An Integrated Approach for Discovering Non-canonical MHC-I Peptides Encoded by Small Open Reading Frames

Lei Chen[2], Yuanliang Zhang[1], Ying Yang[1], Yang Yang[1], Huihui Li[3], Xuan Dong[4], Hongwei Wang[3], Zhi Xie[3], Qian Zhao[1]*

1. State Key Laboratory of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong SAR, China
2. Laboratory for Synthetic Chemistry and Chemical Biology Limited, Hong Kong SAR, China
3. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
4. BGI-Shenzhen, Shenzhen 518083, China

*Email: q.zhao@polyu.edu.hk

KEYWORDS: MHC-I peptides, small open reading frames, Ribo-seq, database search, *de novo* sequencing

ABSTRACT: MHC-I peptides are a group of important immunopeptides presented by major histocompatibility complex (MHC) on the cell surface for immune recognition. The majority of reported MHC-I peptides are derived from protein coding sequences, and non-canonical peptides translated from small open reading frames (sORF) are largely unknown due to the lack of accurate and sensitive detection methods. Herein we report an efficient approach that implements complementary bioinformatic strategies to improve the identification of non-canonical MHC-I peptides. In a database search strategy, non-canonical immunopeptides mapping was optimized by combining three complementary pipelines to construct predicted sORF databases from Ribo-seq data. In a *de novo* peptide sequencing strategy, MS data search results were filtered against sORF databases to pin down additional non-canonical immunopeptides. In total, 308 non-canonical immunopeptides were identified from two tumor cell lines with selected ones vigorously validated. Our approach is a handy solution to identify non-canonical MHC peptides with Ribo-seq and MS data. Meanwhile, the novel non-canonical immunopeptides identified with this method could shed insights on fundamental immunology as well as cancer immunotherapies.

INTRODUCTION

MHC-I peptides that are presented on the cancer cell surface are key factors for recognition by $CD8^+$ T cells to trigger immune response in cancer immunotherapy.[1-5] MHC-I peptides that are unique to cancer, or neoantigens, are considered ideal immunotherapy targets and thus are very attractive. Initially, most neoantigens were identified through computational prediction of mutation-bearing MHC-I peptides encoded by protein coding sequences obtained from whole exome sequencing and RNA-seq.[2] As this approach failed to provide neoantigens with high efficacy for patients with low tumor mutation burden (TMB),[6] recent efforts have focused on aberrantly expressed MHC-I peptides from alternative spliced RNAs,[7] introns,[8] non-coding regions,[9, 10] and even epigenetic changes.[11, 12]

Protein coding DNA sequences constitute 1.5% of the whole genome,[13] whereas up to 75% of the genome can be transcribed yet are presumably "non-coding".[14] In recent years, it has come to light that many of the non-coding RNAs have coding potential to produce polypeptides.[15] Thousands of ncRNAs, representing 40% of long ncRNAs (lncRNAs) and pseudogene RNAs, and 35% of untranslated regions (UTRs) in messenger RNAs, are potentially translated.[16] Computational, ribosomal profiling and mass spectrometry (MS) studies have collectively demonstrated the existence of a plethora of peptides encoded by unannotated small open reading frames (sORFs) that are about 100-codon-long. The sORFs-encoded small peptides (SEPs) or non-canonical peptides have specific subcellular distribution, similar abundance to their canonical counterparts and even essential biological functions.[9, 17-21] The increasing knowledge of sORFs offers a great opportunity for discovering MHC-I peptides or even potent cancer neoantigens from non-canonical SEPs.

Systematic identification of sORFs and their translation products remains a promising and yet challenging task. While computational approaches predict the existence of thousands of sORFs,[22] mass spectrometry is the only and final proof of non-canonical peptide expression. Like all database-dependent proteomics studies, sensitive and accurate detection of SEPs heavily relies on the presence of high-quality databases. Custom sORF database construction has been performed with various methods including in-silico six-frame translation of whole-genome sequences,[21, 23] three-frame translation from Refseq,[24] exome-seq,[25] RNA-seq[22] and Ribo-seq.[17, 26] Most prior works combined these sequencing methods to build comprehensive sORF databases. However, in-silico translation of genome or

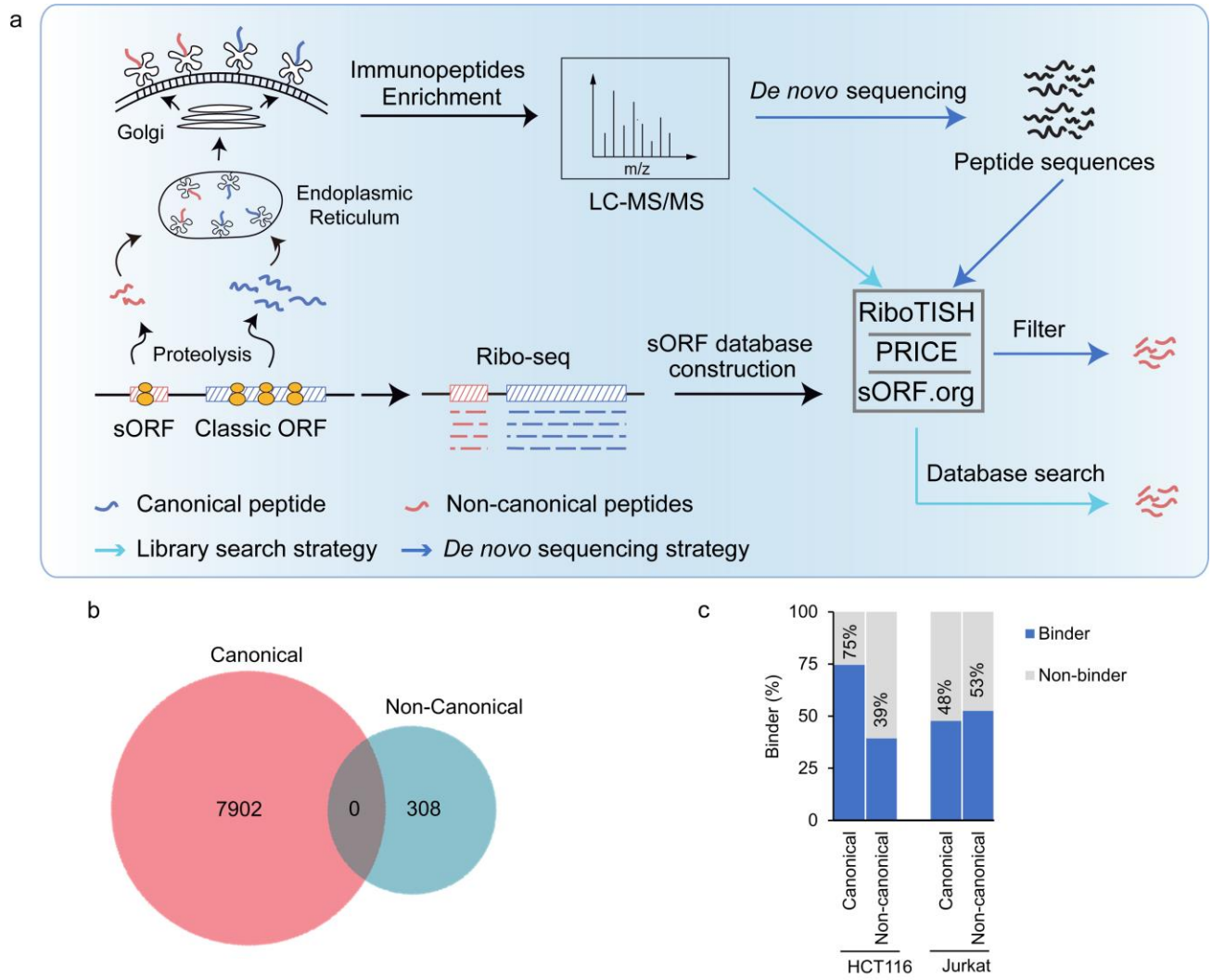


Figure 1. Methods for identifying MHC I peptides from canonical and non-canonical peptides. (a) Schematic overview of the approach. (b) Number of canonical and non-canonical immunopeptides. (c) The percentage of predicted canonical and non-canonical MHC-I peptides as binders.

transcriptome would result in a large number of predicted sORF sequences. Searching MS data against such inflated databases may increase the chance of false discovery. Meanwhile, substantial amount of sequencing work and strong bioinformatics skills are required to identify hundreds of non-canonical immunopeptides.

Ribo-seq offers superior reliability for sORF prediction because it is a snapshot of the translational events and can avoid generating inflated databases. So far, various bioinformaticspipelines with respective strength have been developed for predicting ORFs from Ribo-seq data. However, the performance of the pipelines used to construct custom database of SEPs have not been fully evaluated. Among them, some emerging pipelines enable sORF prediction based on user-defined Ribo-seq data. For example, RiboTISH uses statistical tests to assess translational events.[27] It enables efficient *de novo* prediction of sORF with either AUG or near-cognate start codons. PRICE, which is based on supervised learning, is capable of modeling the noise in Ribo-seq and improving resolution for sORF prediction.[28] RiboTISH and PRICE have been reported to outperform other methods and therefore would be adopted in our study. In addition, sORF.org, a readily available sORF database that contains a large number of sORF predicted from Ribo-seq data,[29] would also be evaluated here considering its easy accessibility to common users.

The database search strategy heavily relies on available protein sequence references, which limits its performance for a less defined proteome background as in the case of non-canonical peptides. On the contrary, *de novo* peptide sequencing method circumvents the need of a reference database by inferring the amino acid sequences directly from experimental MS/MS spectra. Combining these two complementary strategies, database search and *de novo* sequencing, could improve detection of immunopeptides. There were several successful endeavors to identify canonical immunopeptide with *de novo* sequencing.[30-34] However, *de novo* sequencing has not been used to find SEPs or non-canonical immunopeptides encoded by sORFs to the best of our knowledge. An obvious obstacle is the considerable high proportion of false positive prediction in *de novo* sequencing. A systematic approach with a posteriori error control is urgently needed to tackle this obstacle, so that *de novo* sequencing can be used in non-canonical peptide identification. Recently, several research groups had independently reported identification of non-canonical peptides as tumor immunopeptides using

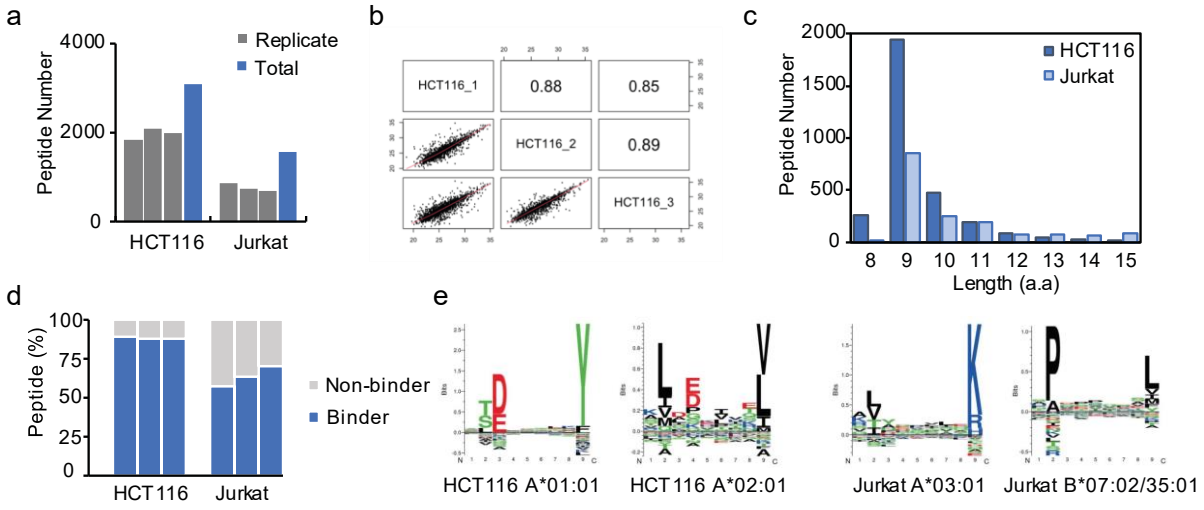


Figure 2. Comprehensive and reproducible identification of canonical immunopeptides. (a) Number of MHC-I peptides identified in each sample. (b) Pearson correlation of reproducibility among the same cell type. (c) Length distribution of MHC-I peptides. (d) Predicted affinity of peptides to MHC molecules (triplicates). (e) Clusters and sequence motifs of immunopeptides.

vastly different bioinformatics or proteomics approaches.[25, 35] Based on these prior studies, we have streamlined the workflow to improve the identification of non-canonical immunopeptide with Ribo-seq and MS analysis. We found that combining three bioinformatics pipelines to predict sORF from Ribo-seq data could improve the database-dependent peptide search. Meanwhile, *de novo* peptide sequencing followed by custom sORF database filtering could further expand the repertoire of non-canonical immunopeptides. With this approach, 308 non-canonical MHC-I peptides were identified in a colorectal carcinoma cell line HCT116 and an acute T cell leukemia cell line Jurkat (E6-1). The characteristics of these non-canonical immunopeptides highly resembled those of their canonical counterparts. Among the novel non-canonical immunopeptides, representative ones were extensively validated with multiple approaches including MS methods and MHC-I presentation in live cells.

MATERIALS AND METHODS

Chemicals and Materials. Jurkat (E6-1) cells were maintained in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS) and 1% antibiotics. HCT116 cells were in DMEM medium supplemented with 10% FBS and 1% antibiotics. T2-A02:01 monoallelic cells were maintained in IMDM medium supplemented with 10% fetal bovine serum (FBS) and 1% antibiotics. W6/32 monoclonal antibody was purchased from AtaGenix Laboratories Co., Ltd. (Wuhan). HLA-I alleles were examined using high-resolution genotyping (BGI). Peptide standards were purchased from GenScript Biotech. FcX™ (422301) was purchased from Biolegend, and anti-HLA-A, B, C (W6/32) antibody conjugated with PE (12-9983-42) was purchased from eBioscience.

MHC-I Immunopeptidome Sample Preparation. MHC-I peptidomes were obtained from established cell lines as described previously.[36] In each group, 1×10^8 cells were used for immunopeptides isolation. In brief, cell pellets were dissociated with lysis buffer with 0.25% sodium deoxycholate, 1% *n*-octyl glucoside, 100 mM PMSF, 0.2 mM iodoacetamide and protease inhibitors cocktail in Gibco's Dulbecco's phosphate-buffered saline (DPBS). Lysate were further cleared by centrifugation for 50 min at 17,000 g at 4 °C. The supernatant was purified with W6/32 antibody covalently bound to Protein-A Sepharose CL-4B beads. Beads were then washed with buffer A (150 mM NaCl, 20 mM Tris HCl) and 400 mM NaCl, 20 mM Tris HCl. The MHC-I complex was eluted with 1% TFA. Eluate was then loaded on Sep-Pak tC18 cartridges (Waters, 100 mg) and wash with 0.1% TFA and 2% ACN in 0.1% TFA, sequentially. The peptides were separated from MHC-I complexes on the tC18 cartridges by eluting with 28% ACN in 0.1% TFA and dried using vacuum centrifugation.

Immunopeptides Sequencing with Data-Dependent Acquisition. The fractionation was performed with a Shimadzu HPLC system based on a previous research.[37] The enriched immunopeptides were injected into a self-packed capillary column and separated by 70 min gradient of buffer A (5% ACN in ammonia) and buffer B (95% ACN in ammonia, pH 9.8). Each sample was fractionated into 6 fractions and dried on vacuum centrifuge. The fractionated immunopeptides were sequenced using Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific) coupled to an UltiMate 3000 HPLC (Thermo Fisher Scientific) based on our previous study.[38] Briefly, each sample was separated on a self-packed capillary column at a flowrate of 500 nL/min by 65 min gradient of buffer A (2% ACN and 0.1% formic acid) and buffer B (98% ACN and 0.1% formic acid). The peptides were ionized by 2K spray voltage in Orbitrap Fusion Lumos Tribrid mass spectrometer. The full scan spectra were measured with a resolution of 60,000 and auto gain control (AGC) of 1E5 within 50 ms max injection time, followed by top 30 MS2 scans with a resolution of 15,000 and AGC of 2E4 within the same 50 ms max injection time. The isolation window of MS2 scan was set to 1.6 *m/z* and only ions with 2-6 charges were triggered for the MS2 event. The normalized collision energy (NCE) was set as 30. The dynamic exclusion was set as 30s.

Validation of Immunopeptides by Parallel Reaction Monitoring Mode. To confirm the existence of immunopeptides detected from DDA data, both selected native immunopeptides and synthetic peptides were analyzed using Orbitrap Fusion

Table 1. HLA Typing of HCT116 and Jurkat Cells

	HLA-A*	HLA-A*	HLA-B*	HLA-B*
HCT116	01:01	02:01	18:02	45:01
Jurkat (E6-1)	03:01	-	07:02	35:01

Lumos Tribrid mass spectrometer coupled to an UltiMate 3000 UPLC. The samples were separated on a column (75 μm × 25 cm, 2 μm particle size) at a flowrate of 300 nL/min by 90 min gradient of the same buffer A and B. The full scan spectra were measured with a resolution of 60,000 and auto gain control (AGC) of 4E5 within 30 ms max injection time, followed by targeted peptides MS2 scans with a resolution of 15,000 and AGC of 1E5 within 50 ms max injection time under the 1.4 m/z isolation window. The normalized collision energy (NCE) was set as 30. The PRM data were processed with Skyline (20.2.0.343) software.

Identification of Canonical and Non-canonical Immunopeptides with *De Novo* Peptide Sequencing. PEAKS (Studio 10.5) was used to conduct *de novo* sequencing and peptide annotations. 10 ppm and 0.2 Da were set as error tolerances for the precursor ions and the fragment ions, respectively. No enzyme was selected. Carbamidomethylation of cysteines was set as a fixed modification and oxidation of methionine was set as a variable modification. The maximum number of modifications was two. For each spectrum, only the highest-ranking candidate peptide reported by PEAKS was further used. To ensure the non-canonical immunopeptides are unique, peptides that had 100% similarity against a database composed of RiboTISH, PRICE and sORF, and ≤ 80% similarity against annotated proteome (UniProtKB/SwissProt, Jan-2020) were considered as non-canonical immunopeptides using Blastp (2.9.0+). The peptides with > 80% similarity against SwissProt were considered as canonical immunopeptides.

Identification of Canonical and Non-canonical Immunopeptides with Database Search. The DDA data were separately searched against the library generated by the PRICE, sORF, RiboTISH, and SwissProt database (UniProtKB/SwissProt, Jan-2020) using MaxQuant (1.6.17.0). The common parameters were set as below: no contaminant database was included; the MS/MS match tolerance was set as 20 ppm; the digestion mode was set as unspecific; oxidation (M), acetyl (protein N-term) and carbamidomethyl (C) were included as variable modifications; the FDR was set as 1%. The identified peptides from MaxQuant were mapped against SwissProt using Blastp (2.9.0+) and only peptides with less than 80% overlapping rate were considered as non-canonical immunopeptides.

Ribo-seq Analysis and Database Construction. Four publicly available Ribo-seq datasets downloaded from the NCBI SRA database were used.[39] Preprocessing of Ribo-seq raw data consisted of adaptor removal using Cutadapt (v1.8.1),[40] low-quality trimming using Sickle (v1.33),[41] and removal of rRNA and tRNA contaminants using Bowtie2 (v2.3.5.1).[42] All remaining reads were mapped to the human reference genome (GENCODE, Release 28: GRCh38.p12) using STAR (v2.5.2)[43] with default parameters and further uniquely mapped reads were extracted. Non-canonical sORF detection was subsequently performed using RiboTISH (v0.2.1)[27] and PRICE (v1.0.3b)[44] with default parameters. To increase the statistical power of the sORF calling, the aligned BAM files for replicates of each cell line were merged with 'samtools merge' (v1.6). Finally, nucleic acid sequences of all actively translated sORFs were converted into amino acid sequences in the FASTA format for construction of MS/MS protein searching databases.

Assessment of Peptide-HLA binding. T2-A*02:01 cells were maintained in RPMI 1640 medium without any supplements. Peptides (10 μg/mL) were added into the culture medium and DMSO was used as a control. Antigen peptide MART-1 (ELAGIGILTV) was used as the positive control. A peptide binding to HLA-A*11:01 (SVSTVLTSK) was used as the negative control. Peptides were incubated with T2-A*02:01 cells at 37 °C for 2 h. Cells were collected for Fc receptor blockade, and stained by anti-HLA-A, B and C antibody. DAPI was used for counter-staining of live cells, and fluorescent signal were obtained and analyzed by flow cytometry. Data were normalized to DMSO group for quantification of peptide-HLA binding.

Prediction of Peptide-HLA Affinity. To evaluate the binding affinity of immunopeptides, netMHCpan 4.0 prediction software was run on all immunopeptides with length ranging from 8 to 15 amino acids.[45] Peptides with a rank ≤ 2% were considered as binders, and peptides with a rank ≤ 0.5% were considered as strong binders.

Hydrophobicity Index Calculation. Sequence-specific HI was calculated with the SSRCalc vQ.0 tool, which was available online at http://hs2.proteome.ca/SSRCalc/SSRCalcQ.html. Only unmodified peptides were included. Parameters were set as 100 Å C18 column, 0.1% formic acid separation system and without cysteine protection. Observed RTs were obtained from MaxQuant. If a peptide was detected multiple times in the same sample, the mean RT was used. Peptides and their mean RTs were plotted against the predicted HIs.

RESULTS AND DISCUSSION

An Efficient Approach to Identify Non-canonical MHC-I Immunopeptides. To achieve sensitive detection of non-canonical immunopeptides, we developed an integrative approach based on Ribo-seq, database search and *de novo* sequencing proteomics (Figure 1a). For this purpose, a colorectal carcinoma cell line HCT116, which was defined as mutation burden high (TMB-H) and microsatellite instability high (MSI-H) and thereby had been used previously to study cancer neoantigens, was chosen in our study as a reference.[46, 47] Another acute T cell leukemia cell line Jurkat, which was also TMB-H/MSI-H but displayed distinct HLA allotypes, was chosen as another representative cell type in our study.[48] Considering that these two cell types displayed distinct HLA allotypes, we expected to identify non-canonical peptides from different repertoires.

First, immunopeptides were enriched by co-immunoprecipitation with MHC antibody and analyzed with tandem mass spectrometry. Canonical immunopeptides were detected by searching MS data against UniProt human protein database. Canonical immunopeptide served as a benchmark to indicate successful enrichment of MHC-I peptides. Next, two complementary strategies, namely database search strategy and *de novo* sequencing, were applied to identify non-canonical immuno-

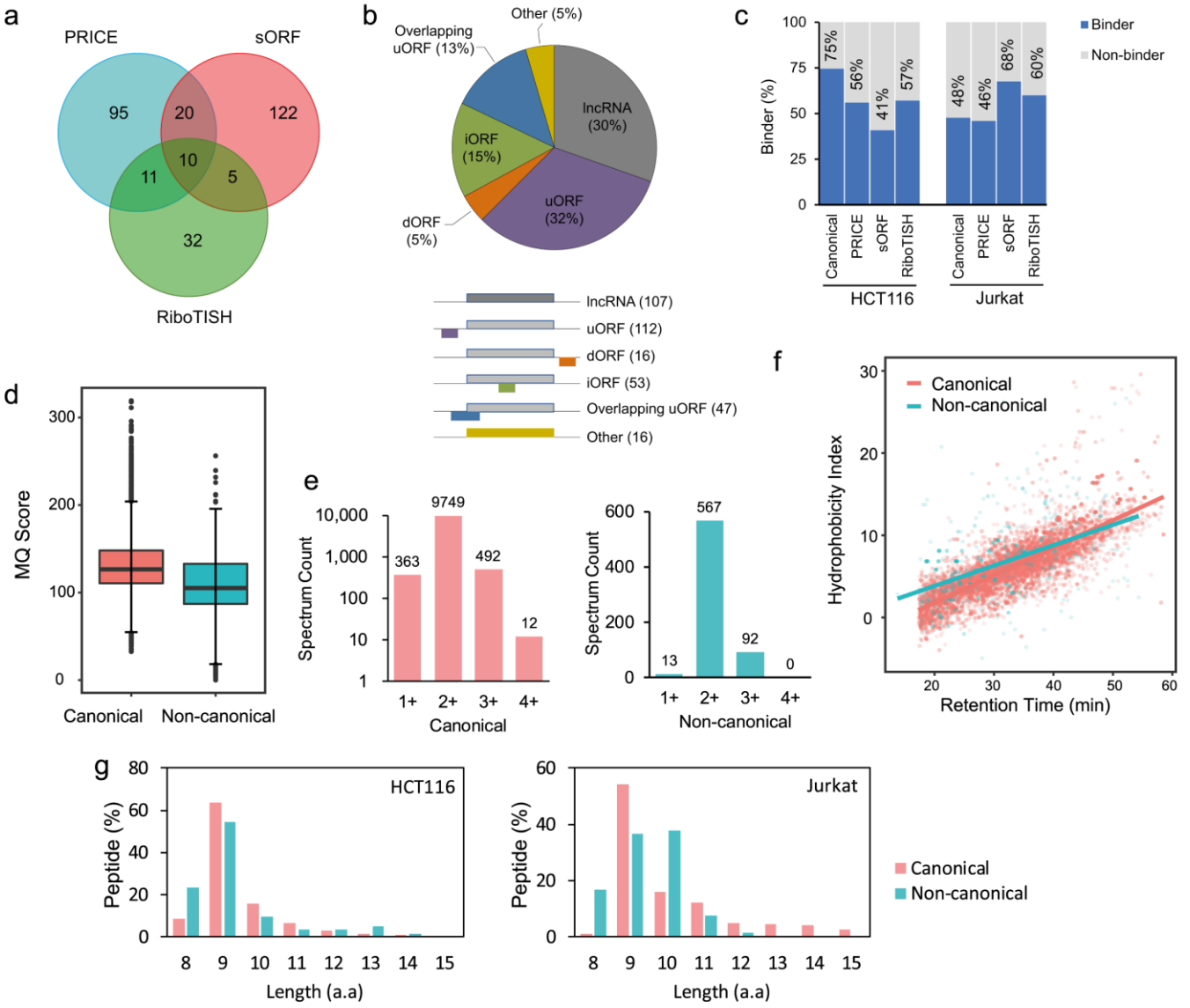


Figure 3. Non-canonical immunopeptides are encoded by sORFs in cancer cells. (a) Non-canonical immunopeptides are identified by database searching with the three different library construction pipelines (b) Non-canonical immunopeptides are translated from sORF at various genome regions, and (c) predicted MHC-I binder percentage. Comparison of canonical and non-canonical immunopeptides in terms of (d) MaxQuant score, (e) spectrum charge state, (f) retention time and (g) peptide length distribution.

peptides. As reference databases are critical in SEPs identification, they were constructed by using different algorithms to predict sORFs based on Ribo-seq results. Meanwhile, *de novo* peptide sequencing was used to provide additional non-canonical immunopeptides. The returned peptide sequences were further filtered against the custom sORF database and only those that were the most confident non-canonical immunopeptides were kept. Such a highly integrative approach enabled us to identify 308 non-canonical immunopeptides along with 7902 canonical and 3 mutation-bearing immunopeptides (Figure 1b). A considerable portion of these peptides were theoretical MHC binders (Figure 1c). Notably, two of the non-canonical immunopeptides had been reported previously.[9] The detailed workflow of the approach was elaborated as follows.

Canonical MHC-I Peptides as a Benchmark. MHC-I peptides were specifically enriched through immunoprecipitation with monoclonal MHC-I antibody.[36] With a peptide fractionation method, we identified 3,067 and 1,571 peptides in HCT116 and Jurkat, respectively (Figure 2a). The detection sensitivity and quantitative reproducibility of the canonical immunopeptides were demonstrated by high correlation between biological replicates (Figure 2b). The 8-11mers, which contributed to above 90% of the total canonical immunopeptides, fell in the typical length range of MHC-I peptides[36] (Figure 2c). According to the calculation with netMHCpan 4.0,[49] over 77% of these peptides were precited as MHC-I binders (Figure 2d) while 62% were as strong binders, indicating robust enrichment and detection of MHC-I peptides. Clustering[50, 51] based on peptide sequence similarity revealed 3 motifs from HCT116 and 4 motifs from Jurkat. Genotyping was carried out to ensure the HLA alleles of these two cells lines were identical to what had been recorded in previous database. Selected motifs with characteristics for 9-mer peptides were highly consistent with the expected amino acid distribution at anchor positions of HLA-

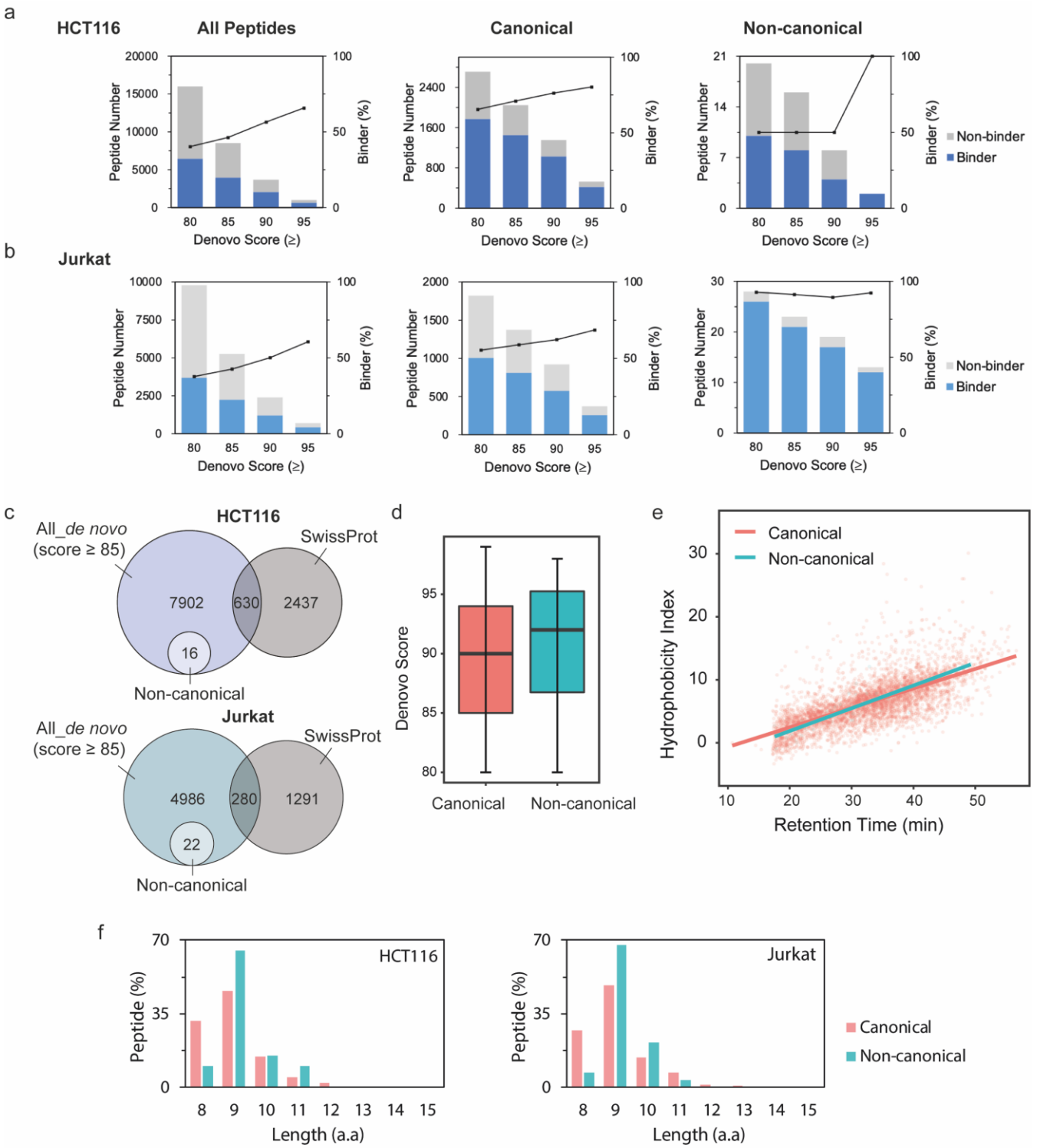


Figure 4. Identification of canonical and non-canonical MHC-I peptides with *de novo* sequencing. Number of peptides identified with *de novo* sequencing and their predicted affinity to MHC molecules from (a) HCT116 cells and (b) Jurkat cells. (c) Overlap between peptide sequences identified by *de novo* sequencing (*de novo* score≥85) and database search with SwissProt. Comparison between canonical and non-canonical immunopeptides in terms of (d) *de novo* score, (e) retention time, and (f) length identified by using *de novo* sequencing.

allele genotyping of these two cell lines[46, 52] (Figure 2e, Table 1).

Mutation-bearing neoantigen is also a very important class of immunopeptides. Therefore, we searched MS data against a library of computational predicted mutant immunopeptides.[53] Three mutation-bearing immunopeptides (QTDQMVFNTY, DEYTKFIPP, and EEEKFYLEP) were identified from HCT116, the TMB-H/MSI-H cell line, with 1% FDR and stringent manual spectra inspection (Figure S2). All three peptides were calculated as strong MHC binders with K*d* values less than 260 nM. Two of the three (QTDQMVFNTY) and (EEEKFYLEP) had been reported previously as mutated

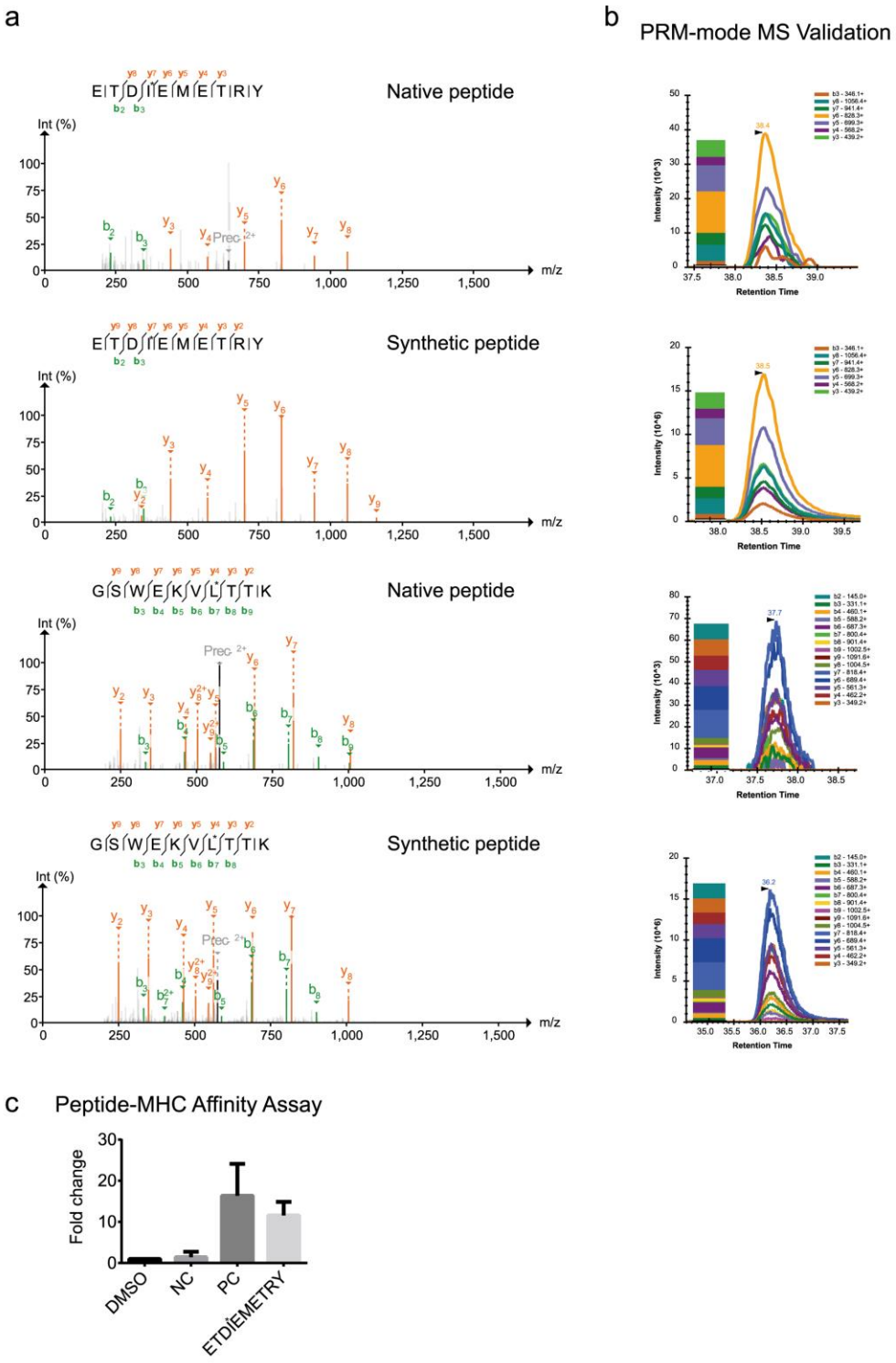


Figure 5. Validation of novel non-canonical peptides with MS and biochemical approaches. (a) Evidence of identification with synthetic peptides. (b) Representative peptide validation with targeted MS method. (c) Peptide-MHC affinity assay (NC, negative control; PC, positive control). *Possible I/L substitutions.

HLA-I peptides.[46] Altogether, successful identification of canonical and mutated immunopeptides demonstrated our sample preparation and data acquisition were robust enough for further discovery of non-canonical immunopeptides.

Detection of sORF-encoded Non-canonical Immunopeptides by Database Search. As good custom databases are critical for successful identification of sORF-encoded peptides, we first evaluated the performance of one established sORF database and two pipelines for predicting sORF from Ribo-seq data (Figure 3a). 157 non-canonical peptides were identified by searching MS data against the human database from sORF.org. To achieve cell-type specific detection, we also analyzed the

Ribo-seq data of HCT116 and Jurkat with both RiboTISH and PRICE pipelines (Figure S3). To our surprise, different sORF perdition algorithms of RiboTISH and PRICE led to distinct groups of sORFs, although their data pre-processing procedures were largely the same (Figure 3a). Subsequently, 58 and 136 peptides were identified by using these two sORF databases respectively, with 21 peptides in common. Besides, 10 non-canonical peptides were identified in common by using three databases, suggesting the complementary roles of different algorithms. With a smaller database size, PRICE offered highest identification number of non-canonical immunopeptide in the present study (Figure 3a). Consistent with a previous report,[16] lncRNAs and uORFs took the large majority of translated sORFs. iORFs from frame shift also contributed to 15% of novel peptide translation (Figure 3b). Affinity prediction by using netMHCpan indicated that 41% to 68% of non-canonical peptides were binders of MHC-I molecules. We observed that the percentage of binders in total immunopeptides was lower for non-canonical peptides compared to canonical peptides (Figure 3c). This could be explained by the fact that the original training dataset in netMHCpan was canonical peptides and thus led to a plausible bias. The non-canonical peptides highly resembled their canonical counterparts in terms of identification confidence, hydrophobicity, motif characters and length distribution (Figure 3d, f-g). These observations collectively implied that SEPs could have undergone the same antigen presentation pathway and been displayed on cell surface. It is noteworthy that the charge distribution varied between canonical and non-canonical immunopeptides, possibly because the non-canonical peptides had lower abundance in general and thus were more likely to be detected at doubly or triply charged state (Figure 3e).

***De novo* Peptide Sequencing Expanded the Repository of Non-canonical MHC-I Peptides.** So far, there is no such universal sORF database for non-canonical peptides and therefore many of them could not be identified through the database search strategy. As an alternative approach, *de novo* sequencing was used to analyze our MS data, resulting in the identification of 15,993 and 9,794 peptides with a score above 80 from HCT116 and Jurkat cells, respectively (Table S2). Non-canonical immunopeptides were identified by filtering them against our custom sORF databases. Only sequences that were 100% identical in sORF databases and had less than 80% similarity with annotated proteome (UniProtKB/SwissProt) were kept. Possibly due to smaller databases and lower conservation of sORFs, 20 and 28 non-canonical peptides were identified. For comparison, the whole peptide list was also aligned to UniProtKB/SwissProt protein database to find canonical immunopeptides. Additional 2,807 and 1,819 canonical peptides were identified from HCT116 and Jurkat cells, respectively (Figure 4a). Next, the affinity of all identified peptides to MHC molecules was calculated. We observed (i) the binder percentage increased along with *de novo* score cutoff, regardless of the immunopeptide type; (ii) sequences that could be matched with either UniProtKB /SwissProt or sORF databases had a higher binder percentage than those that were exclusively identified via *de novo* sequencing; (iii) over 90% of the non-canonical peptides identified via *de novo* sequencing were theoretical MHC binders when we set a stringent cut-off (*de novo* score ≥ 95) (Figures 4a and 4b).

Next, immunopeptides identified with *de novo* sequencing were compared with those identified via the conventional database search strategy. To our surprise, they were largely complementary with each other, with a small fraction in common. The number of canonical immunopeptides commonly identified with the two methods decreased from 702 to 121 when a more stringent cut-off was applied (Figure S4, *de novo* scores from 80 to 95). A score over 85 that resulted in reasonable identification number and MHC binding affinity was a relatively balanced criterion, and therefore was applied in all follow-up analyses (Figure 4c). Next, we compared the canonical and non-canonical immunopeptides in terms of their identification confidence, hydrophobicity and length distribution (Figure 4d-f). In contrast to the MaxQuant scores in database search strategy, scores of non-canonical peptides were generally higher than those of canonical peptides in *de novo* sequencing. The use of more stringent criteria in *de novo* sequencing to exclude false detection guaranteed the confident discovery of non-canonical peptides. Our observations above indicated that *de novo* sequencing had great potential in non-canonical peptide identification.

Validation of Non-canonical Immunopeptides with Multiple Approaches. We selected 8 peptides that were commonly identified with PRICE, sORF and RiboTISH for extensive validations. First, 6 out of the 8 peptide sequences were verified by matching their spectra with those of synthetic standard peptides side by side (Figure 5a and S6). Using MS method in data-dependent mode (DDA), these peptides were detected with the same retention time and *m/z* as their synthetic peptide standards. MS/MS spectra of these peptides were manually checked to assign the fragment ions. Next, the selected peptides were further validated by using an alternative MS method in targeted mode named parallel reaction monitoring (PRM) (Figure 5b and S7). As we expected, novel non-canonical immunopeptides provided us with identical precursor ions and transitions compared with their synthetic standards. Among the 8 chosen non-canonical immunopeptides picked, 7 were strong MHC-I binders as calculated by netMHCpan 4.0. The novel non-canonical immunopeptide ETDIEMETRY from HCT116 cells was assessed on its affinity to MHC-I molecule in live cells. T2-A*02:01 mono-allelic cells, which lacked endogenous antigen presentation pathway, were used to demonstrate the interaction between peptides and MHC molecules. Peptides were incubated with T2 cells and subsequently stained with MHC-I antibody for analysis by flow cytometry. The peptide ETDIEMETRY generated a considerable fluorescent signal change, suggesting that it was a strong binder of MHC-I molecule (Figure 5c).

CONCLUSION AND DISCUSSION

This work provides an efficient approach combining Ribo-seq and mass spectrometry to detect novel MHC-I peptides directly. In the present approach, database search and *de novo* sequencing were combined to identify hundreds of non-canonical immunopeptides. As there is a lack of universal sORF databases at the moment, we demonstrated that Ribo-seq data were effective for generating custom sORF databases by using complementary bioinformatics pipelines, without having to perform genomic sequencing, RNA-seq or exome-seq. In the *de novo* sequencing strategy, a workflow was designed to filter *de novo* results against the custom sORF databases, which improved the confidence of discovery.

This study is, to our best knowledge, the first one that uses *de novo* sequencing for identification of sORF-encoded peptides. We identified 308 non-canonical immunopeptides that were translated from sORFs. Many of them such as 5'-UTR or

lncRNA, were previously presumably non-coding regions. Considering that the total number of sORFs was estimated to be several folds higher than annotated ORFs, the non-canonical immunopeptides that were identified so far could be just a tip of the iceberg. We noticed that non-canonical peptides were not proportional to their canonical counterparts in the two cell lines. More canonical immunopeptides were identified in HCT116 while more non-canonical immunopeptides were identified in Jurkat cells. There were at least two possibilities. (i) The two cells have distinct HLA allotypes, which may have certain preference on non-canonical immunopeptides. (ii) There is a HLA haplotype loss in Jurkat.[48] The downregulated level of MHC molecules may resulted in a lower canonical immunopeptide number with the same cell input. Novel methods for non-canonical peptides identification are emerging, and they will expand the databases and the peptide list. With expanded knowledge, the selectivity of HLA alleles towards non-canonical immunopeptides will be investigated in due course.

This study evaluated the experimental settings for non-canonical immunopeptide identification and provided better options for following studies. We need to point out that there is still room for improvement. For example, a large portion of peptides from *de novo* sequencing could not be assigned. With the identification of more and more non-canonical peptide sequences, the machine learning methods may get better trained for *de novo* sequencing in future. The novel sequences of canonical and non-canonical immunopeptide identified here may also be useful for such training purpose.

ASSOCIATED CONTENT

Supporting Information

The data that support the findings of this study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD024415, and CNGB Sequence Archive (CNSA)[54] of China National GeneBank DataBase (CNGBdb)[55] with accession number CNP0001645.

The Supporting Information is available free of charge on the ACS Publications website.

Figure S1: Pearson correlation analysis for peptides across biological triplicates in Jurkat cells.
Figure S2: MS/MS spectra of mutated immunopeptides identified from HCT116.
Figure S3: Predicted sORF sequences in PRICE and RiboTISH-generated databases.
Figure S4: Peptide sequences identified by *de novo* sequencing (colored) with indicated cut-offs and bottom-up database search (gray) against SwissProt.
Figure S5: Contribution of different methods for identification of non-canonical immunopeptides.
Figure S6: MS/MS spectra of non-canonical immunopeptides and spectra of corresponding synthetic standards.
Figure S7: PRM-mode MS validation of the non-canonical immunopeptides.
Table S1: All canonical and non-canonical peptides identified by database search.
Table S2: All canonical peptides and non-canonical peptides identified by *de novo* sequencing.

AUTHOR INFORMATION

Corresponding Author

Qian Zhao - *State Key Laboratory of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong SAR, China; Phone: 852-34008711*
Email: q.zhao@polyu.edu.hk ORCID ID: 0000-0003-2244-6516

Author Contributions

All authors have given approval to the final version of the manuscript and declare no conflict of interest.

ACKNOWLEDGMENT

We acknowledge the funding support from NSFC 21705136, Research Grants Council (RGC)-Early Career Scheme 25301518, RGC-CRF Equipment C5033-19E, RGC-RIF R5050-18. We acknowledge the funding support from Laboratory for Synthetic Chemistry and Chemical Biology Limited under the Health@InnoHK Program launched by ITC, HKSAR. We thank the support from the State Key Laboratory of Chemical Biology and Drug Discovery in PolyU, the PolyU Research Facilities in Chemical and Environmental Analysis (UCEA) and Life Sciences (ULS), and China National GeneBank.

REFERENCES

(1) Hu, Z.; Ott, P. A.; Wu, C. J., Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nature Reviews Immunology* **2018,** *18* (3), 168.
(2) Hilf, N.; Kuttruff-Coqui, S.; Frenzel, K.; Bukur, V.; Stevanović, S.; Gouttefangeas, C.; Platten, M.; Tabatabai, G.; Dutoit, V.; van der Burg, S. H., Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature* **2019,** *565* (7738), 240-245.
(3) Keskin, D. B.; Anandappa, A. J.; Sun, J.; Tirosh, I.; Mathewson, N. D.; Li, S.; Oliveira, G.; Giobbie-Hurder, A.; Felt, K.; Gjini, E., Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **2019,** *565* (7738), 234-239.
(4) 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., An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **2017,** *547* (7662), 217-221.
(5) Sahin, U.; Derhovanessian, E.; Miller, M.; Kloke, B.-P.; Simon, P.; Löwer, M.; Bukur, V.; Tadmor, A. D.; Luxemburger, U.; Schrörs, B., Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **2017,** *547* (7662), 222-226.
(6) Rajasagi, M.; Shukla, S. A.; Fritsch, E. F.; Keskin, D. B.; DeLuca, D.; Carmona, E.; Zhang, W.; Sougnez, C.; Cibulskis, K.; Sidney, J., Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* **2014,** *124* (3), 453-462.
(7) Wang, R.-F.; Johnston, S. L.; Zeng, G.; Topalian, S. L.; Schwartzentruber, D. J.; Rosenberg, S. A., A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames. *The journal of Immunology* **1998,** *161* (7), 3596-3606.
(8) Robbins, P. F.; El-Gamil, M.; Li, Y. F.; Fitzgerald, E. B.; Kawakami, Y.; Rosenberg, S. A., The intronic region of an incompletely spliced gp100 gene transcript encodes an epitope recognized by melanoma-reactive tumor-infiltrating lymphocytes. *The Journal of Immunology* **1997,** *159* (1), 303-308.
(9) Laumont, C. M.; Vincent, K.; Hesnard, L.; Audemard, É.; Bonneil, É.; Laverdure, J.-P.; Gendron, P.; Courcelles, M.; Hardy, M.-P.; Côté, C.; Durette, C.; St-Pierre, C.; Benhammadi, M.; Lanoix, J.; Vobecky, S.; Haddad, E.; Lemieux, S.; Thibault, P.; Perreault, C., Noncoding regions are the main source of targetable tumor-specific antigens. *Science Translational Medicine* **2018,** *10* (470), eaau5516.
(10) Laumont, C. M.; Daouda, T.; Laverdure, J.-P.; Bonneil, É.; Caron-Lizotte, O.; Hardy, M.-P.; Granados, D. P.; Durette, C.; Lemieux, S.; Thibault, P., Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nature communications* **2016,** *7* (1), 1-12.

(11) 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.; Bachireddy, P.; Zervantonakis, I. K.; Rosenbluth, J. M.; Ouspenskaia, T.; Law, T.; Justesen, S.; Stevens, J.; Lane, W. J.; Eisenhaure, T.; Lan Zhang, G.; Clauser, K. R.; Hacohen, N.; Carr, S. A.; Wu, C. J.; Keskin, D. B., A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology* **2019**.

(12) Liepe, J.; Marino, F.; Sidney, J.; Jeko, A.; Bunting, D. E.; Sette, A.; Kloetzel, P. M.; Stumpf, M. P. H.; Heck, A. J. R.; Mishto, M., A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* **2016,** *354* (6310), 354.

(13) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J.-F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blöcker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H.-C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G. R.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F. A.; Stupka, E.; Szustakowki, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S.-P.; Yeh, R.-F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Patrinos, A.; Morgan, M. J.; International Human Genome Sequencing, C.; Whitehead Institute for Biomedical Research, C. f. G. R.; The Sanger, C.; Washington University Genome Sequencing, C.; Institute, U. D. J. G.; Baylor College of Medicine Human Genome Sequencing, C.; Center, R. G. S.; Genoscope; Cnrs, U. M. R.; Department of Genome Analysis, I. o. M. B.; Center, G. T. C. S.; Beijing Genomics Institute/Human Genome, C.; Multimegabase Sequencing Center, T. I. f. S. B.; Stanford Genome Technology, C.; University of Oklahoma's Advanced Center for Genome, T.; Max Planck Institute for Molecular, G.; Cold Spring Harbor Laboratory, L. A. H. G. C.; Biotechnology, G. B. G. R. C. f.; *Genome Analysis, G.; Scientific management: National Human Genome Research Institute, U. S. N. I. o. H.; Stanford Human Genome, C.; University of Washington Genome, C.; Department of Molecular Biology, K. U. S. o. M.; University of Texas Southwestern Medical Center at, D.; Office of Science, U. S. D. o. E.; The Wellcome, T., Initial sequencing and analysis of the human genome. *Nature* **2001,** *409* (6822), 860-921.

(14) Djebali, S.; Davis, C. A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; Xue, C.; Marinov, G. K.; Khatun, J.; Williams, B. A.; Zaleski, C.; Rozowsky, J.; Röder, M.; Kokocinski, F.; Abdelhamid, R. F.; Alioto, T.; Antoshechkin, I.; Baer, M. T.; Bar, N. S.; Batut, P.; Bell, K.; Bell, I.; Chakrabortty, S.; Chen, X.; Chrast, J.; Curado, J.; Derrien, T.; Drenkow, J.; Dumais, E.; Dumais, J.; Duttagupta, R.; Falconnet, E.; Fastuca, M.; Fejes-Toth, K.; Ferreira, P.; Foissac, S.; Fullwood, M. J.; Gao, H.; Gonzalez, D.; Gordon, A.; Gunawardena, H.; Howald, C.; Jha, S.; Johnson, R.; Kapranov, P.; King, B.; Kingswood, C.; Luo, O. J.; Park, E.; Persaud, K.; Preall, J. B.; Ribeca, P.; Risk, B.; Robyr, D.; Sammeth, M.; Schaffer, L.; See, L.-H.; Shahab, A.; Skancke, J.; Suzuki, A. M.; Takahashi, H.; Tilgner, H.; Trout, D.; Walters, N.; Wang, H.; Wrobel, J.; Yu, Y.; Ruan, X.; Hayashizaki, Y.; Harrow, J.; Gerstein, M.; Hubbard, T.; Reymond, A.; Antonarakis, S. E.; Hannon, G.; Giddings, M. C.; Ruan, Y.; Wold, B.; Carninci, P.; Guigó, R.; Gingeras, T. R., Landscape of transcription in human cells. *Nature* **2012,** *489* (7414), 101-108.

(15) Aspden, J. L.; Eyre-Walker, Y. C.; Phillips, R. J.; Amin, U.; Mumtaz, M. A. S.; Brocard, M.; Couso, J.-P., Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **2014,** *3*, e03528.

(16) Ji, Z.; Song, R.; Regev, A.; Struhl, K., Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **2015,** *4*, e08890.

(17) Prensner, J. R.; Enache, O. M.; Luria, V.; Krug, K.; Clauser, K. R.; Dempster, J. M.; Karger, A.; Wang, L.; Stumbraite, K.; Wang, V. M.; Botta, G.; Lyons, N. J.; Goodale, A.; Kalani, Z.; Fritchman, B.; Brown, A.; Alan, D.; Green, T.; Yang, X.; Jaffe, J. D.; Roth, J. A.; Piccioni, F.; Kirschner, M. W.; Ji, Z.; Root, D. E.; Golub, T. R., Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nature Biotechnology* **2021**.

(18) Koh, M.; Ahmad, I.; Ko, Y.; Zhang, Y.; Martinez, T. F.; Diedrich, J. K.; Chu, Q.; Moresco, J. J.; Erb, M. A.; Saghatelian, A.; Schultz, P. G.; Bollong, M. J., A short ORF-encoded transcriptional regulator. *Proceedings of the National Academy of Sciences* **2021,** *118* (4), e2021943118.

(19) Jackson, R.; Kroehling, L.; Khitun, A.; Bailis, W.; Jarret, A.; York, A. G.; Khan, O. M.; Brewer, J. R.; Skadow, M. H.; Duizer, C.; Harman, C. C. D.; Chang, L.; Bielecki, P.; Solis, A. G.; Steach, H. R.; Slavoff, S.; Flavell, R. A., The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **2018,** *564* (7736), 434-438.

(20) Slavoff, S. A.; Heo, J.; Budnik, B. A.; Hanakahi, L. A.; Saghatelian, A., A Human Short Open Reading Frame (sORF)-encoded Polypeptide That Stimulates DNA End Joining *. *Journal of Biological Chemistry* **2014,** *289* (16), 10950-10957.

(21) Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M. N.; Ma, J.; Levin, J. Z.; Karger, A. D.; Budnik, B. A.; Rinn, J. L.; Saghatelian, A., Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology* **2013,** *9* (1), 59-64.

(22) Ma, J.; Ward, C. C.; Jungreis, I.; Slavoff, S. A.; Schwaid, A. G.; Neveu, J.; Budnik, B. A.; Kellis, M.; Saghatelian, A., Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of proteome research* **2014,** *13* (3), 1757-1765.

(23) Wang, S.; Tian, L.; Liu, H.; Li, X.; Zhang, J.; Chen, X.; Jia, X.; Zheng, X.; Wu, S.; Chen, Y.; Yan, J.; Wu, L., Large-Scale Discovery of Non-conventional Peptides in Maize and Arabidopsis through an Integrated Peptidogenomic Pipeline. *Molecular Plant* **2020,** *13* (7), 1078-1093.

(24) Oyama, M.; Kozuka-Hata, H.; Suzuki, Y.; Semba, K.; Yamamoto, T.; Sugano, S., Diversity of translation start sites may define increased complexity of the human short ORFeome. *Molecular & cellular proteomics : MCP* **2007,** *6* (6), 1000-6.

(25) Chong, C.; Müller, 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.; Simó-Riudalbas, L.; Planet, E.; Lubiński, J.; Bryśkiewicz, M.; Wiznerowicz, M.; Xenarios, I.; Zhang, L.; Trono, D.; Harari, A.; Ohler, U.; Coukos, G.; Bassani-Sternberg, M., Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nature Communications* **2020,** *11* (1), 1293.

(26) Martinez, T. F.; Chu, Q.; Donaldson, C.; Tan, D.; Shokhirev, M. N.; Saghatelian, A., Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology* **2020,** *16* (4), 458-468.

(27) Zhang, P.; He, D.; Xu, Y.; Hou, J.; Pan, B. F.; Wang, Y.; Liu, T.; Davis, C. M.; Ehli, E. A.; Tan, L.; Zhou, F.; Hu, J.; Yu, Y.; Chen, X.; Nguyen, T. M.; Rosen, J. M.; Hawke, D. H.; Ji, Z.; Chen, Y., Genome-wide identification and differential analysis of translational initiation. *Nat Commun* **2017,** *8* (1), 1749.

(28) Erhard, F.; Halenius, A.; Zimmermann, C.; L'Hernault, A.; Kowalewski, D. J.; Weekes, M. P.; Stevanovic, S.; Zimmer, R.; Dölken, L., Improved Ribo-seq enables identification of cryptic translation events. *Nature Methods* **2018,** *15* (5), 363-366.

(29) Olexiouk, V.; Van Criekinge, W.; Menschaert, G., An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* **2017,** *46* (D1), D497-D502.

(30) Li, S.; DeCourcy, A.; Tang, H., Constrained De Novo Sequencing of neo-Epitope Peptides using Tandem Mass Spectrometry. *Res Comput Mol Biol* **2018,** *10812*, 138-153.

(31) Wei, X.; Wang, S.; Li, Z.; Li, Z.; Qu, Z.; Wang, S.; Zou, B.; Liang, R.; Xia, C.; Zhang, N., Peptidomes and Structures Illustrate Two Distinguishing Mechanisms of Alternating the Peptide Plasticity Caused by Swine MHC Class I Micropolymorphism. *Frontiers in Immunology* **2021,** *12* (155).

(32) Faridi, P.; Li, C.; Ramarathinam, S. H.; Vivian, J. P.; Illing, P. T.; Mifsud, N. A.; Ayala, R.; Song, J.; Gearing, L. J.; Hertzog, P. J.; Ternette, N.; Rossjohn, J.; Croft, N. P.; Purcell, A. W., A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Science Immunology* **2018,** *3* (28), eaar3947.

(33) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Shan, B.; Li, M., Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence* **2020,** *2* (12), 764-771.

(34) Karunratanakul, K.; Tang, H.-Y.; Speicher, D. W.; Chuangsuwanich, E.; Sriswasdi, S., Uncovering Thousands of New Peptides with Sequence-Mask-Search Hybrid De Novo Peptide Sequencing Framework. *Molecular & Cellular Proteomics* **2019,** *18* (12), 2478-2491.

(35) Ouspenskaia, T.; Law, T.; Clauser, K. R.; Klaeger, S.; Sarkizova, S.; Aguet, F.; Li, B.; Christian, E.; Knisbacher, B. A.; Le, P. M.; Hartigan, C. R.; Keshishian, H.; Apffel, A.; Oliveira, G.; Zhang, W.; Chow, Y. T.; Ji, Z.; Shukla, S. A.; Bachireddy, P.; Getz, G.; Hacohen, N.; Keskin, D. B.; Carr, S. A.; Wu, C. J.; Regev, A., Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. *bioRxiv* **2020**, 2020.02.12.945840.

(36) Chong, C.; Marino, F.; Pak, H.; Racle, J.; Daniel, R. T.; Muller, M.; Gfeller, D.; Coukos, G.; Bassani-Sternberg, M., High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferongamma-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. *Molecular & cellular proteomics : MCP* **2018,** *17* (3), 533-548.

(37) Zhang, Y.; Lin, Z.; Hao, P.; Hou, K.; Sui, Y.; Zhang, K.; He, Y.; Li, H.; Yang, H.; Liu, S.; Ren, Y., Improvement of Peptide Separation for Exploring the Missing Proteins Localized on Membranes. *Journal of proteome research* **2018,** *17* (12), 4152-4159.

(38) Zhang, Y.; Zhang, K.; Bu, F.; Hao, P.; Yang, H.; Liu, S.; Ren, Y., D283 Med, a Cell Line Derived from Peritoneal Metastatic Medulloblastoma: A Good Choice for Missing Protein Discovery. *Journal of proteome research* **2020,** *19* (12), 4857-4866.

(39) Kodama, Y.; Shumway, M.; Leinonen, R., The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **2012,** *40* (Database issue), D54-6.

(40) Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **2011,** *17* (1), 10-12.

(41) Joshi, N.; Fass, J., Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software]. 2011.

(42) Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nat Methods* **2012,** *9* (4), 357-9.

(43) Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R., STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **2013,** *29* (1), 15-21.

(44) Erhard, F.; Halenius, A.; Zimmermann, C.; L'Hernault, A.; Kowalewski, D. J.; Weekes, M. P.; Stevanovic, S.; Zimmer, R.; Dölken, L., Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **2018,** *15* (5), 363-366.

(45) Jurtz, V.; Paul, S.; Andreatta, M.; Marcatili, P.; Peters, B.; Nielsen, M., NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of immunology (Baltimore, Md. : 1950)* **2017,** *199* (9), 3360-3368.

(46) Bassani-Sternberg, M.; Pletscher-Frankild, S.; Jensen, L. J.; Mann, M., Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular & cellular proteomics : MCP* **2015,** *14* (3), 658-73.

(47) Chen, R.; Fauteux, F.; Foote, S.; Stupak, J.; Tremblay, T.-L.; Gurnani, K.; Fulton, K. M.; Weeratna, R. D.; Twine, S. M.; Li, J., Chemical Derivatization Strategy for Extending the Identification of MHC Class I Immunopeptides. *Analytical Chemistry* **2018,** *90* (19), 11409-11416.

(48) Masuda, K.; Hiraki, A.; Fujii, N.; Watanabe, T.; Tanaka, M.; Matsue, K.; Ogama, Y.; Ouchida, M.; Shimizu, K.; Ikeda, K., Loss or down‐regulation of HLA class I expression at the allelic level in freshly isolated leukemic blasts. *Cancer science* **2007,** *98* (1), 102-108.

(49) Jurtz, V.; Paul, S.; Andreatta, M.; Marcatili, P.; Peters, B.; Nielsen, M., NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology* **2017,** *199* (9), 3360.

(50) Andreatta, M.; Alvarez, B.; Nielsen, M., GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res* **2017,** *45* (W1), W458-W463.

(51) Andreatta, M.; Lund, O.; Nielsen, M., Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics (Oxford, England)* **2013,** *29* (1), 8-14.

(52) Nicastri, A.; Liao, H.; Muller, J.; Purcell, A. W.; Ternette, N., The Choice of HLA-Associated Peptide Enrichment and Purification Strategy Affects Peptide Yields and Creates a Bias in Detected Sequence Repertoire. *Proteomics* **2020,** *20* (24), 2070175.

(53) Boegel, S.; Löwer, M.; Bukur, T.; Sahin, U.; Castle, J. C., A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *OncoImmunology* **2014,** *3* (8), e954893.

(54) Guo, X.; Chen, F.; Gao, F.; Li, L.; Liu, K.; You, L.; Hua, C.; Yang, F.; Liu, W.; Peng, C.; Wang, L.; Yang, X.; Zhou, F.; Tong, J.; Cai, J.; Li, Z.; Wan, B.; Zhang, L.; Yang, T.; Zhang, M.; Yang, L.; Yang, Y.; Zeng, W.; Wang, B.; Wei, X.; Xu, X., CNSA: a data repository for archiving omics data. *Database : the journal of biological databases and curation* **2020,** *2020*.

For Table of Contents Only

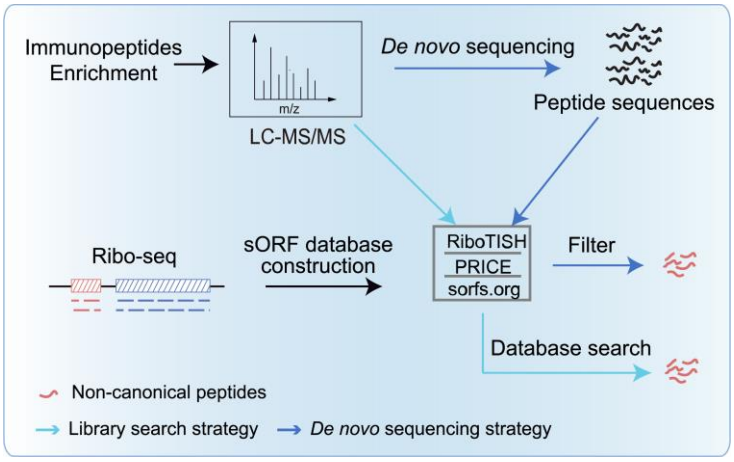
An Integrated Approach for Discovering Non-canonical MHC-I Peptides Encoded by Small Open Reading Frames

Lei Chen[2], Yuanliang Zhang[1], Ying Yang[1], Yang Yang, Huihui Li[3], Xuan Dong[4], Hongwei Wang[3], Zhi Xie[3], Qian Zhao[1*]

1. State Key Laboratory of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong SAR, China
2. Laboratory for Synthetic Chemistry and Chemical Biology Limited, Hong Kong SAR, China
3. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
4. BGI-Shenzhen, Shenzhen 518083, China

*Email: q.zhao@polyu.edu.hk

KEYWORDS: MHC-I peptides, small open reading frames, Ribo-seq, database search, *de novo* sequencing



An efficient approach that implements complementary bioinformatic strategies to improve the identification of non-canonical MHC-I peptides.