

pXg: Comprehensive Identification of Noncanonical MHC-I-Associated Peptides From *De Novo* Peptide Sequencing Using RNA-Seq Reads

Authors

Seunghyuk Choi, and Eunok Paek

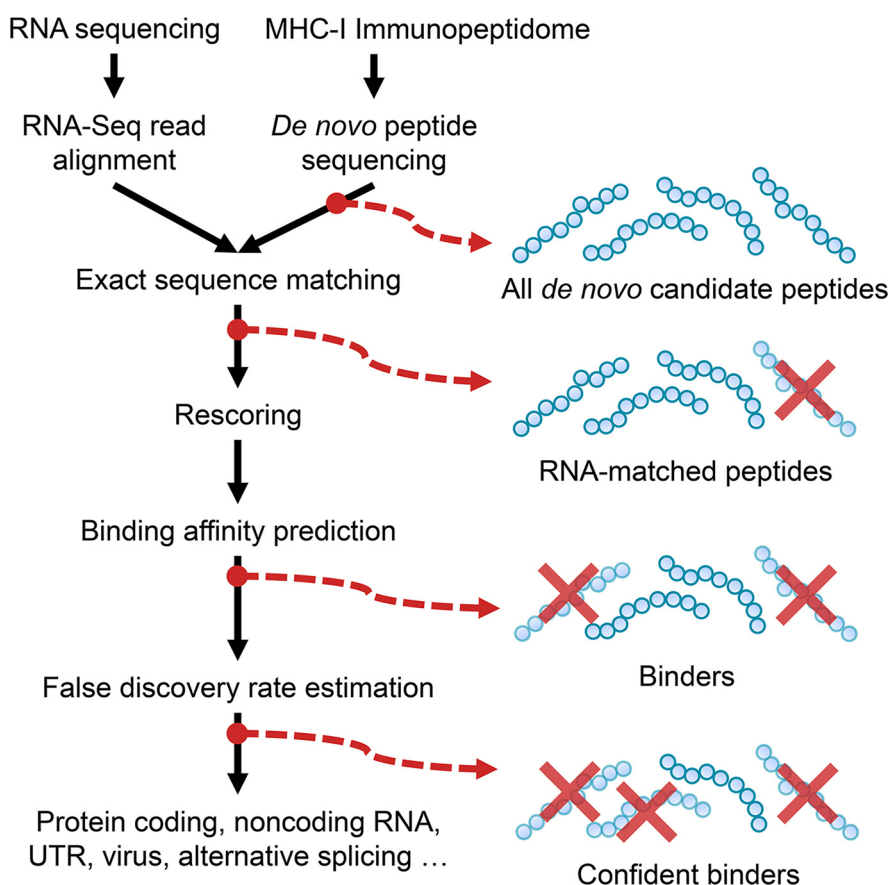
Correspondence

eunokpaek@hanyang.ac.kr

In Brief

Identification of noncanonical peptides is an essential task for discovery novel immunotherapeutic targets. Due to the limited search space by prior knowledge, database searches struggle to identify comprehensive noncanonical peptides. In order to completely and accurately identify immunopeptides using both *de novo* peptide sequencing and RNA-Seq reads, we describe a unique pipeline called pXg. We show that pXg can confidently identify noncanonical peptides, including viral proteins in immunopeptidomics. It demonstrates its potential use in the search for new immunotherapeutic targets.

Graphical Abstract



Highlights

- We implemented pXg to confidently identify *de novo* peptides using RNA-Seq reads.
- pXg incorporates proteogenomic features into Percolator in immunopeptidomics.
- RNA-Seq quantity/quality has a clear relationship with PSM quality.
- pXg enables identifying every potential immunopeptidome including viral proteins.

pXg: Comprehensive Identification of Noncanonical MHC-I-Associated Peptides From *De Novo* Peptide Sequencing Using RNA-Seq Reads

Seunghyuk Choi¹  and Eunok Paek^{1,2,*} 

Discovering noncanonical peptides has been a common application of proteogenomics. Recent studies suggest that certain noncanonical peptides, known as noncanonical major histocompatibility complex-I (MHC-I)-associated peptides (ncMAPs), that bind to MHC-I may make good immunotherapeutic targets. *De novo* peptide sequencing is a great way to find ncMAPs since it can detect peptide sequences from their tandem mass spectra without using any sequence databases. However, this strategy has not been widely applied for ncMAP identification because there is not a good way to estimate its false-positive rates. In order to completely and accurately identify immunopeptides using *de novo* peptide sequencing, we describe a unique pipeline called proteomics X genomics. In contrast to current pipelines, it makes use of genomic data, RNA-Seq abundance and sequencing quality, in addition to proteomic features to increase the sensitivity and specificity of peptide identification. We show that the peptide-spectrum match quality and genetic traits have a clear relationship, showing that they can be utilized to evaluate peptide-spectrum matches. From 10 samples, we found 24,449 canonical MHC-I-associated peptides and 956 ncMAPs by using a target-decoy competition. Three hundred eighty-seven ncMAPs and 1611 canonical MHC-I-associated peptides were new identifications that had not yet been published. We discovered 11 ncMAPs produced from a squirrel monkey retrovirus in human cell lines in addition to the two ncMAPs originating from a complementarity determining region 3 in an antibody thanks to the unrestricted search space assumed by *de novo* sequencing. These entirely new identifications show that proteomics X genomics can make the most of *de novo* peptide sequencing's advantages and its potential use in the search for new immunotherapeutic targets.

Major histocompatibility complex I (MHC-I) molecules are cell surface glycoproteins that present peptide antigens and

act as key factors in CD8⁺ T-cell recognition, which initiates immune responses against pathogens. The accurate typing of MHC-I classes and the identification of MHC-I-associated peptides (MAPs) that bind to these MHC-I molecules are essential for understanding immune responses and can greatly contribute to the development of cancer immunotherapies.

Recent advancements have significantly improved the accuracy of MHC-I class typing approximately to 0.95 (1–3). However, identification of MAPs presents challenges due to their diversity, which includes noncanonical peptides originating from various aberrant transcriptional or translational events. These events encompass alternative RNA splicing, frameshifts, and translation of noncoding RNAs (ncRNAs), UTRs, intronic regions (IRs), intergenic regions, and even antisense RNAs (4–10). The fact that there were an order of magnitude less identified noncanonical MAPs (ncMAPs) per sample compared to canonical MAPs (cMAPs) suggests that identifying ncMAPs can be particularly challenging. However, because noncanonical peptides exhibit tissue- and sample-specific expression patterns, they hold promise as valuable targets for precision medicine in immunotherapy (11).

Because next-generation sequencing and mass spectrometry (MS)-based MHC-I immunopeptidomics are the only analytical methods that can identify a significant number of ncMAPs at once, previous research has used a variety of bioinformatics workflows to find ncMAPs following a proteogenomic approach. For instance, Laumont and colleagues constructed a custom sequence database (4) using peptides with lengths of 8 to 11 amino acids, supported by at least 10 RNA reads, and then employed Mascot (12) for peptide identification from tandem mass spectra. Similarly, Cuevas and colleagues generated a sequence database using Ribo-Seq (10) and used PEAKS DB (13) for the search. These so-called “genomics-to-proteomics” methods have commonly

From the ¹Department of Computer Science, Hanyang University, Seoul, Republic of Korea; ²Institute for Artificial Intelligence Research, Hanyang University, Seoul, Republic of Korea

*For correspondence: Eunok Paek, eunokpaek@hanyang.ac.kr.

been used to identify ncMAPs but are known for their limited sensitivity and high false-positive rates, partly due to the large size of genomics-based databases. Furthermore, if the target database was built entirely based on prior knowledge, it cannot discover noncanonical peptides, such as those arising from unexpected viral infections.

On the other hand, “proteomics-to-genomics” techniques using *de novo* peptide sequencing offer a promising approach for ncMAP identification. *De novo* methods are known for their accuracy in inferring shorter peptides from mass spectra (14), making them an attractive tool for discovering ncMAPs, typically ranging from 8 to 15 amino acids. Some studies (15, 16) have adopted “proteomics-to-genomics” strategies using PEAKS (<https://www.bioinform.com/peaks-studio>) (17), well-known commercial *de novo* peptide sequencing software, and matching peptides to target/decoy sequence database to estimate false discovery rate (FDR). These methods do not need RNA-Seq information and can be used with just MS-based immunopeptidomics. As Erhard and colleagues demonstrated, however, mutant peptides tend to show greater FDRs, making them ineffective for identifying tumor-specific antigens. Additionally, their reliance on presumptive sequence databases results in constraints similar to “genomics-to-proteomics” methods, which are dependent on prior knowledge of potential MAPs.

In general, peptide spectrum match (PSM) rescoring and FDR estimation strategies are critical for improving the sensitivity and the specificity of peptide identification. Because scoring functions of PSMs are not optimized for immunopeptidome data (18), re-evaluating PSMs based on data characteristics can enhance MAP identification in a data-dependent manner. Percolator (19) has been widely used to rescore PSMs in global proteomics, and recent studies have demonstrated that integrating deep learning-based features, such as predicted spectrum similarity and retention time (RT), can improve immunopeptide identification rates (20, 21). For FDR estimation, tools like MHCquant (22) and MAPDP (23) have employed subset FDR (24) by filtering MHC-I-bound peptides predicted by various tools, such as NetMHCpan (25) and MHCflurry (26). This approach yielded more binder peptides by filtering out unnecessary peptides (nonbinders). It has been emphasized that canonical and noncanonical PSMs should be separately controlled for precise FDR estimation (27, 28). For example, Woo and colleagues (28) showed that under 1% FDR estimation, combined FDR resulted in severely underestimated FDR for noncanonical peptides (36% FDR) and overestimated FDR for canonical peptides (0.03% FDR). Despite these findings, many immunopeptidomic workflows, including MHCquant and MAPDP, have overlooked this distinction and applied combined FDR estimation. The discovery of immunotherapy targets (tumor-associated or tumor-specific antigens) requires the confident identification of ncMAPs, including mutant peptides. Therefore, it is imperative to develop comprehensive and reliable identification methods

for ncMAPs that are not restricted by prior knowledge or database.

For comprehensive and reliable discovery of ncMAPs, we developed proteomics X genomics (pXg), a pipeline that enables the confident identification of both cMAPs and ncMAPs from *de novo* peptide sequencing, incorporating RNA-Seq data to address the aforementioned challenges. In our quest to improve both sensitivity and specificity in immunopeptide identifications, pXg utilizes Percolator. For the first time, we employed genomic features, RNA-Seq abundance and sequencing quality, in addition to the typical proteomic features used as input to Percolator for rescoring immunopeptide sequencing. Our results demonstrate that these newly introduced genomic characteristics significantly impact target/decoy classification using 10 immunopeptidome tandem mass spectrometry (MS/MS) and RNA-Seq datasets. Furthermore, pXg offers comprehensive identification of ncMAPs, including rare events that are particularly challenging to infer from the reference sequence database. Across the 10 samples, we identified 24,449 cMAPs and 956 ncMAPs, with 1611 cMAPs and 387 ncMAPs representing new identifications not reported elsewhere. Notably, we discovered 11 immunopeptides derived from a squirrel monkey retrovirus in human leukemia monocytic cell line, thanks to the unrestricted search space offered by *de novo* peptide sequencing. This approach provides users with an exhaustive search space, in contrast to previous proteomics-to-genomics approaches, allowing them to decipher aberrantly transcribed/translated peptides while maintaining sensitivity.

EXPERIMENTAL PROCEDURES

Experimental Design and Statistical Rationale

pXg integrates *de novo* peptide sequencing with RNA-Seq results at the individual read level to identify confident cMAPs and ncMAPs based on the following steps (Fig. 1): (1) remove undesirable candidate peptides from *de novo* peptide sequencing results, (2) find exact matches between peptides and reads, (3) generate proteogenomic features, (4) rescore the peptide-read-spectrum match (PRSM) scores using Percolator and (5) estimate the FDR to report confident PRSMs. pXg begins by taking *de novo* peptides, RNA-Seq results, genome annotation and reference protein sequences. Optionally, undesired candidate peptides (e.g., those below a certain rank or peptides whose lengths are outside of the known range of typical MHC peptides) can be filtered out. The pipeline then finds exact matches between peptides and both target and decoy read sequences where decoy sequences are generated by reversing the target sequences while preserving their genomic loci and read alignment. Only matched peptides are retrieved, and their predicted spectrum and RT are obtained from Prosit (29) and DeepLC (30), respectively. Proteogenomic features are generated and subsequently used as Percolator input (details in Fig. 2). After rescoring, the best scored PRSM per spectrum is selected. Only PRSMs predicted as eluted ligands by NetMHCpan 4.1 are subjected to FDR estimation (subset FDR), while canonical and noncanonical PRSMs are separately estimated (separate FDR) using target-decoy competition (TDC) (31). For the conservative identification of noncanonical peptides, peptides with sequences matching a

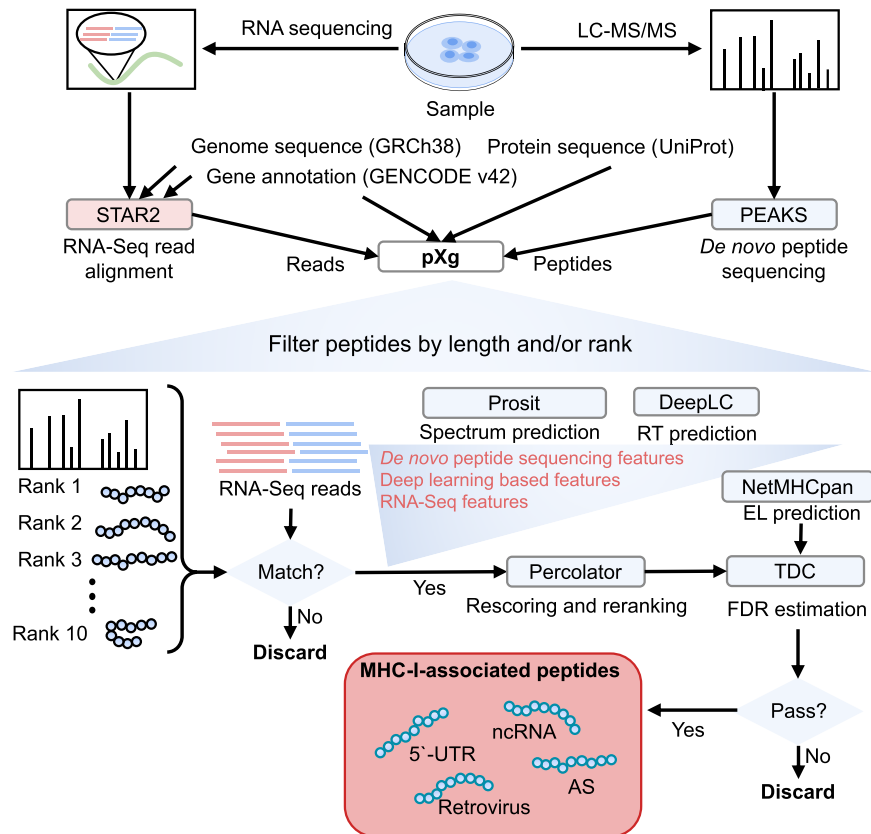


FIG. 1. Overview of the pXg pipeline. RNA-Seq alignment and *de novo* peptide sequencing are used as pXg input. Peptides can be filtered by length and/or rank. It then integrates *de novo* peptide sequencing with RNA-Seq results (both mapped and unmapped reads) and generates feature vectors. Based on the feature vectors, peptide-read-spectrum matches (PRSMs) are rescored by Percolator. The best scored PRSM per spectrum is selected and then PRSMs predicted as eluted ligands are subjected to target decoy competition (TDC) for FDR estimation. Lastly, confident PRSMs are reported with their genomic information. 5'-UTR, ncRNA and AS indicate 5'-UTR, noncoding RNA and alternative splicing, respectively. AS, alternative RNA splicing; FDR, false discovery rate; ncRNA, noncoding RNA; pXg, proteomics X genomics.

given reference protein sequence are considered canonical peptides. PRSMs passing 5% FDR threshold are considered confident and reported along with their genomic information. pXg can handle unmapped reads as well as mapped reads, allowing it to output even MAPs that are difficult to infer from the reference sequence database, such as antibody and host-infecting viral RNA. To demonstrate its performance, we used ten high-resolution immunopeptidome datasets (Table 1) and their matched paired-end RNA-Seq datasets (Table 2) and validated 16 ncMAPs using synthetic peptides in this study.

Core Algorithm in pXg Pipeline for identification of MAPs

De novo peptide sequencing algorithms assign sequences based solely on mass spectra. Thus, when given a low-quality tandem mass spectrum, ambiguous sequencing results may be reported. It is not unusual for a candidate peptide of a lower rank to be the correct sequence, rather than the top-ranked one. For example, LC-MS/MS assays often lack both N- and C-terminal fragment ions (i.e., b1 and y1 ions), leading to ambiguity in candidate sequence order near peptide termini. To overcome such difficulties, pXg considers all candidates up to rank n , where n is a user-specified parameter (in this study, we used a value of 10). This approach allows us to identify higher confidence peptides by leveraging RNA-Seq results obtained from the same (matched) sample during the proteomic analysis.

Possible candidate peptides (with lengths of 8–15 amino acids for MAPs) are aggregated to build a keyword-trie using the Aho–Corasick algorithm (32). To integrate peptide sequence data with genomic evidence, reads are converted to peptide sequences (target sequences) by three- or six-frame translation (Fig. 2A). Candidate peptide sequences are subsequently compared to the target sequences using the Aho–Corasick algorithm, which matches multiple patterns (peptide sequences) against text (target reads) in linear time ($O(\text{maximum length of peptides} + \text{target sequence length} + \text{the number of matches})$). This linkage between peptide and reads allows for the inference of genomic information for each peptide (it is the reason why we need the alignment of reads to genome). We assign transcriptional/translational interpretations based on the genomic information for each peptide (supplemental Note 1). This approach covers all possible events that can be inferred from each read in a reasonable amount of time.

Matching reads can drastically reduce false positives for canonical peptides. The probability that a *de novo* peptide sequence will match reads is generally lower than 0.01. For example, the chance of an arbitrary eight amino acid peptide matching protein coding sequences by chance is approximately $\frac{\text{all length 8 peptides in protein coding sequences}}{\text{all possible length 8 peptides}} \cong \frac{1.12 \times 10^7}{1.67 \times 10^{10} (\cong 19^8)} \cong 6.71 \times 10^{-4}$ (for 106,142 protein coding transcripts in GENCODE v38). On the other hand, the search space of noncanonical peptides of length eight is

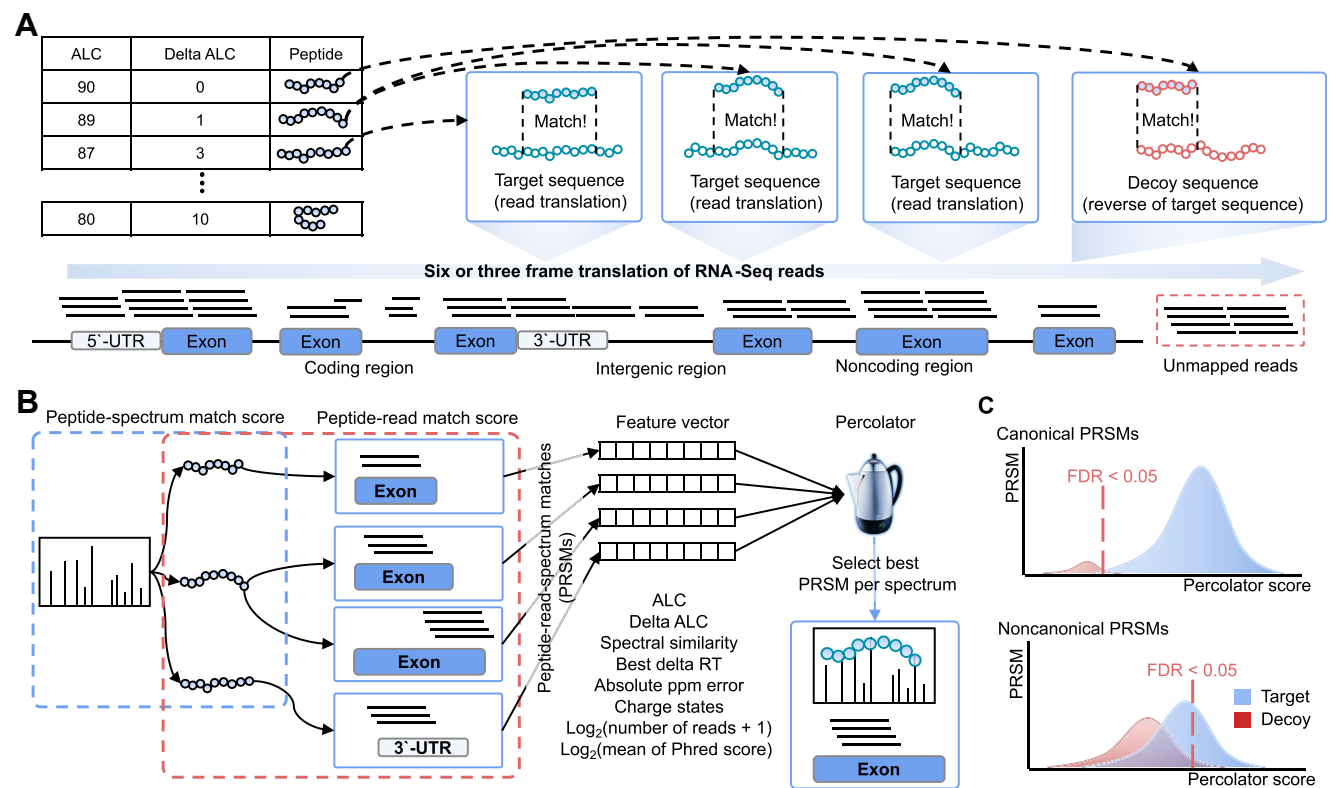


FIG. 2. Core algorithms in pXg pipeline. A, aggregating candidate peptides from *de novo* peptide sequencing and then finding exact matches between peptides and mapped/unmapped reads. Target and decoy sequences are generated by six- or three-frame translation of reads and their reverse sequences, respectively. B, machine learning approach to rescore peptide-read-spectrum matches (PRSMs). Average local confidence and retention time are denoted as ALC and RT, respectively. C, canonical and noncanonical PRSMs are separately estimated at 5% false discovery rate (FDR). pXg, proteomics X genomics.

$\sim 2.55 \times 10^8$ in the six-frame translation of all reference transcripts, including protein coding and ncRNAs (a total of 237,011 transcripts in GENCODE v38), yielding a higher probability of matching ($\sim 1.53 \times 10^{-2}$) than in the case of canonical peptides. Because pXg includes intronic regions and intergenic regions as well as unmapped regions, the probability is even higher, implying that the simple existence of matching reads is not sufficient to determine reliable noncanonical peptides. To reduce false positives caused by random matches, we applied a TDC commonly employed in proteomics data analyses. Decoy sequences

were generated by reversing the target sequences while preserving their genomic loci and read alignment. The decoy sequences were then matched to the peptides in the same manner.

Spectra that do not match any candidate peptides in both target and decoy sequences are discarded from further consideration. The peptide-spectrum pair and the peptide-read pair are connected via PRSM (Fig. 2B). Using Prosit and DeepLC, we generated predicted spectra and predicted RTs, respectively. We calculate the normalized spectral contrast angle (SA) between experimental and predicted

TABLE 1
MS/MS datasets

Sample	Instrument	Collision energy	The number of MS/MS spectra	Data access number	Publication
B-LCL1	LTQ-Orbitrap	35	58,771	PASS00270	(34)
B-LCL2	Elite		60,286		
B-LCL3			84,600	PXD001898	(4)
B-LCL4			103,801		
DOHH2	Q-Exactive HF	25	11,827	PXD020620	(10)
HBL1			29,786		
SHUDL4			9833		
THP1-1	Orbitrap Fusion	32	228,591	PXD015039	(35)
THP1-2			202,332		
THP1-3			322,714		

LTQ, linear trap quadrupole; MS/MS, tandem mass spectrometry.

TABLE 2
RNA-Seq datasets

Sample	Strand specificity	Reads	The number of reads ^a	Data access number	Publication
B-LCL1_1	Unstranded	PE100	16,745,792	SRR1925276	(4)
B-LCL1_2			17,331,638	SRR1925277	
B-LCL2_1			16,810,288	SRR1925278	
B-LCL2_2			17,060,387	SRR1925279	
B-LCL3_1	Unstranded	PE100	60,216,195	SRR1925280	(4)
B-LCL3_2			57,726,943	SRR1925281	
B-LCL4_1			59,326,152	SRR1925282	
B-LCL4_2			57,619,493	SRR1925283	
DOHH2_D1	Stranded	PE75	89,281,893	SRR12285182	(10)
DOHH2_D2			94,491,243	SRR12285181	
HBL1_D1			59,423,443	SRR12285180	
HBL1_D2			64,668,535	SRR12285175	
SHUDL4_D1	Stranded	PE100	129,892,001	SRR12285190	(6)
SHUDL4_D2			121,938,651	SRR12285189	
THP1_1_RNA			20,230,753	SRR13279451	
THP1_2_RNA			25,711,033	SRR13279452	
THP1_3_RNA			33,355,104	SRR13279453	

^aAfter trimming low-quality reads and removing adaptor sequences.

spectra (33) as a measure of spectral similarity. Additionally, we define the best delta RT as the smallest absolute difference between experimental and predicted RTs for each peptide. This results in the use of six proteomic features, including 1) *de novo* peptide sequencing score (e.g., average local confidence [ALC] score provided by PEAKS), 2) delta score (difference from the ALC score of the top ranked peptide), 3) absolute ppm error of a precursor ion *m/z*, 4) SA as spectral similarity, 5) best delta RT, and 6) charge state. These proteomic features have been shown to discriminate between true- and false-positive PSMs (a linkage between spectrum-peptide pair), but they do not account for the linkage between peptide-read pair. To evaluate peptide-read pairs, we utilize RNA abundance, as Chong and colleagues demonstrated that the proportion of long noncoding RNA-derived binder peptides increases with higher fragments per kilobase of transcript per million mapped reads (FPKM) values (5). RNA abundance per peptide is quantified by counting the number of matched reads, instead of using FPKM values, because FPKM is difficult to define for intronic, intergenic, and unmapped reads. We also use RNA-seq quality as a feature under the assumption that poorly sequenced reads are more likely to yield false positives. RNA sequencing quality is calculated as $\log_2(\text{mean of Phred score in the matched reads})$. Each PRSM feature is vectorized before being input to Percolator. Subsequently, only the best PRSM for each spectrum is selected and subjected to eluted ligand prediction by NetMHCpan 4.1. PRSMs with % Rank values lower than 2.0 are considered as eluted ligands and subsequently used for TDC.

In the TDC step (Fig. 2C), canonical and noncanonical PRSMs are separately validated (separate FDR estimation) because their score distributions are reportedly different (27). PRSMs passing 5% FDR threshold are deemed confident PRSMs and are reported along with their genomic information.

Immunopeptidome MS/MS and RNA-Seq Datasets

We used previously published 10 Immunopeptidome MS/MS datasets and their corresponding RNA-Seq datasets. Briefly, the first seven datasets are pan-HLA class I and were acquired from Epstein-Barr virus-transformed B-lymphoblastoid cell lines (B-LCL1, B-LCL2, B-LCL3, and B-LCL4) (4, 34) using linear trap quadrupole-Orbitrap

Elite and three human diffuse large B cell lymphomas cell lines (DOHH2, HBL1, and SUDHL4) using Q-Exactive HF (10). B-LCL1, B-LCL2, and B-LCL4 had four biological replicates each and B-LCL3 had three replicates. DOHH2, HBL1, and SUDHL4 had two replicates each. The last three datasets are mono-allelic HLA class I and were acquired from three biological replicates of acute myeloid leukemia cell lines (THP1-1, THP1-2, and THP1-3) using Orbitrap Fusion (35).

For RNA-Seq datasets, two biological replicates of each Epstein-Barr virus-transformed B-lymphoblastoid cell line were sequenced as 2 × 100-nt paired end reads using Illumina HiSeq 2000. Two replicates of Human diffuse large B cell lymphomas cell lines were sequenced as 2 × 75-nt paired end reads using Illumina NextSeq 550. Lastly, three biological replicates of acute myeloid leukemia cell lines were sequenced as 2 × 100-nt paired end reads using DNBSEQ-G400 (6).

De Novo Peptide Sequencing and RNA-Seq Alignment

Each MS dataset was searched using the commercial software PEAKS Studio 10.6 with the following search parameters: 10.0 ppm for parent mass tolerance, 0.02 Da for fragment mass tolerance, no enzyme specificity and reporting up to the top 10 candidates per spectrum. We extracted the “all candidates result” from PEAKS for input into pXg pipeline.

For paired-end RNA-Seq datasets, we used Trimmomatic v0.39 (36) to trim low-quality reads and remove adaptor sequences. The fastq data were then aligned to the human reference genome (GRCh38) with gene annotations (GENCODE v42) using the STAR2 (37) aligner with The Cancer Genome Atlas two-pass alignment option (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline). SAMtools v1.10 (38) was used to merge aligned results from replicates for each sample.

Spectral Similarity Between Experimental and Predicted Spectra

Prosit v1.1.2 was employed to predict MS/MS spectra of candidate peptides (29). We downloaded the Prosit hla_hcd model (https://figshare.com/articles/dataset/Prosit_Non_tryptic_-_Model_-_Fragmentation/12936947) to cover the prediction of nontryptic peptides. An SA was calculated for each pair of spectra, comparing experimental

and predicted spectra (33). It is noteworthy that we considered only b- and y-ions in the experimental spectra when calculating SA, because the fragment ion prediction was limited to these ion types.

Prediction of RT Using DeepLC

DeepLC v1.1.2 was used to predict PSM RTs (30). We calibrated the predicted RTs using cMAPs with the pygam_calibration option. Since a MAP can be detected by multiple spectra, we used the smallest absolute value of delta RT per MAP and denoted the feature as “best delta RT” in the manuscript.

pXg Parameter and Generation of Proteogenomic Feature

All reads from the replicates of each sample were combined. These reads, along with PEAKS de novo peptide sequencing results were input to pXg. The UniProt human protein database (version 2022.04, 102,601 entries), combined with 179 common contaminant proteins, was used as input to pXg. For B-LCL samples, we also concatenated the UniProt Epstein-Barr virus strain B95-8 protein database (version 2022.04, 88 entries) since we knew that the samples were infected by Epstein-Barr virus. We considered only the top 10 ranked candidates and peptide lengths of 8 to 15 amino acids to focus on the identification of MAPs. We used three-frame translation mode for DOHH2, HBL1, SUDHL4, and THP1 cell lines. For B-LCLs, we used six-frame translation mode due to the lack of its strand specificity.

After matching peptides to reads, we calculated their predicted spectra and RTs, which were used to generate deep learning-based features such as spectral similarity and best delta RT. We generated a feature vector using eight proteogenomic features, including (1) de novo peptide sequencing score, (2) delta score, which is a difference from the top-rank score, (3) absolute ppm error of a precursor m/z, (4) precursor charge state, (5) spectral similarity, (6) best delta RT, (7) log2(matched read count + 1), and (8) log2(mean of Phred score). These features were used to rescore and rerank PRSMs using Percolator, and the top-ranked PRSM for each spectrum was selected.

HLA Typing and Binding Prediction

For DOHH2, HBL1, SUDHL4, and THP1 cell lines, we used the same HLA types as described in the previous research. For the four B-LCL cell lines, we used OptiType (1) in RNA-Seq mode to identify the HLA-A, HLA-B, and HLA-C types because the previous research only reported HLA-A and B. Note that B-LCL1 and B-LCL2 share identical HLA types (Table 3). Afterward, both peptides and HLA types were input to NetMHCpan 4.1 (25) to predict HLA binding affinity. Among them, only peptides with %Rank <2.00 were selected as MAPs. cMAP and ncMAP were separately estimated at a 5% FDR.

Construction of Known MAP Database

To construct a known database, we collected MAPs from several resources as follows: first, we downloaded 739,670 MAPs from the Immune Epitope Database (2023-04) (39) with human-host only. We also downloaded 239,861 and 90,428 MAPs from IEAtlas (40) and HLA Ligand Atlas (41), respectively. IEAtlas is a database consisting of ncMAPs across 15 cancer types and 30 noncancerous tissues; HLA Ligand Atlas contained naturally presented MAPs only. Although IEAtlas included both MHC-I and MHC-II classes, we used all of them to construct the known database, because it was difficult to extract only MAPs from the resource. We also downloaded 21,322 MAPs reported by previous publications (4, 6, 10). The collection of MAPs yielded 921,951 MAPs in total.

Validation of Noncanonical Peptides by MS/MS Analyses of Synthetic Peptides

Noncanonical peptides were selected for validation using three criteria. First, the predicted binding affinity (%Rank) of the noncanonical peptides from NetMHCpan 4.1 were less than 2.00. Second, peptides included in the known MAP database were excluded to avoid biased validation. Lastly, we selected a total of 18 peptides such as (1) 14 peptides from unmapped reads (the reads were not aligned to the human genome), (2) three peptides from alternative splicing, and (3) one peptide from intergenic region harboring two mutations (1 SNV and 1 insertion). To validate the 18 peptides, we synthesized the peptides using a solid phase peptide synthesis method and analyzed the synthetic peptides by MS/MS.

Briefly, in order to load Fmoc-amino acid on the resin, a swollen Wang resin (1.10 mmol/g, 1.10 mmol) was treated with a solution of amino acid (0.60 equiv), N,N'-diisopropylcarbodiimide (4.0 equiv), and dimethylaminopyridine (0.20 equiv) in dichloromethane/dimethylformamide (DCM:DMF = 3:1, 8.0 ml). After reacting for 3.5 h at room temperature, the resin product was collected by filtration and washed with DMF (5.0 ml × 5) and DCM (5.0 ml × 5). The resultant resin was subsequently treated with acetic anhydride (50 equiv) in 5% pyridine in DMF (9.0 ml) to cap the remaining hydroxyl groups on the resin. After 3.5 h at room temperature, the resin product was collected by filtration and washed with DMF (5.0 ml × 5), DCM (5.0 ml × 5), and methanol (5.0 ml × 5). The final resin was dried by blowing with nitrogen gas for 3 h. Peptide chain elongation was conducted in an automated microwave peptide synthesizer (Liberty PRIME equipped with HT12, CEM Corporation) at the Center for Proteogenome Research. The synthetic scale was 0.005 mmol. Deprotection was performed with a solution of 2% pyrrolidine and ethyl cyano(hydroxyimino)acetate (Oxyma Pure) in DMF. Coupling reactions were carried out with Fmoc-amino acid (5.0 equiv), N,N'-diisopropylcarbodiimide (10 equiv), and Oxyma Pure (5.0 equiv) in DMF. Upon completing the peptide synthesis, the resin peptide was cleaved using trifluoroacetic acid (TFA)/water/triisopropylsilane (95:2.5:2.5 v/v) cocktail solution for

TABLE 3
HLA types of ten samples

Sample	HLA-A1	HLA-A2	HLA-B1	HLA-B2	HLA-C1	HLA-C2
B-LCL1	A03:01	A29:02	B08:01	B44:03	C07:01	C16:01
B-LCL2	A03:01	A29:02	B08:01	B44:03	C07:01	C16:01
B-LCL3	A02:01	A29:02	B57:01	B44:03	C07:01	C16:01
B-LCL4	A02:01	A01:01	B18:01	B39:24	C07:01	C05:01
DOHH2		A01:01	B08:01	B44:02	C07:01	C07:04
HBL1		A02:06		B51:01		C14:02
SUDHL4	A02:01	A31:01	B15:01			C03:04
THP1		A02:01		NA		

3 h and precipitated the peptide products with diethyl ether overnight at -20°C . Among 18 peptides, two peptides containing methionine were failed to synthesize due to the highly prevalent methionine oxidation and excluded for further analysis.

The resultant crude peptides were purified by a preparative HPLC using an Agilent 1260 Infinity II Prep LC system (Agilent Technologies) at the Center for Proteogenome Research, equipped with a Gemini 5 μm C18 110 column (250×21.1 mm, P/N 00G4454-P0-AX, Phenomenex). Solvent A (sol A) is 0.1% TFA in water, and solvent B (sol B) comprised 0.1% TFA in acetonitrile. A linear gradient of either 5 to 35% or 10 to 40% sol B over 30 min was used at a flow rate of 15 ml/min.

After the purification, each of the peptides was MS/MS analyzed by using Q Exactive (Thermo Fisher Scientific) equipped with a home-built electrospray ionization (ESI) source at the Center for Proteogenome Research. The ESI source connected a nano-ACQUITY UPLC system (Waters) with the mass spectrometer. Using 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B), the UPLC system generated a flow rate of 1 $\mu\text{l}/\text{min}$ of 50% solvent A and 50% solvent B that drive the peptide sample in a 1.5 μl sample loop for direct infusion. The ESI spray voltage was set to 2.4 kV and the ion transfer tube temperature was set to 250°C . For the MS data acquisition, a resolution of 140,000

and a scan range of m/z 133.4–2000 were used with an automatic gain control target of 1×10^6 and a radio frequency lens of 50%. For MS/MS data acquisition, an isolation width of 1.0 Th and a higher energy collisional dissociation energy of either 32 or 35 were used with a resolution of 17,500 and an automatic gain control target of 1×10^6 .

Similarity scores between MS/MS spectra of the endogenous peptides and their synthetic counterparts were determined for 16 peptides by calculating SA.

RESULTS

Discriminative Power of RNA-Seq Reads for PRSM Identification

We analyzed 946,182 spectra from 10 proteogenomic samples using the pXg pipeline. Out of these, 802,440 spectra contained at least one candidate peptide with lengths between eight and 15. Among them, 259,085 had candidate peptides that were fully covered by reads. This observation aligns with our expectations, as the majority of randomly sequenced peptides are filtered out due to a lack of RNA

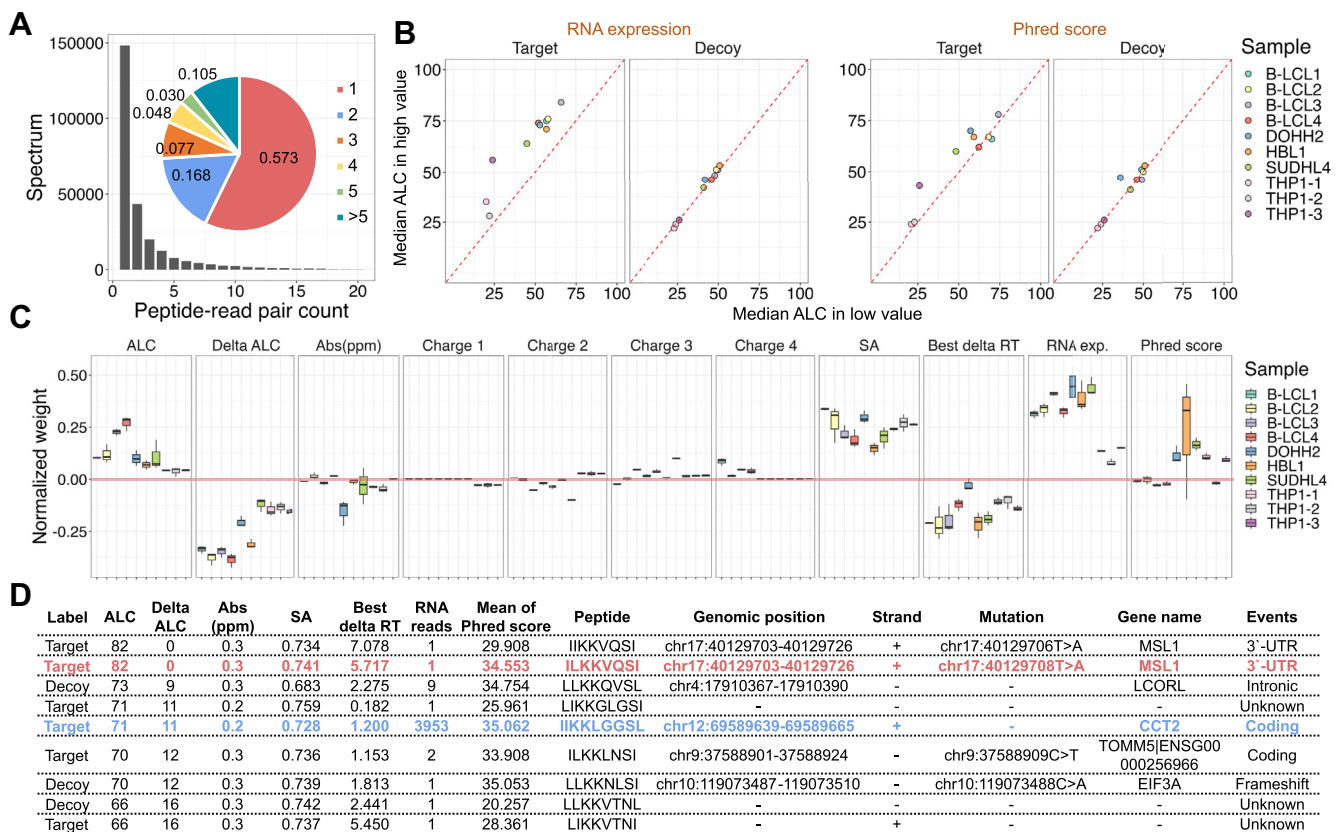


FIG. 3. Effects of RNA-Seq reads for peptide-read-spectrum match quality and their genomic interpretation. A, the number of spectra according to peptide-read pair count. B, median average local confidence (ALC) of high versus low RNA expression. A "high" value is defined as the top 50% of RNA expression in the left and Phred score in the right panels, respectively. Otherwise, it is a "low" value. Note that the RNA expression is log2 transformation of the number of reads matching each peptide. C, Normalized weights for the eight proteogenomic features of Percolator. RT and RNA exp. indicate retention time and RNA expression, respectively. Abs(ppm) means absolute value of ppm error. D, an example of peptide-read pairs for a single spectrum. The red color indicates the best-scored peptide-read-pair after rescoring by Percolator using proteomic features only. The blue color indicates the best-scored peptide-read pair after rescoring by Percolator using proteogenomic features. RT, retention time.

evidence. Notably, 110,735 out of 259,085 spectra (43.7%) had two or more peptide-read pairs (Fig. 3A), indicating potential ambiguous interpretations for many spectra. In such cases, integrating both transcriptomic and proteomic evidence becomes crucial in identifying the most accurate peptide sequence depicting the spectrum. As illustrated in Figure 3D, for example, a single spectrum has nine peptide-read pairs. The first and second peptides had identical PEAK's ALC scores. If only proteomic features were used, the second peptide ("ILKKVQSI") would be assigned to the spectrum due to its superior SA score and best delta RT value compared to the first peptide ("IIKKVQSI"). However, when proteogenomic features were used for rescoring, the fifth peptide ("IIKKLKGGSL") received the highest score.

Next, we examined whether the quantity and quality of reads matching each peptide correlated with its ALC score, a metric indicating peptide sequencing quality in PEAKS (Fig. 3B). First, we categorized peptide-read pairs into four groups based on their RNA expression levels (log2 transformed number of reads matching the peptide) and their status as either target or decoy. These groups were target/high, target/low, decoy/high, and decoy/low. Median ALC scores were calculated for each group and compared between high and low groups within target and decoy categories (left panel in Fig. 3B). The target/high group displayed higher median ALC scores than the target/low group, whereas the decoy group showed no such difference. Similarly, we compared high/low mean Phred scores, which represent RNA-Seq read quality (right panel in Fig. 3B). While the difference was less pronounced compared to RNA expression, the target/high group still exhibited increased median ALC.

To assess feature importance, we compared normalized weights of the support vector machine in Percolator (Fig. 3C). As expected, ALC and SA were positively weighted, while delta ALC and best delta RT were negatively weighted. Remarkably, in all samples, RNA expression was found to be a more powerful discriminative feature than ALC, with seven samples assigning it the highest positive weight. Mean Phred score was less dominant than ALC but was comparable to or stronger than other conventional proteomic features such as abs(ppm) and charge states. Figure 3D provides an example where the fifth peptide-read pair was assigned to the spectrum because of significantly higher RNA expression when proteogenomic features were considered. This highlights that the interpretation of the fifth row peptide-read pair is more reliable than that of other pairs because it originates from a protein coding sequence.

Confident Identification From De Novo Peptide Sequencing

We assessed whether the use of Percolator with our proteogenomic features in the TDC of *de novo* peptide sequencing could effectively distinguish correct and incorrect PRSMs. After rescoring PRSMs using Percolator, we retained

the top-scoring PRSM per spectrum. We then predicted their eluted ligand scores, %Rank, using NetMHCpan 4.1. Peptides with %Rank values below 2 were categorized as binders, while others were considered nonbinders. When comparing the score distributions of binder and nonbinder peptides between target and decoy PRSMs (Fig. 4A), the target PRSMs with score less than 0 showed similar binder and nonbinder ratio with those of decoy PRSMs. In contrast, target PRSMs with scores equal to or greater than 0 showed a notably higher proportion of binder peptides. When we used only proteomic features, the target and decoy PRSMs were less discriminant (supplemental Fig. S1). This result demonstrates that Percolator, equipped with our proteogenomic features, effectively discriminates between correct and incorrect PRSMs.

Recent immunopeptidomic pipelines such as MHCquant (22) and MAPDP (23) have applied a subset FDR strategy ("Sub"), as suggested by Sticker and colleagues (24), to increase the identifications of MAPs. In contrast to the "All" strategy, which subjects all PRSMs to TDC, the "Sub" strategy exclusively applies TDC to PRSMs of interest, that is, binder PRSMs in this study. To evaluate whether the "Sub" strategy could improve identification in the pXg pipeline, we compared the proportion of target PRSMs between the "All" and "Sub" strategies (Fig. 4B). The average target proportions for canonical PRSMs in ten samples were 0.984 for "All" and 0.997 for "Sub," indicating only slight improvements with the "Sub" strategy. However, in the case of noncanonical PRSMs, the average target proportion for "Sub" was greater than that for "All" (0.662 and 0.612, respectively). We further compared the number of binders in noncanonical PRSMs while varying the FDR threshold from 0.01 to 0.1 in increments of 0.01 (Fig. 4C). B-LCL4, DOHH2, HBL1, and SUDHL4 samples exhibited marginal or no improvements, while the other samples showed significant improvements. Thus, we adopted the "Sub" strategy and separated canonical and noncanonical PRSMs to estimate an FDR of 5%. This resulted in the identification of 131,082 canonical PRSMs and 3201 noncanonical PRSMs, yielding 24,449 cMAPs and 956 ncMAPs across 10 samples (Table 4 and supplemental Data 1). Notably, the FDRs for canonical PRSMs were less than 1% due to lower probability of random matches. It is important to highlight that the number of identifications varies according to a user-specified parameter rank *n* (supplemental Fig. S2).

The spectrum match and RNA-seq qualities of the identified canonical and noncanonical PRSMs were comparable (Fig. 4D), showing similar characteristics. There was no significant difference in length distribution between the two groups (Fisher's exact test; Fig. 4E). Length 9 peptides were enriched in both canonical and noncanonical MAPs, consistent with the characteristics of MAPs. These findings underscore the reliability of the pXg pipeline in distinguishing between correct and incorrect PRSMs obtained through *de novo* peptide sequencing.

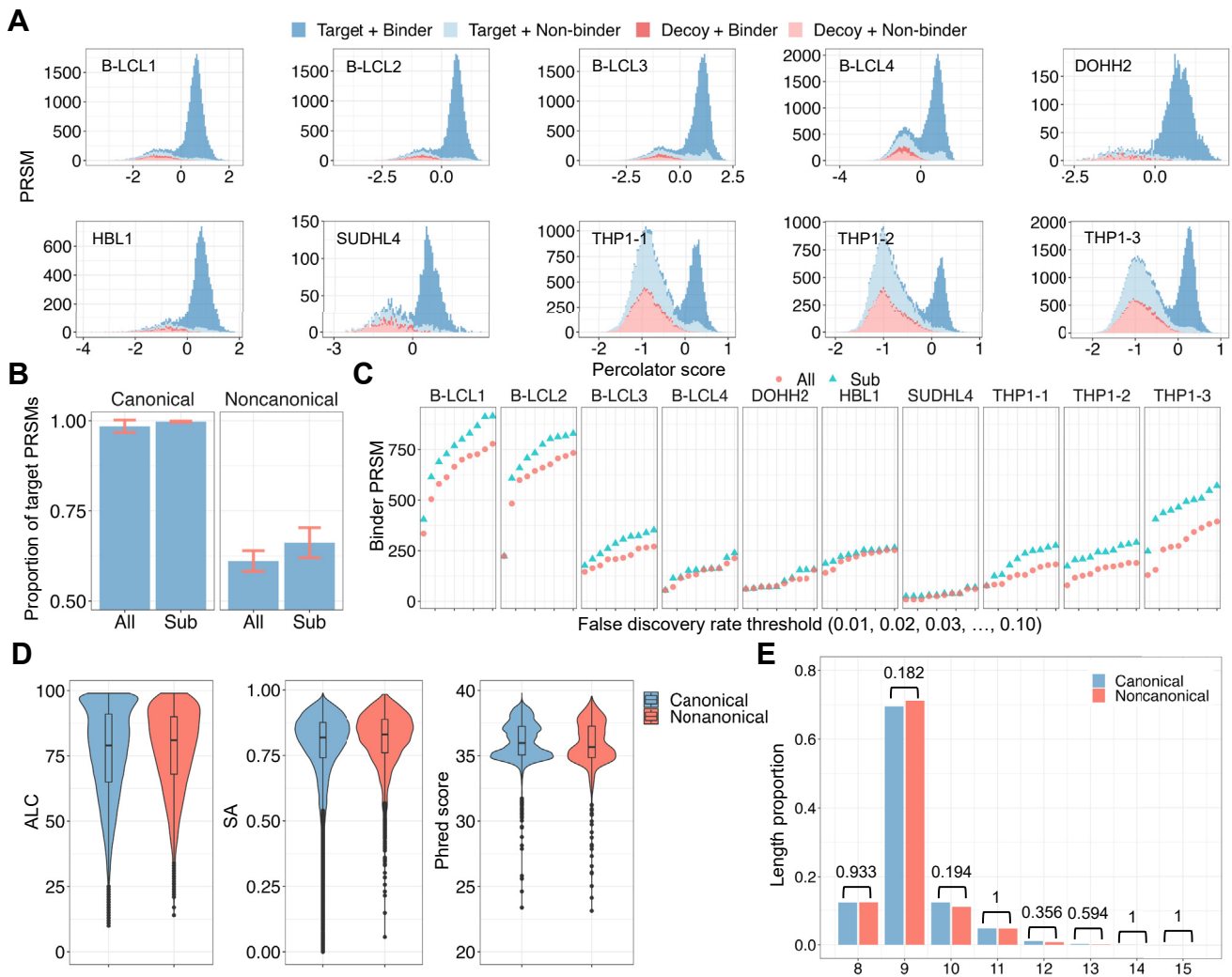


FIG. 4. Analysis of target-decoy competition in *de novo* peptide sequencing. A, target and decoy distributions of binder and nonbinder peptides are presented by cumulative histogram. B, proportions of target peptide-read-spectrum matches (PRSMs) according to “All” and “Sub” strategies are shown in canonical and noncanonical PRSMs, separately. A “All” means that all PSMs are used, and a “Sub” means that only binder PRSMs are used. Error bars represent SD. C, the number of binders in noncanonical PRSMs is plotted according to false discovery rate threshold from 0.01 to 0.1. D, comparison peptide/RNA-Seq quality between confident canonical and noncanonical binder PRSMs. E, length distribution of confident canonical and noncanonical MHC-I-associated peptides (MAPs). The number indicates *p*-value of Fisher’s exact test. MHC-I, major histocompatibility complex; PSM, peptide spectrum match.

Comprehensive Identification of ncMAPs

One of the primary advantages of *de novo* peptide sequencing is its potential to discover new peptides. We built a known MAP database by aggregating data from six public resources, including the Immune Epitope Database (39), IEAtlas (40), HLA Ligand Atlas (41), and three publications (Fig. 5A) (4, 6, 10). Out of the 24,449 cMAPs identified by pXg, 1611 cMAPs (6.59%) were not found in the known database. On the other hand, a significant portion of ncMAPs (390 of 956; 40.79%) were absent from the known database. Note that when we compared with the previous results from three publications (4, 6, 10), pXg pipeline outnumbered the

identification in general (supplemental Figs. S3 and S4). We further categorized the ncMAPs by 782 WT and 174 mutant MAPs (Fig. 5B). As expected, well-studied and/or straightforward events like frameshift, 5'-UTR and ncRNA translations were more likely to be reported for WT ncMAPs than for less well-studied or complex events (left panel in Fig. 5B). For example, 24 of 33 ASs (72.73%) identified were not found in the known database. Two noteworthy ncMAPs involved alternative splicing in PAK1 and RNF181. PAK1 has been suggested as a therapeutic target in BRAF WT melanoma (42), while the depletion of RNF181 was known to inhibit breast cancer progression *in vivo* and *in vitro* (43). Mutant ncMAPs

TABLE 4
Summary of identifications in ten samples

Sample	Canonical			Noncanonical		
	FDR	PRSM	MAP	FDR	PRSM	MAP
B-LCL1	<0.01	19,192	6360	0.048	768	296
B-LCL2		19,059	5936	0.049	733	289
B-LCL3		20,577	3884	0.049	286	109
B-LCL4		20,962	3442	0.045	155	54
DOHH2		3214	1659	0.042	72	43
HBL1		8339	3209	0.046	237	102
SUDHL4		1635	926	0.032	31	18
THP1-1		9930	4835	0.048	210	128
THP1-2		7028	3455	0.045	244	133
THP1-3		21,146	7186	0.049	465	213

FDR, false discovery rate; MAP, MHC-I-associated peptide; PRSM, peptide-read-spectrum match.

are more challenging than the WT ncMAPs because combinations of point mutations are involved. Reflecting the difficulty, the proportion of “not found” ncMAPs in mutants was higher than that in WT (67.24% and 34.91%, respectively).

Only 12 ncMAPs (12 of 956; 1.26%) were assigned multiple genomic annotations, primarily due to synonymous mutations or the assignment of multiple genomic regions and/or overlapping transcripts in the same genomic region (supplemental Fig. S5). Sixteen ncMAPs matched unmapped reads (labeled as “Unknown” in Fig. 5B). The unmapped reads were annotated using blastn (44) or MiXCR (45) in case those tools were available to ascertain their genomic origins (supplemental Fig. S6). Unexpectedly, 11 ncMAPs in THP1 cell line were significantly matched to the squirrel monkey retrovirus (Fig. 5C). Three of these ncMAPs originated from intergenic regions or point mutations in the retrovirus. Furthermore, a total of 392,228 reads matched these ncMAPs, confirming their translation in the THP1 cell line (supplemental Figs. S7 and S8). Notably, we observed 65 mutations in the coding regions of the virus genome while none were nonsense mutations, indicating intact full-length translation (supplemental Data 2). The remaining ncMAPs were associated with a complementarity-determining region 3 in an antibody, SATB1, ATG13, or UBC (supplemental Fig. S6). It is noteworthy that the reads from SATB1, ATG13, and UBC exhibited low qualities (mean Phred score <32.37), making them difficult to align to the human genome. However, the pXg pipeline successfully rescued the reads based on their proteomic evidence.

We examined 174 mutant MAPs in more detail. We excluded six mutant MAPs annotated as “Multiple” because their origins were ambiguous. When comparing 168 mutant MAPs to the dbSNP (46) and COSMIC (47) databases (Figs. 5D), 45 combinations of mutations in 56 known mutant MAPs were annotated in at least one mutation resource. On the other hand, only 30 of 112 mutant MAPs in “not found” category were annotated. This highlights that mutant MAPs in “not found” category are likely to represent new findings at both genomic and proteomic levels.

From the identification, we further validated the presence of 16 peptides including 10 retroviral peptides by synthesizing peptides (supplemental Fig. S9). By comparing between MS/MS spectra of the endogenous peptides and their synthetic counterparts, we confirmed that 14 of 16 peptides were truly expressed (SA >0.75). As example of a retroviral peptide that showcases the power of pXg is provided in Figure 5E using the Universal Spectrum Explore (48).

DISCUSSION

In conventional genomics-to-proteomics approaches, where MS-based proteomics data are typically searched against transcriptome-derived databases, researchers must construct custom proteogenomic databases tailored to their specific research objectives. For example, if one aims to identify peptides originating from UTRs and ncRNAs, separate sequence databases for UTRs and ncRNAs must be constructed and combined prior to matching them with the proteomics data. Thus, peptide identification is constrained by predefined search space of the database. To overcome this limitation, the proteomics-to-genomics approach using *de novo* peptide sequencing was introduced to identify unrestricted ncMAPs (15, 16). However, even in this approach, one must inevitably rely on predefined sequence databases to find the genomic origins of peptides. Consequently, the efficacy of peptide identification is heavily dependent on the comprehensiveness of these sequence databases, resulting in the omission of peptides derived from unexpected biological processes such as viral infection.

The adoption of sequence similarity search tools like blast may help finding the origins of peptides. However, immunopeptides are typically short, ranging from 8 to 15 amino acids, resulting in unreliable matches due to higher e-values. Furthermore, blast-based methods have been associated with higher false-positive rates when applied to mutant peptides, limiting their effectiveness in the discovery of tumor-specific antigens.

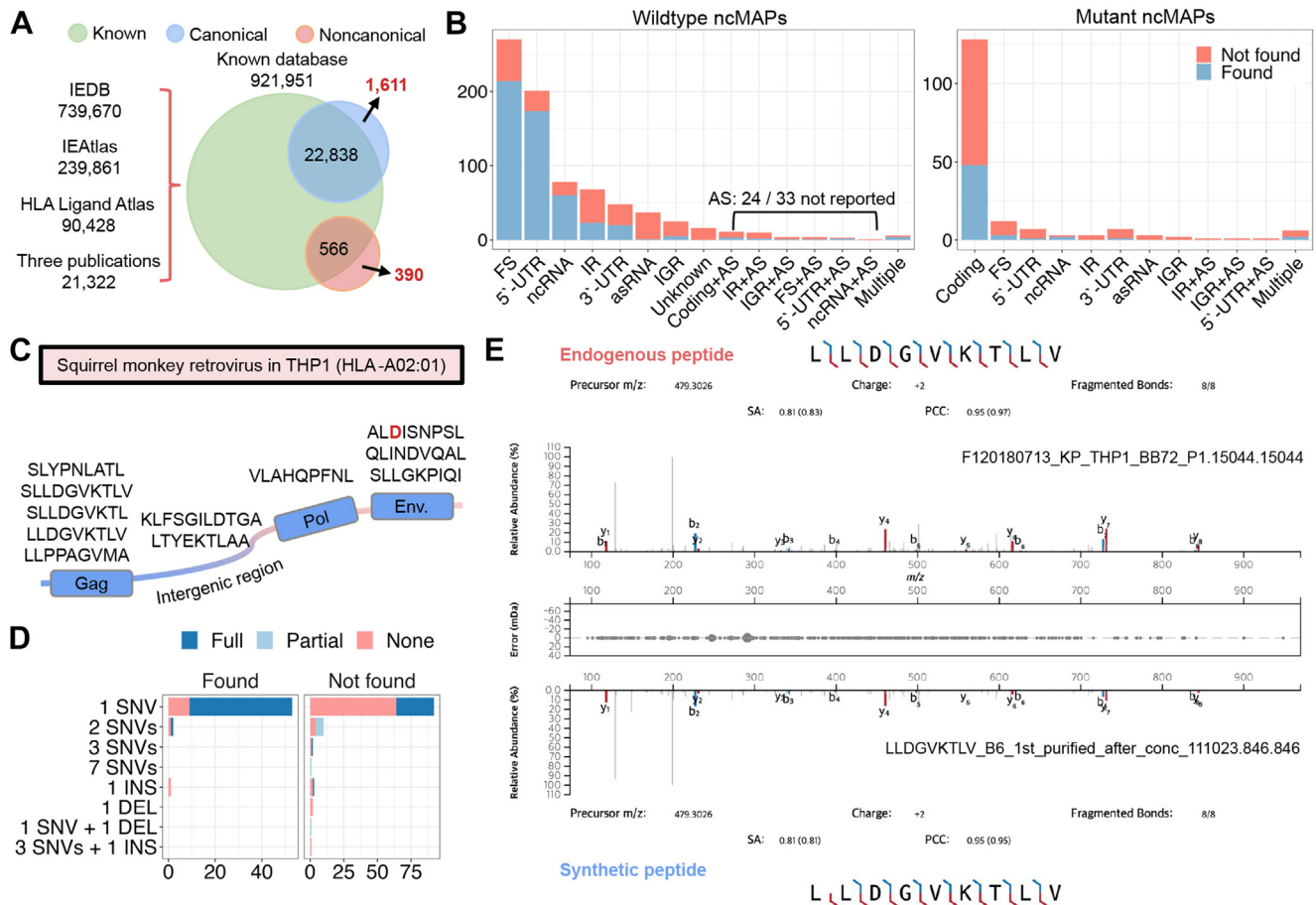


FIG. 5. Comprehensive identification of noncanonical MHC-I-associated peptides. A, comparison between our findings in MHC-I-associated peptides (MAPs) and known MAPs. Our findings are denoted as “Canonical” and “Noncanonical” depending on their existence in a reference protein sequence database. B, each proportion of ncMAPs found is shown according to events. If a MAP is contained in the existing databases, it is denoted as “Found.” Otherwise, it is denoted as “Not found.” Frameshift (FS), UTR, intron (IR), antisense RNA (asRNA), intergenic region (IGR), protein coding sequence (coding), and alternative splicing (AS) are shown abbreviated. When a peptide matches unmapped reads, it is denoted “Unknown.” If a peptide can be interpreted as more than a single event, it is denoted “Multiple.” C, identification of 11 ncMAPs derived from squirrel monkey retrovirus in THP1 cell line. Two peptides, “KLFSGILDTGA” and “LTYEKTAA” locate in intergenic region between gag and pol proteins. Red letter indicates a single amino acid change in “ALDISNPSL.” D, combinations of mutations among identified ncMAPs are described. If all mutations in a combination are reported by the dbSNP and/or COSMIC database, then the combination is denoted as “Full.” If all mutations in a combination are not reported by them, “None.” Otherwise, “Partial.” E, comparison mass spectra between endogenous (upper) and synthetic (lower) peptides. MHC-I, major histocompatibility complex; PSM, peptide spectrum match.

In contrast, the pXg workflow capitalizes on RNA-Seq reads to address the aforementioned challenges. Firstly, we can determine the origins of peptides using blast with the reads instead of the peptides. As we demonstrated, this approach can sensitively identify ncMAPs derived from intricate biological processes, including unanticipated viral proteins and intergenic regions. Furthermore, the pXg pipeline excels in recognizing mutant peptides. Unless RNA-Seq fails to capture some transcripts or peptides resulting from non-RNA-mediated processes, the pXg pipeline can identify virtually every potential MAP, regardless of complexity.

The identification of the squirrel monkey retrovirus in THP1 cell line serves as an illustrative example. This was corroborated by (1) the abundance of reads matching the retrovirus

and (2) the successful synthesis of peptides matching these reads. Several retroviruses are known to have direct or indirect links to cancer development (49). However, the role of the retrovirus in THP1 cell remains uncertain and questions of potential laboratory contamination arise.

Regarding the strategy for FDR estimation, we applied a subset FDR method to increase identification rate. However, this method may potentially miss some peptides that are not wellrepresented or welltrained in binding prediction tools. This challenge is not unique to the pXg pipeline but is also faced by other immunopeptidomics pipelines like MHCquant and MAPDP. Hence, future developments should focus on alternative ways to address the potential loss of peptides in FDR estimation.

In summary, the identifications achieved through the pXg pipeline underscore its capability to harness the advantages of *de novo* peptide sequencing. It showcases the prospective applications of this approach in the discovery of new immunotherapeutic targets, providing a promising avenue for advancing our understanding of the immunopeptidome and its relevance in various biological processes and diseases.

DATA AVAILABILITY

Immunopeptidome MS data were previously published and are available as the PASS00270 for B-LCL1 and 2, the PXD001898 for B-LCL3 and 4, the PXD020620 for DOHH2, HBL1, and SHUDL4 and the PXD015039 for THP1. Corresponding RNA-Seq data were previously published and are available as the PRJNA279172 for B-LCLs, the PRJNA647736 for DOHH2, HBL1, and SHUDL4 and the PRJNA686824 for THP1. All PEAKS *de novo* peptide sequencing projects and mass spectra of synthetic peptides are available as the PXD048006 in the ProteomeXchange Consortium via the PRIDE partner repository. The source code of pXg is freely available at <https://github.com/progistar/pXg>. This article contains supplemental data.

Supplemental data—This article contains supplemental data (25, 37, 44, 45, 48, 50).

Acknowledgments—This work was supported by the BK21 Fostering Outstanding Universities for Research (FOUR) program through the National Research Foundation (NRF) funded by the Ministry of Education of Korea, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2019M3E5D3073568) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University), No. 2021-0-02068, Artificial Intelligence Innovation Hub) and a grant from the Korea Basic Science Institute (KBSI) National Research Facilities & Equipment Center (NFEC) that is funded by the Korean government (Ministry of Education) (2019R1A6C1010028).

Author contributions—S. C. and E. P. conceptualization; S. C. and E. P. methodology; S. C. software; S. C. formal analysis; S. C. writing-original draft; E. P. writing-review and editing; E. P. supervision; E. P. funding acquisition.

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: ALC, average local confidence; cMAP, canonical MAP; FPKM, fragments per kilobase of transcript per million mapped reads; DCM, dichloromethane; DFM, dimethylformamide; ESI, electrospray ionization; FDR, false discovery rate; MAP, MHC-I-associated

peptide; MHC-1, major histocompatibility complex; MS, mass spectrometry; MS/MS, tandem mass spectrometry; ncMAP, noncanonical MAP; ncRNA, noncoding RNA; PRSM, peptide-read-spectrum match; PSM, peptide spectrum match; pXg, proteomics X genomics; RT, retention time; SA, normalized spectral contrast angle; TDC, target-decoy competition; TFA, trifluoroacetic acid.

Received July 12, 2023, and in revised form, February 19, 2024
Published, MCPRO Papers in Press, February 23, 2024, <https://doi.org/10.1016/j.mcpro.2024.100743>

REFERENCES

1. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014) OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316
2. Orenbuch, R., Filip, I., Comito, D., Shaman, J., Pe'er, I., and Rabadan, R. (2020) arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* **36**, 33–40
3. Li, X., Zhou, C., Chen, K., Huang, B., Liu, Q., and Ye, H. (2021) Benchmarking HLA genotyping and clarifying HLA impact on survival in tumor immunotherapy. *Mol. Oncol.* **15**, 1764–1782
4. Laumont, C. M., Daouda, T., Laverdure, J. P., Bonnell, É., Caron-Lizotte, O., Hardy, M. P., et al. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238
5. Chong, C., Müller, M., Pak, H., Harnett, D., Huber, F., Grun, D., et al. (2020) Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293
6. Scull, K. E., Pandey, K., Ramarathnam, S. H., and Purcell, A. W. (2021) Immunopeptidogenomics: Harnessing RNA-seq to illuminate the dark immunopeptidome. *Mol. Cell Proteomics* **20**, 100143
7. Weingarten-Gabbay, S., Klaeger, S., Sarkizova, S., Pearlman, L. R., Chen, D. Y., Gallagher, K. M. E., et al. (2021) Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. *Cell* **184**, 3962–3980.e17
8. Qi, Y. A., Maity, T. K., Cultraro, C. M., Misra, V., Zhang, X., Ade, C., et al. (2021) Proteogenomic analysis unveils the HLA class I-presented immunopeptidome in melanoma and EGFR-mutant lung adenocarcinoma. *Mol. Cell Proteomics* **20**, 100136
9. Chen, L., Zhang, Y., Yang, Y., Li, H., Dong, X., Wang, H., et al. (2021) An integrated approach for discovering noncanonical MHC-I peptides encoded by small open reading frames. *J. Am. Soc. Mass Spectrom.* **32**, 2346–2357
10. Ruiz Cuevas, M. V., Hardy, M. P., Holly, J., Bonnell, É., Durette, C., Courcelles, M., et al. (2021) Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815
11. Frankiw, L., Baltimore, D., and Li, G. (2019) Alternative mRNA splicing in cancer immunotherapy. *Nat. Rev. Immunol.* **19**, 675–687
12. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567
13. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., et al. (2012) Peaks DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell Proteomics* **11**, M111.010587
14. McDonnell, K., Howley, E., and Abram, F. (2022) The impact of noise and missing fragmentation cleavages on. *Comput. Struct. Biotechnol. J.* **20**, 1402–1412
15. Faridi, P., Li, C., Ramarathnam, S. H., Vivian, J. P., Illing, P. T., Mifsud, N. A., et al. (2018) A subset of HLA-I peptides are not genomically templated: evidence for cis- and trans-spliced peptide ligands. *Sci. Immunol.* **3**, eaar3947
16. Erhard, F., Dölken, L., Schilling, B., and Schlosser, A. (2020) Identification of the cryptic HLA-I immunopeptidome. *Cancer Immunol. Res.* **8**, 1018–1026
17. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., et al. (2003) PEAKS: powerful software for peptide de novo sequencing by

- tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342
18. Parker, R., Taylor, A., Peng, X., Nicastrì, A., Zerweck, J., Reimer, U., *et al.* (2021) The choice of search engine affects sequencing depth and HLA class I allele-specific peptide repertoires. *Mol. Cell Proteomics* **20**, 100124
 19. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925
 20. Li, K., Jain, A., Malovannaya, A., Wen, B., and Zhang, B. (2020) DeepRescore: leveraging deep learning to improve peptide identification in immunopeptidomics. *Proteomics* **20**, e1900334
 21. Declercq, A., Bouwmeester, R., Hirschler, A., Carapito, C., Degroevé, S., Martens, L., *et al.* (2022) MS² rescore: data-driven rescoring dramatically boosts immunopeptide identification rates. *Mol. Cell. Proteomics* **21**, 100266
 22. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., *et al.* (2019) MHCquant: automated and reproducible data analysis for immunopeptidomics. *J. Proteome Res.* **18**, 3876–3884
 23. Courcelles, M., Durette, C., Daouda, T., Laverdure, J. P., Vincent, K., Lemieux, S., *et al.* (2020) MAPDP: a cloud-based computational platform for immunopeptidomics analyses. *J. Proteome Res.* **19**, 1873–1881
 24. Sticker, A., Martens, L., and Clement, L. (2017) Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nat. Methods* **14**, 643–644
 25. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454
 26. O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020) MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* **11**, 418–419
 27. Nesvizhskii, A. I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125
 28. Woo, S., Cha, S. W., Na, S., Guest, C., Liu, T., Smith, R. D., *et al.* (2014) Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics* **14**, 2719–2730
 29. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., *et al.* (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518
 30. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L., and Degroevé, S. (2021) DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369
 31. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214
 32. Aho, A. V., and Corasick, M. J. (1975) Efficient string matching: an aid to bibliographic search. *Commun. ACM* **18**, 333–340
 33. Toprak, U. H., Gillet, L. C., Maiolica, A., Navarro, P., Leitner, A., and Aebersold, R. (2014) Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol. Cell Proteomics* **13**, 2056–2071
 34. Granados, D. P., Sriranganadane, D., Daouda, T., Zieger, A., Laumont, C. M., Caron-Lizotte, O., *et al.* (2014) Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat. Commun.* **5**, 3600
 35. Pandey, K., Mifsud, N. A., Lim Kam Sian, T. C. C., Ayala, R., Ternette, N., Ramarathnam, S. H., *et al.* (2020) In-depth mining of the immunopeptidome of an acute myeloid leukemia cell line using complementary ligand enrichment and data acquisition strategies. *Mol. Immunol.* **123**, 7–17
 36. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120
 37. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
 38. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008
 39. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., *et al.* (2019) The immune Epitope database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343
 40. Cai, Y., Lv, D., Li, D., Yin, J., Ma, Y., Luo, Y., *et al.* (2023) IEAtlas: an atlas of HLA-presented immune epitopes derived from non-coding regions. *Nucleic Acids Res.* **51**, D409–D417
 41. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D. J., Freudenmann, L. K., Backert, L., *et al.* (2021) HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J. Immunother. Cancer* **9**, e002071
 42. Ong, C. C., Jubb, A. M., Jakubiak, D., Zhou, W., Rudolph, J., Haverty, P. M., *et al.* (2013) P21-activated kinase 1 (PAK1) as a therapeutic target in BRAF wild-type melanoma. *J. Natl. Cancer Inst.* **105**, 606–607
 43. Zhu, J., Li, X., Su, P., Xue, M., Zang, Y., and Ding, Y. (2020) The ubiquitin ligase RNF181 stabilizes ER α and modulates breast cancer progression. *Oncogene* **39**, 6776–6788
 44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
 45. Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., *et al.* (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381
 46. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311
 47. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947
 48. Schmidt, T., Samaras, P., Dorfer, V., Panse, C., Kockmann, T., Bichmann, L., *et al.* (2021) Universal spectrum explorer: a standalone (web-)application for cross-resource spectrum comparison. *J. Proteome Res.* **20**, 3388–3394
 49. Fan, H. (2007) A new human retrovirus associated with prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1449–1450
 50. Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24