L., Mertins, P., Slebos, R., Erdmann-Gilmore, P., Li, S., Gunawardena, H. P., *et al.* (2016) An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell Proteomics* **15**, 1060–1071
 24. Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. (2019) Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**, 63–66
 25. Boegel, S., Lower, M., Schafer, M., Bukur, T., de Graaf, J., Boisguerin, V., Tureci, O., Diken, M., Castle, J. C., and Sahin, U. (2012) HLA typing from RNA-Seq sequence reads. *Genome Med.* **4**, 102
 26. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343
 27. Jurtz, V., Paul, S., Andreatta, M., Marcattili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368
 28. Colaert, N., Helsen, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by ice-Logo. *Nat. Methods* **6**, 786–787
 29. Krokhin, O. V., and Spicer, V. (2009) Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal. Chem.* **81**, 9522–9530
 30. Trapnell, C., Pachter, L., and Salzberg, S. L. (2009) TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111
 31. Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012) Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817
 32. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550
 33. Volders, P. J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019) LNCipedia 5: Towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* **47**, D135–D139
 34. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
 35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079
 36. Quinlan, A. R., and Hall, I. M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842
 37. Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192
 38. Michel, A. M., Fox, G., A. M. K., De Bo, C., O'Connor, P. B., Heaphy, S. M., Mulligan, J. P., Donohue, C. A., Higgins, D. G., and Baranov, P. V. (2014) GWIPS-viz: Development of a ribo-seq genome browser. *Nucleic Acids Res.* **42**, D859–864
 39. Kramer, A., Green, J., Pollard, J., Jr., and Tugendreich, S. (2014) Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530
 40. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015) Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell Proteomics* **14**, 658–673
 41. Williams, T. M. (2001) Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J. Mol. Diagn.* **3**, 98–104
 42. Bjorkman, P. J. (1997) MHC restriction in three dimensions: A view of T cell receptor/ligand interactions. *Cell* **89**, 167–170
 43. Koopman, L. A., Corver, W. E., van der Slik, A. R., Giphart, M. J., and Fleuren, G. J. (2000) Multiple genetic alterations cause frequent and heterogeneous human histocompatibility leukocyte antigen class I loss in cervical cancer. *J. Exp. Med.* **191**, 961–976
 44. Chowell, D., Morris, L. G. T., Grigg, C. M., Weber, J. K., Samstein, R. M., Makarov, V., Kuo, F., Kendall, S. M., Requena, D., Riaz, N., Greenbaum, B., Carroll, J., Garon, E., Hyman, D. M., Zehir, A., *et al.* (2018) Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* **359**, 582–587
 45. Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O., and Rosales-Mendoza, S. (2015) An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. *J. Biomed. Inform.* **53**, 405–414
 46. Hansen, T., and Myers, N. (2003) Peptide induction of surface expression of class I MHC. *Curr. Protoc. Immunol*
 47. Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., and Chen, Y. T. (2002) Cancer/testis antigens: An expanding family of targets for cancer immunotherapy. *Immunol. Rev.* **188**, 22–32
 48. Janitz, M., Fiszer, D., Michalczak-Janitz, K., Lukaszyk, A., Fernandez, N., Skorpinski, W., and Kurpisz, M. (1994) Analysis of mRNA for class I HLA on human gametogenic cells. *Mol. Reprod. Dev.* **38**, 231–237
 49. Gjerstorff, M. F., Andersen, M. H., and Ditzel, H. J. (2015) Oncogenic cancer/testis antigens: Prime candidates for immunotherapy. *Oncotarget* **6**, 15772–15787
 50. Lorient, A., Boon, T., and De Smet, C. (2003) Five new human cancer-germline genes identified among 12 genes expressed in spermatogonia. *Int. J. Cancer* **105**, 371–376
 51. Yen, C. C., Liang, S. C., Jong, Y. J., Chen, Y. J., Lin, C. H., Chen, Y. M., Wu, Y. C., Su, W. C., Huang, C. Y., Tseng, S. W., and Whang-Peng, J. (2007) Chromosomal aberrations of malignant pleural effusions of lung adenocarcinoma: Different cytogenetic changes are correlated with genders and smoking habits. *Lung Cancer* **57**, 292–301

52. Bourdetsky, D., Schmelzer, C. E., and Admon, A. (2014) The nature and extent of contributions by defective ribosome products to the HLA peptidome. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1591–1599
53. Milner, E., Gutter-Kapon, L., Bassani-Strenberg, M., Barnea, E., Beer, I., and Admon, A. (2013) The effect of proteasome inhibition on the generation of the human leukocyte antigen (HLA) peptidome. *Mol. Cell Proteomics* **12**, 1853–1864
54. Berg, M., Parbel, A., Pettersen, H., Fenyo, D., and Bjorkesten, L. (2006) Detection of artifacts and peptide modifications in liquid chromatography/mass spectrometry data using two-dimensional signal intensity map data visualization. *Rapid Commun. Mass Spectrom.* **20**, 1558–1562
55. Liu, Y. D., Goetze, A. M., Bass, R. B., and Flynn, G. C. (2011) N-terminal glutamate to pyroglutamate conversion *in vivo* for human IgG2 antibodies. *J. Biol. Chem.* **286**, 11211–11217
56. Tran, N. H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017) De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8247–8252
57. Erhard, F., Dolken, L., Schilling, B., and Schlosser, A. (2020) Identification of the cryptic HLA-I immunopeptidome. *Cancer Immunol. Res.* **8**, 1018–1026
58. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, E., Bonnell, E., Laverdure, J. P., Gendron, P., Courcelles, M., Hardy, M. P., Cote, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., *et al.* (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**
59. Laumont, C. M., Daouda, T., Laverdure, J. P., Bonnell, E., Caron-Lizotte, O., Hardy, M. P., Granados, D. P., Durette, C., Lemieux, S., Thibault, P., and Perreault, C. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238
60. Maston, G. A., Evans, S. K., and Green, M. R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59
61. Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J. L., Wong, Y. N. S., Rowan, A., Kanu, N., Al Bakir, M., Chambers, T., Salgado, R., Savas, P., Loi, S., Birkbak, N. J., *et al.* (2017) Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: A pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021
62. Zacharakis, N., Chinnasamy, H., Black, M., Xu, H., Lu, Y. C., Zheng, Z., Pasetto, A., Langhan, M., Shelton, T., Prickett, T., Gartner, J., Jia, L., Trebska-McGowan, K., Somerville, R. P., Robbins, P. F., *et al.* (2018) Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat. Med.* **24**, 724–730
63. Leko, V., McDuffie, L. A., Zheng, Z., Gartner, J. J., Prickett, T. D., Apolo, A. B., Agarwal, P. K., Rosenberg, S. A., and Lu, Y. C. (2019) Identification of neoantigen-reactive tumor-infiltrating lymphocytes in primary bladder cancer. *J. Immunol.* **202**, 3458–3467
64. Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., *et al.* (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221
65. Veatch, J. R., Jesernig, B. L., Kargl, J., Fitzgibbon, M., Lee, S. M., Baik, C., Martins, R., Houghton, A. M., and Riddell, S. R. (2019) Endogenous CD4(+) T cells recognize neoantigens in lung cancer patients, including recurrent oncogenic KRAS and ERBB2 (Her2) driver mutations. *Cancer Immunol. Res.* **7**, 910–922
66. Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., and Fugmann, T. (2016) High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics* **16**, 1570–1580
67. Gloger, A., Ritz, D., Fugmann, T., and Neri, D. (2016) Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. *Cancer Immunol. Immunother.* **65**, 1377–1393
68. Anagnostou, V., Smith, K. N., Forde, P. M., Niknafs, N., Bhattacharya, R., White, J., Zhang, T., Adleff, V., Phallen, J., Wali, N., Hruban, C., Guthrie, V. B., Rodgers, K., Naidoo, J., Kang, H., *et al.* (2017) Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov.* **7**, 264–276
69. Chen, N., Fang, W., Zhan, J., Hong, S., Tang, Y., Kang, S., Zhang, Y., He, X., Zhou, T., Qin, T., Huang, Y., Yi, X., and Zhang, L. (2015) Upregulation of PD-L1 by EGFR activation mediates the immune escape in EGFR-driven NSCLC: Implication for optimal immune targeted therapy for NSCLC patients with EGFR mutation. *J. Thorac. Oncol.* **10**, 910–923
70. Saigi, M., Alburquerque-Bejar, J. J., and Sanchez-Cespedes, M. (2019) Determinants of immunological evasion and immuncheckpoint inhibition response in non-small cell lung cancer: The genetic front. *Oncogene* **38**, 5921–5932
71. Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachiredy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., *et al.* (2019) A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.*
72. Klatt, M. G., Kowalewski, D. J., Schuster, H., Di Marco, M., Hennenlotter, J., Stenzl, A., Rammensee, H. G., and Stevanovic, S. (2016) Carcinogenesis of renal cell carcinoma reflected in HLA ligands: A novel approach for synergistic peptide vaccination design. *Oncimmunology* **5**, e1204504
73. Danilova, A., Misyurin, V., Novik, A., Giryuk, D., Avdonkina, N., Nekhaeva, T., Emelyanova, N., Pipia, N., Misyurin, A., and Baldueva, I. (2020) Cancer/testis antigens expression during cultivation of melanoma and soft tissue sarcoma cells. *Clin. Sarcoma Res.* **10**, 3
74. Mendiratta, S. K., Thai, G., Eslahi, N. K., Thull, N. M., Matar, M., Bronte, V., and Pericle, F. (2001) Therapeutic tumor immunity induced by poly-immunization with melanoma antigens gp100 and TRP-2. *Cancer Res.* **61**, 859–863
75. Shraibman, B., Kadosh, D. M., Barnea, E., and Admon, A. (2016) Human leukocyte antigen (HLA) peptides derived from tumor antigens induced by inhibition of DNA methylation for development of drug-facilitated immunotherapy. *Mol. Cell Proteomics* **15**, 3058–3070
76. Weinzierl, A. O., Lemmel, C., Schoor, O., Muller, M., Kruger, T., Wernet, D., Hennenlotter, J., Stenzl, A., Klingel, K., Rammensee, H. G., and Stevanovic, S. (2007) Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface. *Mol. Cell Proteomics* **6**, 102–113
77. Han, X., He, L., Xin, L., Shan, B., and Ma, B. (2011) PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *J. Proteome Res.* **10**, 2930–2936
78. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399
79. Solleder, M., Guillaume, P., Racle, J., Michaux, J., Pak, H. S., Muller, M., Coukos, G., Bassani-Sternberg, M., and Gfeller, D. (2020) Mass spectrometry based immunopeptidomics leads to robust predictions of phosphorylated HLA class I ligands. *Mol. Cell Proteomics* **19**, 390–404
80. Mei, S., Ayala, R., Ramarathinam, S. H., Illing, P. T., Faridi, P., Song, J., Purcell, A. W., and Croft, N. P. (2020) Immunopeptidomic analysis reveals that deamidated HLA-bound peptides arise predominantly from deglycosylated precursors. *Mol. Cell Proteomics* **19**, 1236–1247
81. Tyler-Cross, R., and Schirch, V. (1991) Effects of amino acid sequence, buffers, and ionic strength on the rate and mechanism of deamidation of asparagine residues in small peptides. *J. Biol. Chem.* **266**, 22549–22556
82. Krokhin, O. V., and Spicer, V. (2010) Predicting peptide retention times for proteomics. *Curr. Protoc. Bioinformatics*
83. Faridi, P., Li, C., Ramarathinam, S. H., Illing, P. T., Mifsud, N. A., Ayala, R., Song, J., Gearing, L. J., Croft, N. P., and Purcell, A. W. (2019) Response to Comment on "A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands". *Sci. Immunol.* **4**
84. Day, S. L., Ramsland, P. A., and Apostolopoulos, V. (2009) Non-canonical peptides bound to MHC. *Curr. Pharm. Des.* **15**, 3274–3282
85. Sheynkman, G. M., Shortreed, M. R., Cesnik, A. J., and Smith, L. M. (2016) Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **9**, 521–545
86. Castellana, N., and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteomics* **73**, 2124–2135