Immunopeptidogenomics: Harnessing RNA-Seq to Illuminate the Dark Immunopeptidome

Authors

Katherine E. Scull, Kirti Pandey, Sri H. Ramarathinam, and Anthony W. Purcell

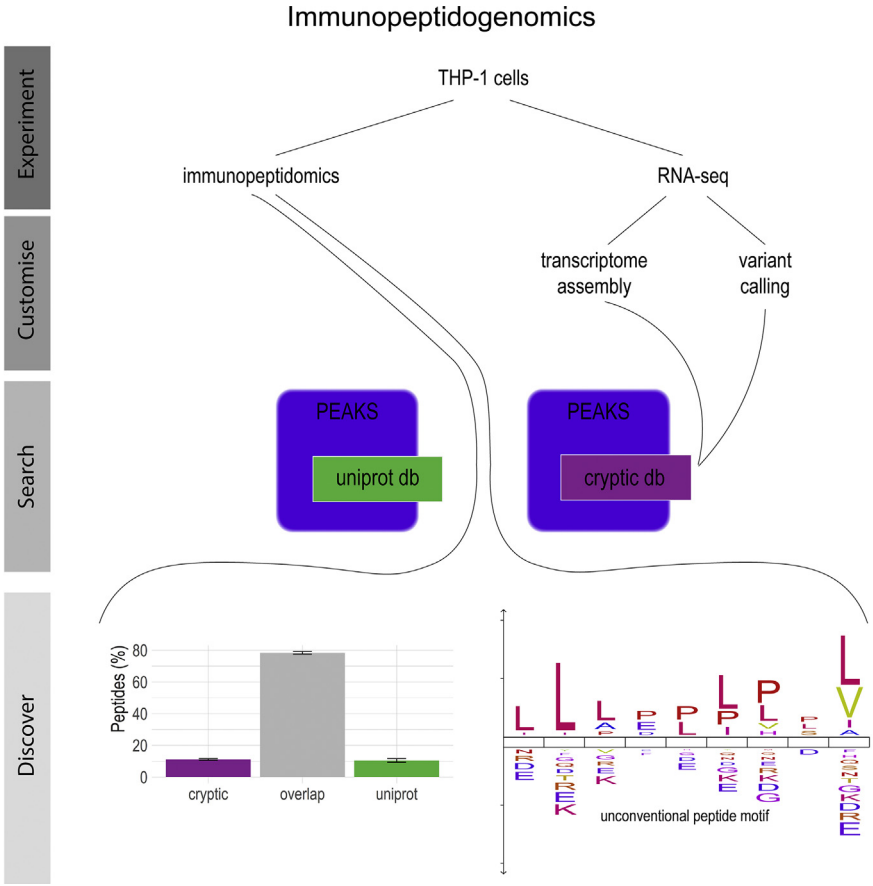
Correspondence

sri.ramarathinam@monash.edu; anthony.purcell@monash.edu

In Brief

Here, we describe a workflow for the inclusion of noncanonical antigen sources in a database for the analysis of immunopeptidomics data. The approach utilizes a series of executable programs that interrogate RNA-Seq data from the tissue of cell-based starting material for the immunopeptidomics experiment to generate a bespoke searchable FASTA database. We show the application of this workflow for the THP-1 leukemia cell line with the identification of a number of validated cancer-associated peptide antigens.

Graphical Abstract



Highlights

- Our workflow uses RNA-Seq data to generate a bespoke database for immunopeptidomics.
- Peptides (7%) in the THP-1 immunopeptidome are derived from noncanonical antigens.
- These include peptides from alternative reading frames or noncoding genomic regions.

2021, Mol Cell Proteomics *20*, 100143

https://doi.org/10.1016/j.mcpro.2021.100143

Immunopeptidogenomics: Harnessing RNA-Seq to Illuminate the Dark Immunopeptidome

Katherine E. Scull, Kirti Pandey, Sri H. Ramarathinam[‡*], and Anthony W. Purcell[‡*]

Human leukocyte antigen (HLA) molecules are cell-surface glycoproteins that present peptide antigens on the cell surface for surveillance by T lymphocytes, which contemporaneously seek signs of disease. Mass spectrometric analysis allows us to identify large numbers of these peptides (the immunopeptidome) following affinity purification of solubilized HLA–peptide complexes. However, in recent years, there has been a growing awareness of the "dark side" of the immunopeptidome: unconventional peptide epitopes, including neoepitopes, which elude detection by conventional search methods because their sequences are not present in reference protein databases (DBs). Here, we establish a bioinformatics workflow to aid identification of peptides generated by noncanonical translation of mRNA or by genome variants. The workflow incorporates both standard transcriptomics software and novel computer programs to produce cell line–specific protein DBs based on three-frame translation of the transcriptome. The final protein DB also includes sequences resulting from variants determined by variant calling on the same RNA-Seq data. We then searched our experimental data against both transcriptome-based and standard DBs using PEAKS Studio (Bioinformatics Solutions, Inc). Finally, further novel software helps to compare the various result sets arising for each sample, pinpoint putative genomic origins for unconventional sequences, and highlight potential neoepitopes. We applied the workflow to study the immunopeptidome of the acute myeloid leukemia cell line THP-1, using RNA-Seq and immunopeptidome data. We confidently identified over 14,000 peptides from three replicates of purified HLA peptides derived from THP-1 cells using the conventional UniProt human proteome. Using the transcriptome-based DB generated using our workflow, we recapitulated >85% of these and also identified 1029 unconventional peptides not explained by UniProt, including 16 sequences caused by nonsynonymous variants. Our workflow, which we term "immunopeptidogenomics," can provide DBs, which include pertinent unconventional sequences and allow neoepitope discovery, without becoming too large to search. Immunopeptidogenomics is a step toward unbiased search approaches that are needed to illuminate the dark side of the immunopeptidome.

A number of recent developments are stirring interest and changing perspectives in the field of immunopeptidomics, which is the study of peptides presented at the cell surface by major histocompatibility complex (MHC) molecules (human leukocyte antigen [HLA] molecules in humans) for immune surveillance. In cancer immunotherapy, immune checkpoint blockade has shown some success in freeing T cells to attack and control tumors, stimulating efforts to discover cancer-specific HLA-binding peptides targeted during immunotherapy for future use in cancer vaccines (see reviews (1–3)). In genomics, next-generation sequencing and ribosomal profiling are dramatically correcting fundamental understandings about how genes work; we now know that a large proportion of the genome is transcribed, and much more RNA is translated than current genome annotations and category labels would suggest (reviewed in Ref. (4)). The resulting peptides may also be available for presentation on HLA and recognition by the immune system, thus broadening the pool of potential immunogenic epitopes (5, 6). Yet such peptides cannot be discovered by mass spectrometry using routine bioinformatics workflows; they belong to the so-called "dark" immunopeptidome (7). Thus, these developments further underpin the exciting potential of illuminating the dark immunopeptidome.

What makes the dark immunopeptidome dark? Immunopeptidomics studies isolate HLA-bound peptides and analyze them by mass spectrometry. Generally, these studies employ techniques and software developed for proteomics, which means that they are suboptimal for immunopeptidomics in various ways. Search algorithms match experimental spectra against theoretical spectra inferred from protein databases (DBs; such as UniProt), relying on the assumption that the protein DB closely reflects the sample being analyzed (8). Dark immunopeptides, by definition, are absent from standard

From the Department of Biochemistry and Molecular Biology and Infection and Immunity Program, Monash Biomedicine Discovery Institute, Monash University, Clayton, Victoria, Australia

‡Joint senior authors.

*For correspondence: Sri H. Ramarathinam, sri.ramarathinam@monash.edu; Anthony W. Purcell, anthony.purcell@monash.edu.

https://doi.org/10.1016/j.mcpro.2021.100143

protein DBs; hence, conventional studies have mostly been unable to identify them. They include peptides derived from somatic mutations, from post-translational peptide splicing, and "cryptic" peptides caused by noncanonical translation such as from alternative reading frames or noncoding genomic regions (7). However, we cannot add every imaginable peptide sequence to the DB used to interrogate mass spectrometry data, since overinflating the size of the DB increases false discovery rates (FDRs) (9). That is, we should achieve the most accurate and statistically powerful analyses for immunopeptidomics using DBs that most closely resemble the source antigen complement of the sample. However, the composition of the dark immunopeptidome remains uncertain. It is unclear what a truly sample-specific immunopeptidome DB would actually look like.

Immunopeptidomics researchers have used various strategies to produce more inclusive and customized DBs while limiting search-space inflation. Most of these studies have examined only one category of dark peptide source at a time. Liepe *et al.* (10) and our own group have pioneered different strategies to make customized DBs including potential post-translationally spliced peptides. Discovering these peptides by mass spectrometric immunopeptidomics is particularly fraught, since allowing splicing of any peptide fragments from any protein source causes an astronomical inflation of the search space. Liepe *et al.* (10) and Faridi *et al.* (11) invented different strategies, which both aimed at including only the most likely potential spliced sequences in DBs. In contrast, researchers seeking neoepitopes by mass spectrometry have used approaches similar to standard (onco-)proteogenomics, deriving both RNA-Seq and mass spectrometric data from tumor samples to create customized DBs, which take into account nonsynonymous single nucleotide variants (ns-SNVs) and noncomplex indels (reviewed in Refs. (12, 13)). Laumont *et al.* (14) used not only similar customized DBs but also included peptides encoded by three-frame translation at variant sites and other tumor-specific RNA sequences. In this way, they found many tumor-specific peptides derived from what were considered noncoding regions of RNA. Most recently, Chong *et al.* (15) used a very comprehensive range of technologies to build various DBs for seeking HLA-binding peptides derived from noncanonical translation in tumors, whereas Erhard *et al.* (16) used a purely bioinformatics strategy to the same end, in which their "stratified" DB imposed presumed likelihoods on various categories of unconventional peptides.

We wished to produce an unbiased bioinformatics workflow, which could be used routinely to foster a broad illumination of the dark immunopeptidome in many tissue types, in health and disease. Our workflow needed to seek various kinds of unconventional as well as conventional peptides simultaneously, yet require minimal extra experimental work and expense. It should also be flexible to accommodate the constantly evolving software and best practices in transcriptomics, with necessary in-house software publicly available to allow general use. Many published proteogenomics software/workflows seek to simplify the bioinformatics for creating customized DBs, yet none precisely fit the dark immunopeptidome's unique requirements (*e.g.*, Refs. (17–26)). The immunopeptidogenomics workflow presented here needs only sample-specific RNA-Seq data to produce a single customized DB incorporating conventional peptides, neopeptides from SNVs/polymorphisms and noncomplex indels, and cryptic peptides derived from noncanonical translation, such as alternative reading frames and noncoding RNA regions/transcripts. We test its utility with the acute myeloid leukemia (AML) cell line THP-1, comparing identifications from the same mass spectrometric immunopeptidome data searched against the customized RNA-Seq–based DB or against a standard Swiss-Prot/TrEMBL DB.

EXPERIMENTAL PROCEDURES

Cell Lines

THP-1 cells were maintained in RF10 (RPMI1640 [Gibco]) supplemented with 2 mM minimum essential medium nonessential amino acid solution (Gibco), 100 mM Hepes (Gibco), 2 mM L-glutamine (Gibco), penicillin/streptomycin (Gibco), 50 μM 2-mercaptoethanol (Sigma–Aldrich), and 10% heat-inactivated fetal calf serum (Sigma–Aldrich). The HLA class I typing of THP-1 cell line was confirmed as homozygous A*02:01, B*15:11, and C*03:03 (Victorian Transplantation and Immunogenetics Service).

RNA-Seq Data

To collect RNA-Seq data, RNA was isolated from three biological replicates of THP-1 cell line using RNeasy mini kit (Qiagen) using 1e5 cells as per the manufacturer's instructions. Sequencing was performed at Micromon Genomics (Monash University), wherein the RNA first passed a quality control check, as evident from high RNA integrity score. The library was prepared using an MGIEasy-stranded mRNA chemistry V2 kit, and sequencing used MGITech MGISEQ2000RS hardware, MGIEasy V3 chemistry, and paired-end 100b reads. The three RNA-Seq replicates each yielded 20 to 34 million reads (replicate 1: 21,230,421 reads; replicate 2: 26,646,795 reads; and replicate 3: 34,146,239 reads) totaling >80 million reads, of which >75 million reads were aligned by STAR.

Experimental Design and Statistical Rationale

We acquired triplicate RNA-Seq data from THP-1 cells to make a cryptic DB using our bioinformatics workflow. We used the DB to analyze triplicate previously published tandem mass spectrometry data from the HLA-A*02:01 immunopeptidome of THP-1 cells (27). In parallel, we searched the data against a standard UniProt DB as described later. To determine statistical significance, two-way ANOVA with Tukey's multiple comparisons were used as described in appropriate sections. An FDR of 5% was used to establish a peptide identification threshold for both DB searches.

Software

The immunopeptidogenomics workflow presented here comprises both standard and novel steps to produce "cryptic" protein DBs (Fig. 1). We have written software for the novel steps and also to aid interpretation of results after searching mass spectrometry data

against both standard and cryptic DBs. The novel software are freely available with full source code and compilation and usage instructions from github.com/kescull/immunopeptidogenomics. Novel software include *alt_liftover*, *curate_vcf*, *db_compare.R*, *filter_FPKM*, *msDot*, *origins*, *revert_headers*, *squish*, and *triple_translate*. *db_compare.R* is an R script; all others are programs written in C. The remaining software must be sourced separately but may be interchanged for the latest versions/algorithms. We used open source software MultiQC, version 1.5 (28), STAR 2.5.2b (29), Cufflinks 2.2.1 (30, 31), gff3sort (32), gffread 0.11.9 (33), RSeQC, version 3.0.0 (infer_experiment.py) (34), the Genome Analysis Toolkit (GATK) 4.1.4.1 for various tools (35), and commercial software PEAKS Studio 10.0 build 20190201 (Bioinformatics Solutions, Inc) (36, 37).

Graphs were produced using R, Python, or GraphPad Prism (GraphPad Software Inc), version 9.0.0 for Windows, apart from Venn diagrams produced with BioVenn (38), and peptide motifs were produced with iceLogo (39). Binding affinity was predicted using NetMHC 4.0 (40, 41).

Immunopeptidogenomics Workflow

As shown in Figure 1, RNA-Seq data were utilized in two ways to produce a cryptic protein DB including both unconventional and known translation products as well as variants. The following sections provide a brief description of our methods; see the Supplemental Methods section for a detailed account of parameters and commands used.

Transcriptome Assembly

STAR in two-pass mode mapped reads from the pooled replicates against the GRCh38 reference genome, then Cufflinks was run in RABT mode to assemble the reads using the GENCODE, version 29, human primary assembly annotation as a guide. Cuffcompare from the Cufflinks suite was used to categorize transcripts in relation to the reference annotations, minimize redundancy, and produce a "tracking" file for later use. At this point, filter_FPKM may be employed to remove transcripts based on low transcript expression levels/evidence. However, for this study, we proceeded with an unfiltered transcriptome assembly, which therefore included many reference transcripts lacking RNA-Seq expression evidence, as well as novel transcripts.

Variant Calling

Variant calling was carried out in accordance with the GATK Best Practices Workflow "RNAseq short variant discovery (SNPs + Indels)" (created January 9, 2018; updated July 11, 2019; accessed November 14, 2019), adapted for use with an unpaired tumor cell line, for example, using Mutect2 instead of HaplotypeCaller to allow for unpredictable ploidy. Thus, STAR in two-pass mode mapped reads for each replicate in turn. Alignment files were preprocessed into analysis-ready BAM files using various GATK tools, which added read group and sample information to each file before the replicates were merged during base recalibration. Variants were called using Mutect2 in tumor-only mode, then filtered using FilterMutectCalls. Command-line text editing (awk) was used to select "PASS" variants.

DB Building

The schematic in Figure 1*B* shows how we built the cryptic protein DB. We converted the transcriptome assembly into a FASTA protein DB by using gffread to write out the transcript complementary DNA sequences, then triple_translate to translate the whole transcripts in all three frames (bar those lacking directionality), and to print out all >7 amino acid sequences, assuming no read through of stop codons. For variant incorporation, the variant file was first curated using curate_vcf to handle cases where a deletion variant removed the site of another variant(s) downstream. This produced two variant files we termed "indel" (which included such deletion variants) and "unmasked" (which ignored these variants in favor of downstream variants). Each variant file was used to produce a parallel transcriptome-based protein DB: first, the GATK tool FastaAlternateReferenceMaker incorporated the variants to form an alternate genome; second, alt_liftover revised the assembly file so that the coordinates referred to this alternate genome. gffread could thus write out the variant-containing transcriptomes using each alternate genome and corresponding assembly. Files containing other known variants might be utilized in the same fashion to produce further transcriptome-based protein DBs. Finally, all transcriptome-based protein DBs written by triple_translate were merged, duplicates were removed, and redundancy was reduced using squish. Furthermore, squish concatenates sequences of $\leq$300 amino acids, inserting a linker sequence of five consecutive Trp (WWWWW) between them, to form pseudoproteins of ~600 amino acids to ensure they are not excluded because of sequence limits for PEAKS DB searches. For the "standard" DB, we downloaded the UniProt *Homo sapiens* proteome (UP000005640; FASTA [canonical and isoform]) on June 5, 2020. Both DBs were appended with the 11 indexed retention time (iRT) peptide sequences (42).

Mass Spectrometry Data Searching

The cryptic and standard DBs were validated for use by PEAKS, showing no invalid entries. All fractions for each experiment were pooled for searching by PEAKS, but each replicate was searched separately against each DB in parallel. The following search parameters were used: parent mass error tolerance—10 ppm; fragment mass error tolerance—0.02 Da; precursor mass search type—monoisotopic; enzyme—none; digest mode—unspecific; variable modifications—oxidation (M), deamidation (N, Q); maximum variable post-translationally modified (PTM) per peptide—3; and FDR estimation—enabled.

Post-PEAKS Analysis

All peptide-spectrum match (PSM) results from PEAKS searches with each DB (cryptic or standard) were exported in text format "DB search psm.csv," and threshold scores for 5% FDR as calculated by PEAKS were noted. We chose the relatively permissive threshold of 5% FDR to avoid unduly filtering out true identifications, as FDR analysis is not ideal for immunopeptidomics. We filtered for 7 to 15mers as a means of removing most nonspecific peptides, since peptides of other lengths are unlikely to bind HLA-A*02:01. We identified and removed some further nonspecific peptides by analyzing the eluate from a precolumn used during THP-1 immunopeptidome purification, when searching the data against the same two DBs. Results from each DB were compared per replicate using db_compare.R, which outputs various comparison graphs and also three txt files listing the confident 7 to 15mers found in both result sets, only cryptic or only standard results, after removing any spectra attributed to iRT peptides in either search. "Cryptic-only" peptide lists from the three replicates were input into the origins program, which searches the transcriptome sequence files (translating each transcript in six frames) and also the standard DB for matches to each peptide sequence. The *origins* program outputs three files: a "discard list" of artificial junction peptides (caused only by use of the pseudoprotein WWWWW linker; see DB Building section); a table reporting when sequences may derive from standard proteins; and a detailed table reporting information on possible origins for each peptide based on the transcriptomes developed with workflow in Figure 1*B*, and also stating whether there was a conventional explanation. For transcripts associated with known

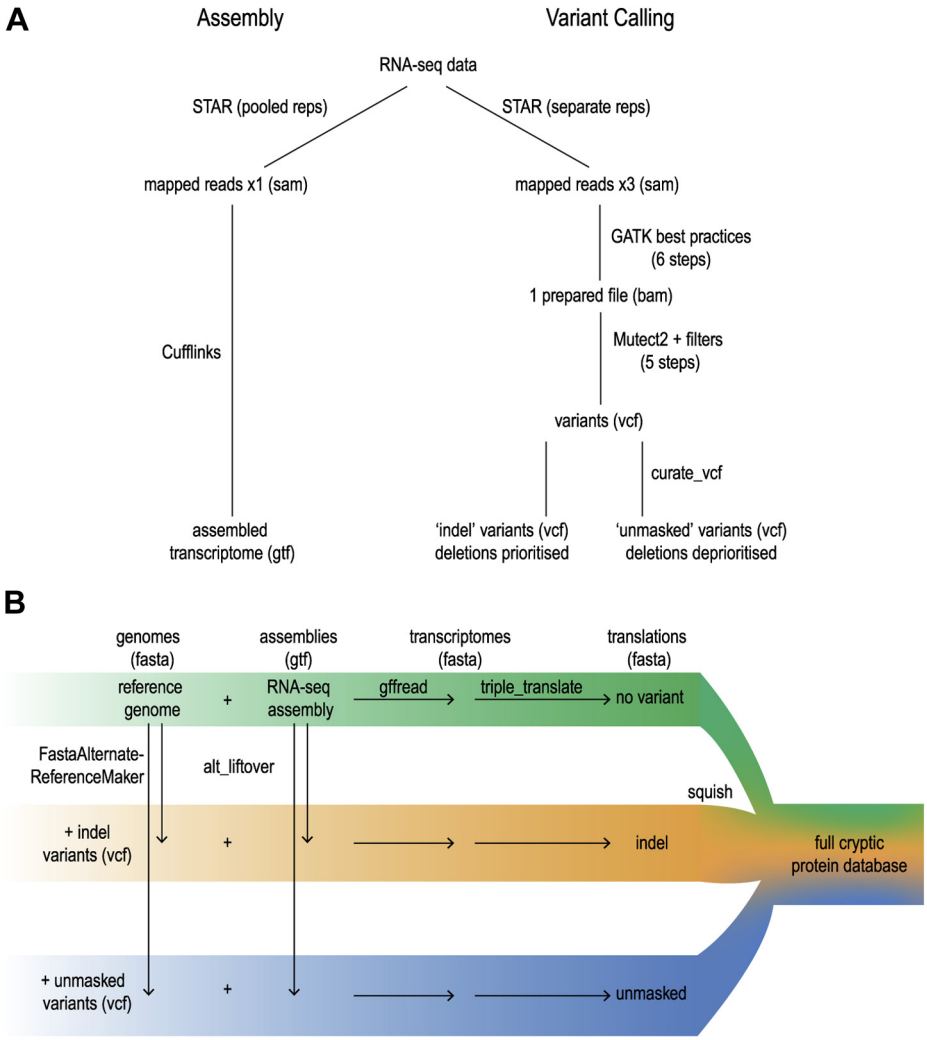


FIG. 1. **Schematic of the immunopeptidogenomics workflow.** *A*, RNA-Seq data from cells of interest were analyzed in two ways: replicates were pooled for *STAR* mapping before assembly into transcriptomes using *Cufflinks* and mapped separately, and the results later pooled as part of the GATK best practices workflow for preparing files for variant calling. Variant calling was performed by GATK's *Mutect2*, and the variant files were curated by our in-house program *curate_vcf*, which ensures that deletion mutations that affect downstream mutation sites are handled separately, thus increasing the possible mutations represented in the final database (DB). *B*, *gffread* wrote out the transcript sequences, based on the reference genome and the transcriptome coordinates from the assembly constructed in (*A*). Furthermore, the curated variant files were used in two ways to create parallel transcriptomes: first, GATK's *FastaAlternateReferenceMaker* incorporated the variants into the reference genome to form alternate genomes; second, our in-house program *alt_liftover* performed "liftover," altering the assembly files so that the listed coordinates corresponded to the revised genomes. Thus, *gffread* was able to write out the transcriptome sequences including variants. In-house software *triple_translate* then translated the transcripts in all three frames, regardless of biologically probable translation start sites and retaining all sequences of eight or more amino acids, to form protein DBs. Finally, in-house software *squish* combined the DBs and reduced redundancy by removing duplicates and any sequences that were wholly incorporated within another sequence. Furthermore, *squish* concatenated sequences of ≤300 amino acids, using a "WWWWW" linker sequence in between them, to form pseudoproteins of ~600 amino acids in the final and searchable "cryptic DB." GATK, Genome Analysis Toolkit.

annotations, *origins* downloads information directly from Ensembl (http://rest.ensembl.org/lookup/id/...) and provides metadata based on the genomic location of the nucleotide sequence encoding the peptide, such as the frame and region relative to canonical translation. *origins* assigns peptides to categories based on this information (as explained in Supplemental Methods) and also reports when sequences arise because of a variant. Finally, artificial junction peptides were removed from the analysis by rerunning db_compare.R with the discard list as additional input.

Validation of Cryptic Peptides by LC–MS/MS

Cryptic peptides were selected for validation using a two-step procedure. First, the binding affinity of the cryptic peptides to HLA A*02:01 was determined using NetMHCpan 4.0 (43), and peptides predicted to bind strongly (ranks of between 0.5 and 2.0) were shortlisted. Second, the shortlisted peptides' spectra were manually inspected to select only those whose MS/MS product ions matched the predicted theoretical b- and y-ion fragmentation series (obtained

from ProteinProspector MS-product [http://prospector.ucsf.edu]). Synthetic peptides for the selected sequences were ordered from Mimotopes as pepsets with ~80% purity. Pepsets were reconstituted in 0.1% formic acid (FA), pooled together, and spiked with a mixture of 11 iRT peptides to aid retention time alignment (42) before LC–MS/MS analysis using the same method as for the HLA peptide analysis (44). That is, peptides were loaded onto a PepMap Acclaim 100 C_{18} trap column 5 μm particle size, 100 μm × 2 cm and 100 Å (Thermo Fisher Scientific) at 15 μl/min using an Ultimate 3000 RSLC nano-HPLC (Thermo Fisher Scientific). After equilibrating the column with 2% acetonitrile and 0.1% FA, peptides were eluted and separated on an in-line analytical column (PepMap RSLC C_{18}, 2 μm particle size, 75 μm × 50 cm and 100 Å; Thermo Fisher Scientific) using a 125-min gradient at a flow rate of 250 μl/min. The gradient started from 2.5% buffer B (80% acetonitrile and 0.1% FA) in buffer A (0.1% FA) and increased to 7.5% buffer B over 1 min followed by a linear gradient to 37.5% buffer B over 90 min, then an increase to 99% buffer B over 10 min. Peptides were introduced using nanoelectrospray ionization method into the Orbitrap Fusion Tribrid MS (Thermo Fisher Scientific) at a source temperature of 275 °C.

All MS spectra (MS1) profiles were recorded from full ion scan mode 375 to 1800 *m/z*, in the Orbitrap at 120,000 resolution with automatic gain control target of 400,000 and dynamic exclusion of 15 s. The top 12 precursor ions were selected using top speed mode at a cycle time of 2 s. For MS/MS, a decision tree allowed distinct selection criteria for peptides of charge state +1 *versus* those with charge +2 to +6. For singly charged analytes, only ions falling within the range of *m/z* 800 to 1800 were selected, whereas for +2 to +6 ions, no such parameter was set. The c-trap was loaded with a target of 200,000 ions with an accumulation time of 120 ms and isolation width of 1.2 amu. Normalized collision energy was set to 32 (high-energy collisional dissociation), and fragments were analyzed in the Orbitrap at 30,000 resolution.

Spectra from the synthetic peptides and their HLA-derived counterparts (previously published data (27)) were compared in terms of normalized retention time, exploiting the iRT standard peptides spiked into each sample, and the similarity of ion fragmentation, determined by our program *msDot*. *msDot* performs peak matching between spectra, imputes peaks of "0" intensity to fill gaps if peaks fail to find a partner, and then converts the spectra to vectors of the peak intensities. The normalized dot product of vectors a and b is calculated as:

$$\text{Normalised dot product} = \mathbf{a} \cdot \mathbf{b} \,/\, (|\mathbf{a}||\mathbf{b}|)$$

where $|\mathbf{a}| = \sqrt{(\mathbf{a} \cdot \mathbf{a})}$ and $|\mathbf{b}| = \sqrt{(\mathbf{b} \cdot \mathbf{b})}$.

Thus, *msDot* outputs a number between 0 and 1 for each comparison, where 1 indicates perfect identity and 0 total dissimilarity.

RESULTS

To help discover sequences arising from unconventional translation and/or from genomic variants in immunopeptidomics studies, we have developed an immunopeptidogenomics workflow, which produces sample-specific and nonredundant transcriptome-based protein DBs. We provide further code to help compare the resulting identifications with standard results and investigate the possible biological origins of identified unconventional sequences. We have tested this workflow for discovering unconventional HLA-binding peptides in the AML cell line THP-1.

As described previously and shown in Figure 1, we employed THP-1 RNA-Seq data for both transcriptome assembly and variant calling and converted the results into a three-frame–translated transcriptome-based protein DB including sequences caused by genomic variants. While our workflow allows for filtering the transcriptome based on RNA-Seq expression evidence for each transcript, in this study, our RNA-Seq data were only of medium depth, and the peptides we identified were ultimately linked to transcripts with a wide range of observed expression, including many transcripts with no RNA evidence (supplemental Fig. S1). Therefore, we chose to use the unfiltered transcriptome for this analysis (the Supplemental Results and supplemental Figs. S2 and S3 explore the results observed using a filtered transcriptome). That is, the THP-1 transcriptome contained all reference transcripts as well as novel transcripts deduced from the RNA-Seq data. The resulting "cryptic" DB was a 247 MB FASTA file including 342,739 sequences and incorporating >300,000 mutations, largely consisting of pseudoproteins formed by concatenating short sequences *via* a penta-tryptophan "WWWWW" linker. Concatenating the many short sequences allowed PEAKS to thoroughly consider them despite its DB search sequence limitations, while 'the linker sequence minimized creation of biologically likely artificial "junction" peptides (for identifying immunopeptides purified from HLA allotypes known to prefer Trp as a C-terminal anchor residue, a linker composed of a different amino acid, *e.g.*, "RRRRR," could be substituted). We also downloaded a "standard" UniProt human proteome DB, which was 48 MB with 96,832 sequences. To directly compare the DBs another way, we counted the unique 8-mers within all the proteins for both DBs. The cryptic DB contained 118×10^6 unique 8mers compared with only 11.5×10^6 unique 8mers in the standard DB. To test the usefulness of the larger cryptic DB for identifying HLA-binding peptides, including unconventional sequences, from mass spectrometric data, we used PEAKS to search previously published triplicate HLA-A*02:01 immunopeptidome data from THP-1 cells (27) against both the cryptic and standard DBs in parallel. In both searches, we employed a global 5% FDR score threshold for peptide identification. As is not uncommon in immunopeptidomics (45, 46), we generally choose a more permissive threshold than for proteomics, to avoid the high incidence of false negatives seen at more stringent FDRs (45). This avoids "throwing away" biologically important peptides at an early stage, while emphasizing the importance of validating sequences of interest through further experimentation. About 11 peptides identified using the cryptic DB were purely artificial sequences formed *in silico* at the junctions between peptide and linker in the pseudoproteins and were removed, compared with a total of 14,709 biological sequences found in the cryptic DB searches (Fig. 2*B*; the 11 artificial peptides are listed in supplemental Table S10).

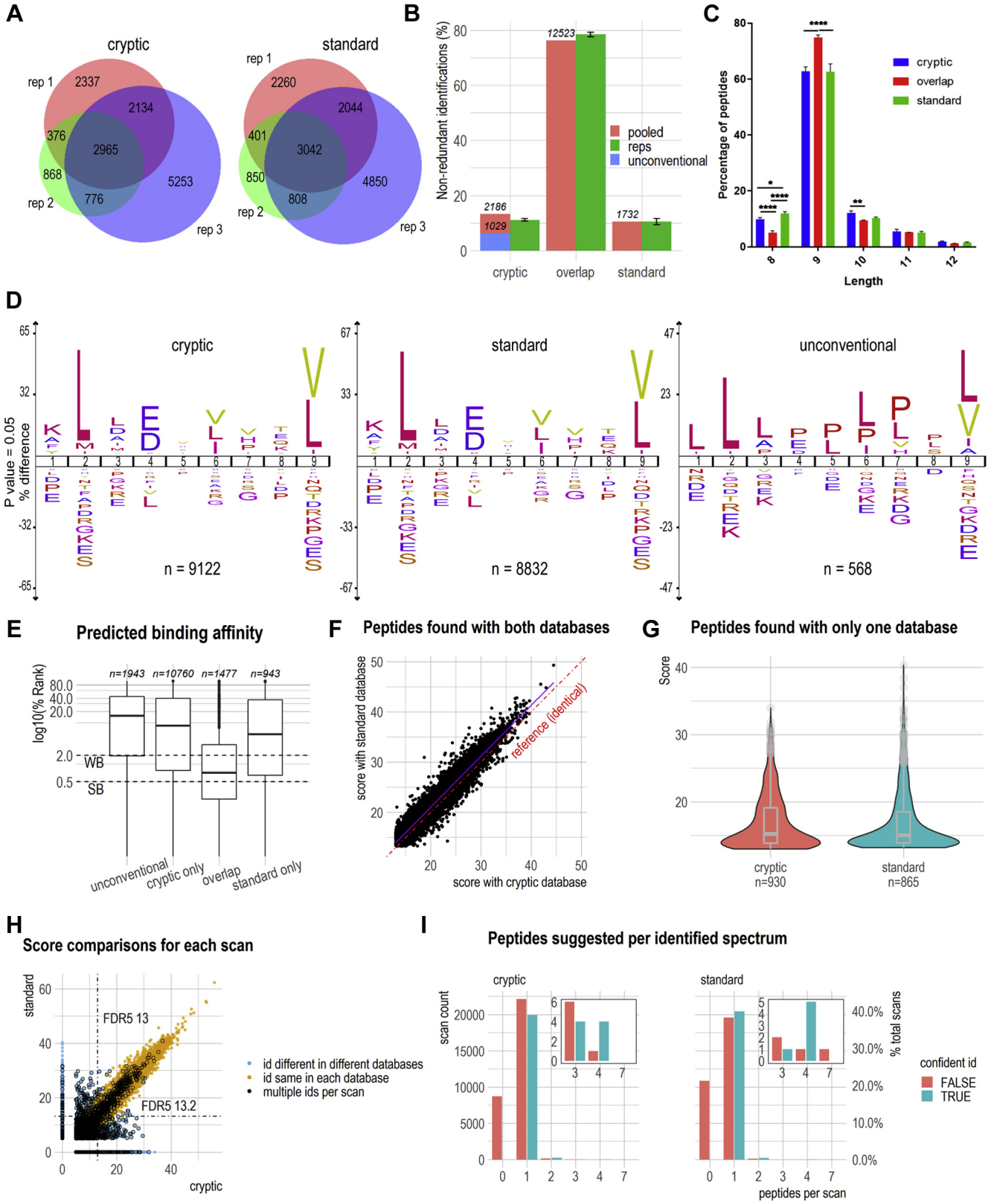


FIG. 2. **The immunopeptidogenomic ("cryptic") and UniProt-Trembl ("standard") databases (DBs) yielded very similar results in terms of peptide numbers and HLA–peptide characteristics, with minimal ambiguity in results from the parallel searches.** HLA-A*02:01 molecules were immunopurified from three replicates of THP-1 cells using BB7.2 Ab. Peptides were separated from the denatured molecules using

Searching Against the Cryptic DB Largely Recapitulated the Standard Results and Identified a Number of Novel and Unconventional Peptides

We scrutinized the two result datasets to assess the effect of using the cryptic DB instead of the standard DB (Fig. 2). First, we compared the overlap in confidently identified 7 to 15mers (5% FDR) between different replicates searched against the same DB (peptide identification data can be found in supplemental Tables S1–S6). The semiquantitative Venn diagrams in Figure 2*A* reveal nearly identical patterns of overlap between replicates using each DB. The triplicates yielded different numbers of confident sequences, reflecting technical and/or biological variation between replicates, but still displayed reasonable overlap. The proportion of unique results in each replicate increased with the size of the dataset (from ~17% in the smallest dataset up to nearly half of the largest set), indicating a shared core of highly reproducible peptides. Next, we analyzed the effect of searching against different DBs for each replicate. As Figure 2*B* shows, nearly 80% of the confidently identified peptides were found using both DBs (labeled "overlap" throughout (Fig. 2)). Searching against the cryptic DB recapitulated 88.2% of standard results (SE = 0.01%; n = 3). Of those identified only using the cryptic DB, 42.6% were unconventional sequences (*i.e.*, absent from the standard DB; SE = 2.1%), totaling 1029 unconventional sequences from the combined replicates.

We compared the peptides found using each DB by assessing characteristics relevant to HLA binding: peptide length (Fig. 2*C*), motif (Fig. 2*D*), and predicted binding affinity (Fig. 2*E*). Figure 2*C* shows that the confident 7 to 15mers follow the expected length pattern for HLA-A*02:01 binding with 9mers dominating the datasets; however, the peptides found using only one DB (cryptic or standard) included significantly smaller proportions of 9mers, whereas 8mers were significantly increased in those sets. Similarly, when we used NetMHC to predict the binding affinity of applicable sequences (8–14mers lacking PTMs), we found strong binding potential in the peptides observed with both DBs, yet weaker results among peptides found only using a single DB. Here, we also assessed the set of unconventional sequences, which are a subset of those found only using the cryptic DB and also had overall weaker predicted binding. Taken together, these data imply a higher likelihood of false discoveries among peptides found only with one DB. Nonetheless, all sets contained peptides predicted to bind very strongly to HLA-A*02:01 (Fig. 2*E*). We used iceLogo to investigate the motif of 9mers lacking PTMs found using the cryptic or standard DB and for the unconventional subset (Fig. 2*D*). Both DBs yielded peptides that adhered to the expected consensus motif for binding HLA-A*02:01.

To directly compare search performance using the two DBs, we investigated each replicate in turn (Fig. 2, *F–I* show the results for replicate 1; similar results for the remaining replicates are displayed in supplemental Fig. S4). First, we compared the maximum scores (−10logP) achieved for each peptide confidently identified using each DB. Peptides found using both DBs could be compared directly (Fig. 2*F*), showing close correlation in scoring between DBs, with only a small decrease in score when using the cryptic DB despite the much larger search space. For peptides found using only one DB, we compared the scoring distributions with violin and overlaid box plots, which demonstrated no appreciable difference between DBs (Fig. 2*G*). Further investigation determined that over 75% of the spectra of peptides found only using one DB were not assigned any peptide sequence in the other DB

HPLC and subjected to LC–MS/MS. The data were searched using PEAKS X against the immunopeptidogenomics DB ("cryptic") and a UniProt-Trembl DB ("standard") in parallel, with FDR estimation enabled; 5% FDR was used as a threshold for confident identifications. *A*, Venn diagrams show the overlap in confidently identified 7 to 15mers between replicates when searching with each DB. *B*, the percentages of 7 to 15mers identified confidently using both DBs (overlap), or only when using one particular DB, were calculated for each replicate separately (reps) and after pooling the replicates from each search (pooled). Peptide numbers are displayed above the "pooled" bars. Unconventional peptides among the pooled "cryptic-only" set are highlighted by an overlaid bar, labeled with the number of unconventional peptides. The proportions of 8 to 12mers among the peptides in these sets were also examined, resulting in *C*. The *star* ratings indicate *p* values for the difference between sets at certain lengths found using two-way ANOVA with Tukey's multiple comparisons, where four, two, and one *stars* stand for $p < 0.0001$, $p = 0.043$, and $p = 0.0327$, respectively. Both *B* and *C* plot the means of the percentages for the three replicates with standard error. *D*, uses IceLogo to reveal the motif of the 9mers confidently identified with each DB (pooled replicates, PTMs removed) and also the motif of the "unconventional" 9mers (*i.e.*, sequences absent from the standard DB). *E*, the binding affinity predicted by NetMHC4.0 for the 8 to 14mers (pooled replicates, PTMs removed) confidently identified with both DBs (overlap), only the cryptic/standard DB, and the "unconventional" set in terms of their percentage rank. Peptide numbers per set are indicated above the boxplot. <0.5% and 0.5 to 2% rank indicate strong and weak binding, respectively. Plots *F–I* were produced for each replicate; data for replicate 1 are shown here with the remainder shown in supplemental Fig. S1. *F* and *G*, the maximum PEAKS score for each peptide confidently identified using each DB was determined. *F*, directly compares scores for peptides confidently found with both DBs while the violin and boxplots in *G* compares the scores of peptides that were only found using one DB. *H* and *I*, assess the ambiguity in identifications from the same scan both within and between searches with the two DBs. *H*, plot scores for each scan identified with each DB and shows the FDR5 confidence threshold for each search as a *dashed line*. To compare scores from the two searches, ids for each scan were paired up first by matching any identical ids (*yellow dots*), then by pairing up different ids or substituting a 0 score if the one search lacked a partner id (*blue dots*). Thus, scans may be represented by >1 dot if there were multiple ids. *Black-ringed dots* indicate scans that received multiple possible identifications either within one search or between the searches. *I*, focuses on the number of identifications per scan within the same PEAKS search, and whether these ids were confident (multiple ids within searches always score the same). *Insets* zoom in on the scans with >2 ids. FDR, false discovery rate; HLA, human leukocyte antigen; PTM, post-translational modification.

search (supplemental Fig. S5). Finally, we wished to investigate the level of ambiguity in the results—that is, the extent to which individual spectra received multiple confident identifications. The data presented in Figure 2, *H* and *I* were not filtered by peptide length or FDR, include iRT peptides, and were compared by scan number (*i.e.*, spectrum) rather than by sequence. Figure 2*H* plot scans by the score of the PSM observed when searching against each DB, imputing a score of 0 where one DB lacked a match (PEAKS' minimum score is 5, resulting in the gap adjacent to each axis). Color coding indicates same/different identifications for the same scan in different searches, and black rings indicate multiple potential identifications either in different searches or in the same search. The latter results from the PEAKS DB search assigning multiple confident identifications to one scan, so some scans are represented by more than one dot in Figure 2*H*. These ambiguous results have matching scores and often involve indistinguishable I/L variations; Figure 2*I* plots the frequency and degree of this occurrence and shows no significant change in ambiguous identifications when searching against the different DBs. Importantly, Figure 2*H* shows that most scans with different identifications when using different DBs (blue dots) are below the confidence threshold for one search. Thus, as expected, the parallel searches contain some ambiguity, but the confidence threshold eliminates most of it. As further investigation using alternative methods would be necessary to determine which of the remaining ambiguous identifications are correct, and such identifications may include biologically important or useful results, we chose to retain all confident identifications in our further analysis.

Since PEAKS performs *de novo* identification prior to the PEAKS DB search, we were interested to assess whether PEAKS' "*de novo*–only" identifications from the standard DB search contained our unconventional peptides. From the triplicates, we found that the *de novo*–only results included an average 20.1% (SE = 1.0%) of the unconventional sequences found using the cryptic DB.

Unconventional Peptides Were Commonly Attributed to Translation From Noncoding RNAs, Noncanonical Frames, or UTRs

We sought to understand the origin of peptides that had no conventional explanation (*i.e.*, those not present in UniProt). Our program *origins* searched for the sequences within our THP-1 transcriptomes. Cufflinks assigned each assembled transcript a class code and also linked them to reference annotations where possible, allowing *origins* to source information on the reference transcripts directly from Ensembl and calculate the location of peptide-coding sequences in relation to the reference transcript and any canonical translated regions within it. *origins* reports this information both with a detailed statement entitled "metadata" and by assigning each potential origin to a "category," with terminology based on class code or region, with the coding frame in relation to canonical translation of the reference transcript where applicable, as seen in Figure 3*A*. (See Supplemental Methods section for a detailed explanation about the categories). It was immediately clear that many peptides could be readily assigned to multiple transcripts, because of both the relatively short sequences (7–15 mers) of HLA class I peptides, and the fact that many genes were represented by multiple possible transcript isoforms (no doubt exacerbated by the difficulty of assembling isoforms accurately from short-read RNA-Seq data). Therefore, peptides could also be assigned to multiple categories. Currently, there is no basis for classifying each peptide into any one category over others; so to produce Figure 3*A*, we counted each peptide toward each assigned category. Noncoding transcripts formed the single largest category of unconventional sequences, followed by canonical

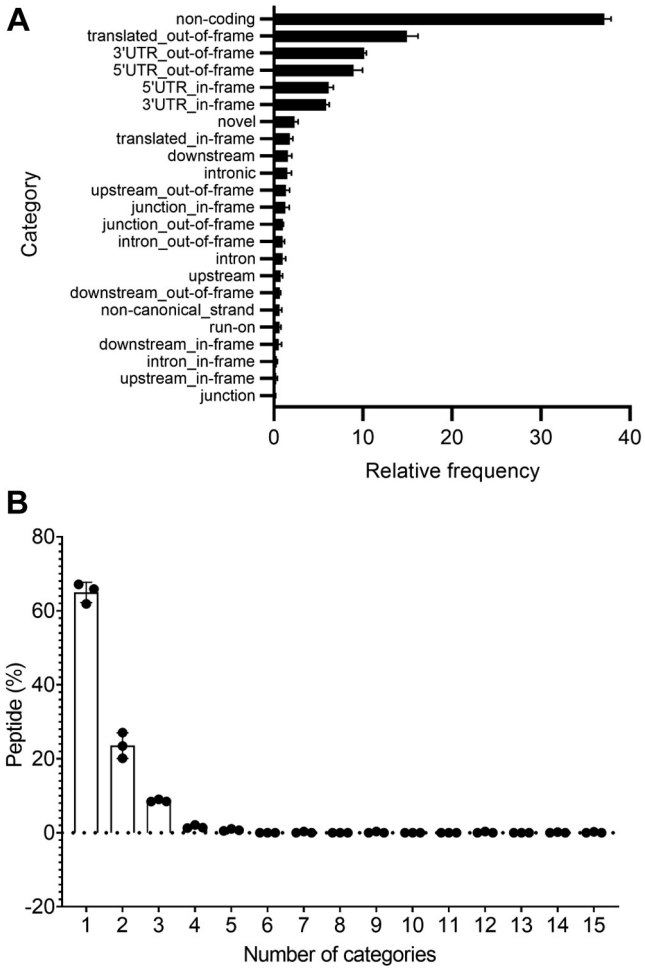


FIG. 3. **The peptides with no conventional explanation mostly may have been generated by translation from noncoding RNA, a noncanonical frame, or from UTRs.** Peptides were split across 23 categories determined by the *origins* program, as defined in the Supplemental Methods section. *A*, each peptide was allowed to count toward multiple categories. *B*, plots how many categories individual peptides counted toward. The error bars represent ±SEM, n = 3.

coding regions translated in the "wrong" frame. However, the various categories representing the 5′- and 3′-UTRs also made up a large proportion when taken together. Other peptides were assigned to sequences in introns, translation that spanned different regions ("junction" category) or translation from the "wrong" strand, upstream or downstream of known transcripts, or novel transcripts not associated with known genes. To check our categorization strategy, we sought to identify the number of peptides that have multiple origin categories. To our surprise, almost two-thirds of the peptides were assigned to a single category, and the vast majority of peptides (>95%) were assigned to ≤3 categories (Fig. 3*B*), which supports the validity of Figure 3*A* as a guide to the probable origins of unconventional sequences in this immunopeptidome.

Most THP-1 Variants Generating Unconventional HLA-Binding Sequences Were Known SNPs, Not Cancer Mutations, but Some Unconventional Sequences Were Associated With Known Cancer Genes

To facilitate neoepitope discovery, the *origins* program also reports when peptide-coding sequences contain variants. We investigated these peptides to determine how many reported variants were nonsynonymous in the frame of translation (ns-SNVs), thus producing neopeptides. The 16 relevant peptides are listed in Table 1, along with a summary of their possible transcript origins. The 14 ns-SNVs involved were spread across 11 different chromosomes, and the possible origins for ten of the sequences included "normal translation" (*i.e.*, translation from a canonical coding region in the "correct" frame). The other sequences were unconventional because of both the presence of a variant and noncanonical translation. Since Mutect2 was used in tumor-only mode with no matched "normal" control, it could not distinguish between germline variants and somatic mutations. Therefore, we checked whether the observed variants were previously known; all but one was present in the dbSNP (Table 1). The exception belonged to a 7mer peptide, which does not fit the HLA-A*02:01 binding motif, and therefore, it is likely to represent a false discovery.

We were interested in the hypothesis that cryptic epitopes can extend the coverage of antigens presented by HLA, and 1000 of the 1029 unconventional sequences were associated with known genes. Therefore, we further investigated these genes to determine how many were linked with UniprotKB proteins, and whether these genes/proteins constituted known cancer-associated antigens (supplemental Table S7). About 727 of the unconventional sequences were associated with 893 distinct UniProtKB accessions, which were linked to 866 Ensembl gene ids. We crossreferenced the gene lists with known cancer-associated gene lists sourced from the TANTIGEN 2.0 Tumor T-cell Antigen DB (47) and the COSMIC Gene Census (48). About 17 unconventional sequences were associated with 16 TANTIGEN genes, 53 with 52 COSMIC genes; five of these genes were present in both cancer gene lists, corresponding to six unconventional sequences.

A Number of Unconventional Peptides Were Validated by Comparison With Synthetic Counterparts

To validate the identification of the unconventional peptides, we synthesized 53 of them, which were predicted to bind to HLA-A*02:01 and analyzed them under mass spectrometric conditions similar to the discovery experiment. The resulting spectra were compared with the original experimental results, in terms of the similarity of fragmentation and retention time, as exemplified in Figure 4*A* (remaining mirror plots and associated data are found in supplemental Fig. S6 and supplemental Table S8, *A*–*C*). We calculated the dot product for the mirrored spectra to judge the similarity in fragmentation and utilized the iRT standard peptides spiked into each sample to calculate the normalized retention times of each peptide in iRT units (42). About 41 of 53 sequences were confirmed by both fragmentation patterns with dot product >0.80 and retention time differences of ±4 iRT, as shown in Figure 4*B*. The remaining 12 sequences matched according to either fragmentation or retention time but not both. The dot product test may have been overly conservative in some cases under the experimental conditions and given the chemistry of fragmentation for the particular peptide, since a simple dot product comparison cannot take fragmentation quality into account. No sequences failed both validation criteria.

For the 41 validated peptides, we investigated even further how they may be generated biologically. We searched for these sequences in our THP-1 cryptic protein DB to extract the three residues flanking the sequence on either side in each potential pseudoprotein. This brought to light that 14 validated peptides could have been generated by a coding sequence followed immediately by a stop codon. Five peptides' potential coding sequences also encoded a possible start codon (M) within three codons of the peptide's code. To double check these very small ORFs, we input the relevant transcript sequences into NCBI's ORF finder, which confirmed that the peptides may derive from ORFs encoding as few as 11 amino acids.

DISCUSSION

In immunopeptidomics, peptide sequences may arise from unconventional translation; yet these will be missed or misidentified when using conventional approaches to search mass spectrometric data due to their absence from the standard protein DBs. To facilitate the discovery of such peptides, we have developed an immunopeptidogenomic workflow, which produces protein DBs from RNA-Seq data. The resulting 'cryptic' protein DB is sample-specific and includes both conventional and unconventional sequences, since whole transcripts are translated in all three frames and

TABLE 1
Peptides with sequences only generated because of nonsynonymous variants in the cryptic DB

Peptide	No. of possible transcripts[a]	Metadata summary[b]	Chromosome	Variant	dbSNP accession number
AAPVFRR	1	Intronic (+1)	19	CCTTT->C@33797693	NF
ALSSVDPEV	8	Normal translation: 6 3′-UTR (+2): 1 Noncoding exon: 1	1	T->C@155765221	rs2297775
EPVAVAQPQ	8	Normal translation: 5 Noncoding exon: 2 Retired transcript (no longer in Ensembl): 1	12	A->C@881746	rs956868
ILPEPSHKV	12	3′-UTR: 2 3′-UTR (+1): 6 3′-UTR (+2): 2 Noncoding exon: 2	12	T->G@1793600	rs2058111
LDTRNNVKV	2	Normal translation	9	T->C@104800523	rs2230808
LICQPHSDPA	5	Noncoding exon: 4 Downstream: 1	6	G->C@31202451	rs9366770
LLQEELEKL	8	Normal translation: 3 (+1): 3 (+2): 1 Noncoding exon: 1	20	C->A@17615510	rs1132274
LNLIFSVPS	1	Noncoding exon	12	C->G@7923034	rs41438344
LTGATALRL	1	5′-UTR	14	G->T@20955528	rs945351
PEPSHKV	12	3′-UTR: 2 3′-UTR (+1): 6 3′-UTR (+2): 2 Noncoding exon: 2	12	T->G@1793600	rs2058111
SCGIFRKSVS	5	Normal translation	2	G->C@233841134	rs3821238
SHYEVKL	2	Normal translation	15	G->A@68312831	rs4777035
SLSHYEVKL	2	Normal translation	15	G->A@68312831	rs4777035
TLDRVLPRV	4	Normal translation: 3 (+1): 1	4	C->A@168394609	rs3749499
TLHDQIFQA	5	Normal translation: 2 Noncoding exon: 3	22	G->A@45332489	rs6007594
VIQERVHSL	4	Normal translation: 2 3′-UTR: 1 Noncoding: 1	12	G->A@68326847	rs962976

[a]Number of THP1 assembly transcripts giving rise to peptide sequence.
[b]Information related to location of peptide-encoding sequence with respect to known Ensembl transcripts associated with THP1 assembly transcript, where possible.

without regard to apparent start sites. The workflow allows users to filter the transcriptome based on RNA-Seq expression evidence (FPKM) to a user-specified threshold. The DB can also incorporate variants (both germ-line and somatic), potentially derived from the same RNA-Seq data. Thus, our workflow may produce a DB that better represents the individual sample/patient for immunopeptidomics discovery experiments, allows identification of cryptic peptides and minor histocompatibility antigens, and highlights potential neoepitopes.

We have demonstrated the usefulness of searching such a DB for discovering peptides of the dark immunopeptidome using the AML cell line THP-1. After generating a cryptic DB from THP-1 RNA-Seq data, we searched THP-1 immunopeptidome mass spectra against both the cryptic DB and a standard UniprotKB proteome. Our in-depth analysis showed that searches against the cryptic DB largely recapitulated standard results, with an intriguing set of 1029 additional peptides which could not be found in a conventional DB. Our *origins* program helped us investigate these peptides by categorising the transcript types and locations of coding sequences (supplemental Table S9). We found few peptides resulting from nonsynonymous variants, and most of these were known SNPs and thus unlikely to represent neoepitopes (Table 1). However, many of the unconventional peptides could be generated by translation from noncoding RNAs, or from cryptic translation of a coding transcript (*e.g.*, translating from a noncanonical frame or from a UTR).

To demonstrate that our workflow can help in identification of tumour specific unconventional peptides we investigated if the cryptic peptides found in our study could be mapped to known tumorigenic proteins. Since AML has low tumour

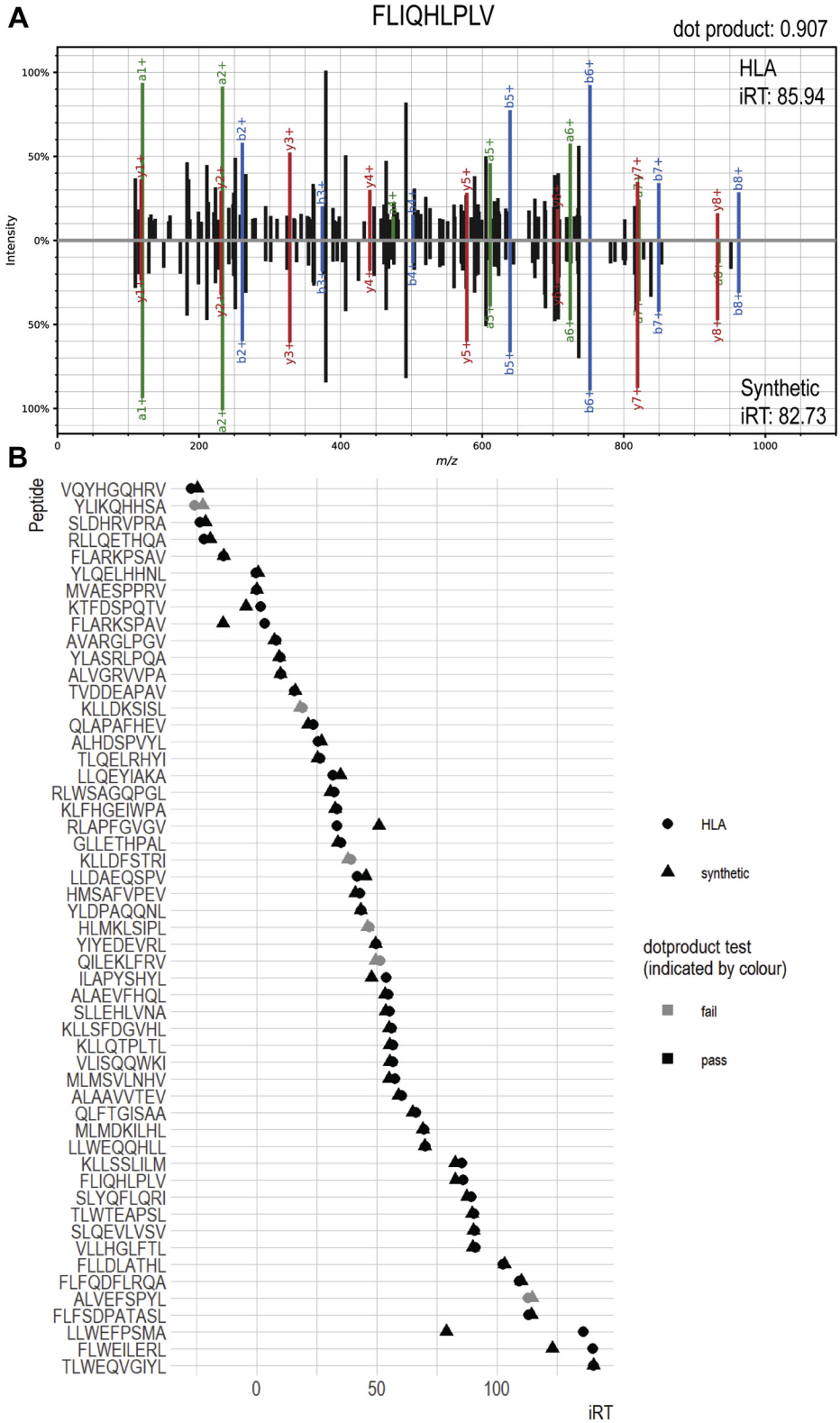


FIG. 4. **A selection of unconventional peptide identifications were validated by comparison to synthetic counterparts.** Synthetic peptides were subjected to mass spectrometry under similar conditions to the original experiments and compared with experimentally identified peptide-spectrum matches (PSMs) in terms of fragmentation pattern and retention time (RT). Similarity of spectra was measured using an in-house program, which matched peaks between spectra, converted them to intensity vectors (imputing intensity 0 where it found no

mutation burden, identifying cryptic peptides from tumorigenic proteins of interest may be a crucial avenue for developing immunotherapy and personalised medicine. For this purpose, we cross-referenced the source proteins of the identified cryptic peptides with cancer-associated genes reported in the COSMIC Gene Census DBs and TANTIGEN 2.0. Interestingly, 64 unconventional peptides were linked to known cancer-associated genes present in the two DBs (supplemental Table S7). This included peptides originating from genes associated with AML including Runt-related transcription factor 1 (RUNX1), Friend leukemia integration 1 transcription factor (FLI1) and Serine/arginine-rich splicing factor 3 (SRSF3); RUNX1 and FLI1 are part of the FLT signalling pathway which is perturbed in AML (49). Some other peptides of interest came from GTPase NRas (NRAS) protein, Nuclear receptor coactivator 2 (NCoA-2), Forkhead box protein P1 (FoxP1) and Transcriptional regulator ATRX. We also compared our data with recently published datasets reporting a combined total of 293 cryptic peptides restricted to different HLA class I alleles (50, 51). One of the unconventional peptides we identified here, LLSSKLLLM, was also reportedly identified as a HLA-A*02:01-restricted peptide in a patient sample from high grade serous ovarian cancer, using a similar proteogenomic method. In agreement with Zhao *et al.* (50), we found that the peptide originated from a noncoding intergenic region present on chromosome 5. Finding this unconventional peptide in separate AML and ovarian cancer studies supports the hypothesis that proteogenomic strategies such as we present here facilitate the discovery of shared unconventional epitopes which are not only valid, but may lead to immunotherapies capable of treating a broad range of cancers and patients.

Another interesting facet of our study was the identification of peptides originating from sORFs. sORFs give rise to microproteins which have been increasingly identified to play a role in several key biological processes including DNA and RNA repair and regulation (52). Also, sORFs may give rise to defective ribosomal products (DRiPs) and/or short-lived proteins (SLiPs) (53), which are known to be funnelled through the antigen processing and presentation pathway and ultimately presented by HLA molecules. We found that of the 41 validated peptides, 14 peptides could have been generated from sORFs. These included FLIQHLPLV, shown in Figure 4, which was also associated with a gene found in the COSMIC Gene Census, TBL1XR1.

Without a tumour-matched normal sample, we have not proven that these peptides are cancer-associated. Nonetheless, these preliminary findings highlight the potential of illuminating the dark immunopeptidome for discovering cancer-associated epitopes, in agreement with previous studies. For example, Laumont *et al.* (14) found a far greater number of cancer-associated presented peptides caused by 'aberrant expression' of supposedly noncoding regions than derived from somatic mutations, and Chong *et al.* (15) noted that cancer-specific cryptic epitopes are more often shared between patients than neoepitopes derived from mutations. Such findings tally with the growing body of evidence that aberrant translation is associated with or even drives cancer, including the use of unconventional 5′ initiation sites (54), intron retention (55–57), alternative splicing (58), and translation of supposedly noncoding RNAs (59). The dark immunopeptidome broadens the HLA peptide repertoire, offering greater hope to clinicians for effective immunotherapies in the future (6).

A deeper understanding of the dark immunopeptidome will be valuable not just for cancer research, but for our fundamental understanding of immune surveillance in health and disease. Cryptic epitopes were originally discovered in disease contexts including cancer (60, 61) and autoimmune disease (62), so we have long known of their immunological relevance. However, without a technique for routine discovery, many questions remain unanswered, such as: What is the proportion of cryptic peptides in the immunopeptidome? Does this vary depending on tissue type, under stress conditions or in disease states other than cancer? Do cryptic epitopes play roles in tolerance breakdown in specific autoimmune diseases, adverse drug reactions or allergies? How many peptides/microproteins with minimal antigen processing requirements result from small ORFs <100 amino acids, such as we found, and how does this change in or affect disease states?

The novel programs and workflow presented here represent a step toward such routine analysis of the dark immunopeptidome, and our THP-1 analysis helped highlight where we might improve our workflow. For example, we utilized a standard PEAKS Studio FDR calculation to determine confident identifications, because FDR is straightforward, accessible, and well understood, but some researchers are moving away from its use in immunopeptidomics because of theoretical concerns (14–16). Also, Nesvizhskii (9) stated that in proteogenomics, novel peptide identifications require stronger evidence than known peptides. It is unclear how to apply this to the immunopeptidome without undue bias when we know so little about its true composition. Various groups are turning to methods of group-specific FDR analysis (15, 16) (*e.g.*, calculating the FDR for the unconventional and annotated peptides separately); however, this calculation similarly demands the separation of protein decoys into unconventional and annotated categories. In its current form, our workflow

matching peak), and calculated the dot product where 1 indicates perfect identity and 0 total dissimilarity. RTs were compared by calculating normalized RT for each spectrum with reference to spiked-in iRT standard peptides. Both metrics are noted in (*A*) a representative mirror plot, where the spectra for the experimental and synthetic peptides are on the *top* and *bottom*, respectively; and (*B*), which summarizes the results from all the peptides tested in this manner.

translates transcripts in three frames, then links the resulting sequences into pseudoproteins to generate a searchable DB, without regard to whether the sequences are canonical or unconventional (we categorize them post-search using *origins*). PEAKS' decoy fusion method generates decoys by scrambling target sequences. Therefore, to classify decoys into "unconventional" or "annotated" categories to allow group-specific recalculation of the FDR, we would first need to separate the translated sequences into group-specific pseudoproteins with appropriate labels. This may be a promising avenue for further workflow refinement. Regardless, our analysis showed that the cryptic DB was still searchable despite its size, and we conclusively validated 41 of 53 unconventional sequences chosen for further investigation.

The workflow as presented here only incorporates SNVs and noncomplex indels, since it uses the Mutect2 variant caller. This means our cryptic DBs lack the fusion proteins formed by large genomic rearrangements, which help drive various cancers (63). For this study, we simply translated known AML fusion transcripts in six frames and appended the proteins to UniProt for searching, which did not yield any THP-1 fusion HLA-binding peptides (data not shown). In future, it will be relatively straightforward to find fusion transcripts in RNA-Seq data using published software, then insert these sequences into our workflow. We will then amend *origins* to find and report any fusion peptides.

The workflow as it stands provides users with considerable flexibility to adapt methods to suit the experimental aims. For example, the workflow was intended to be search-engine agnostic, producing a DB for use with the users' DB-search engine of choice. In practice, however, we recommend PEAKS, as it proved better able to cope with the large DB than other software we trialed (MaxQuant and MSFragger), presumably because of its method of prioritizing certain sections of the DB based on a first round of analysis (37). A recent article has also highlighted the suitability of PEAKS for immunopeptidomics (64). Increased computing power and/or the use of filtered cryptic DBs may facilitate the use of other search engines, if necessary. Similarly, the decision of whether to filter the transcriptome based on RNA-Seq expression evidence may depend on the quality and depth of RNA-Seq data available. Alternatively, users might choose to search against a range of differently filtered cryptic DBs and select consensus identifications for further scrutiny; our results indicate that this may prove a quick and easy way of selecting high confidence identifications for studies, although it may come at a high cost in terms of sensitivity. We recommend searching against both cryptic and standard DBs and then exploiting the software *db_compare.R* to help investigate the results, as a simple quality control indicator. However, given that the cryptic search recapitulated a large majority of the standard results here, parallel searches may prove unnecessary for routine use. We plan to make the software more accessible to nonbioinformaticians, while striking a balance between ease of use and flexibility. Therefore, we intend to develop a simple automated pipeline, as well as maintaining access to our novel programs for users who wish to customize their workflows (*e.g.*, to integrate the latest RNA-Seq tools).

In conclusion, we hope that our immunopeptidogenomics workflow will help change the illumination of the dark immunopeptidome from a novelty requiring special effort and expense to a typical part of any thorough immunopeptidomics study. Only when such analyses become mainstream will we begin to understand the true diversity of the immunopeptidome, as it is presented to cells of the immune system in health and disease.

DATA AVAILABILITY

RNA-Seq data are now available in the Sequence Read Archive, reference PRJNA686824; THP-1 HLA-A*02:01 immunopeptidome mass spectrometry data were previously published and are available as the PXD015039 dataset (27) in the ProteomeXchange Consortium *via* the PRIDE (65) partner repository. As reanalyses of raw data cannot be accommodated in PRIDE, data from the present study were exported from PEAKS at 5% FDR in mgf and mzIdentML formats for submission to MassIVE, along with the sequence DBs used to generate them (https://massive.ucsd.edu: dataset MSV000086922 has the UniProt search results; cryptic search results are in the reanalysis dataset RMSV000000338.1; https://doi.org/10.25345/C5VN5Q). Novel software and usage instructions are available from github.com/kescull/immunopeptidogenomics, and the repository has been archived in Zenodo with https://doi.org/10.5281/zenodo.5348860.

Supplemental data—This article contains supplemental data.

Acknowledgments—The authors acknowledge the Monash Proteomics and Metabolomics Facility for the provision of mass spectrometry instrumentation, training, and technical support. This project was funded in part by the Australian National Health and Medical Research Council project grant 1165490. Computational work was supported by the MASSIVE HPC facility (www.massive.org.au) and R@CMon/Monash Node of the NeCTAR Research Cloud, an initiative of the Australian Government's Super Science Scheme and the Education Investment Fund. The authors acknowledge use of the services and facilities of Micromon Genomics at Monash University and the staff at the Monash Biomedical Proteomics Facility for technical assistance. We thank Dr Leigh Humphries for sharing expertise in mathematics and logic.

Funding and additional information—A. W. P. is supported by the National Health and Medical Research Council Principal Research Fellowship 1137739.

Author contributions—K. E. S., K. P., and S. H. R. conceptualization; K. E. S. and S. H. R. methodology; K. E. S. software; K. P. investigation; K. E. S. writing–original draft; K. P., S. H. R., and A. W. P. writing–review and editing; K. E. S., K. P., and S. H. R. visualization; A. W. P. supervision; A. W. P. funding acquisition.

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: AML, acute myeloid leukemia; DB, database; FA, formic acid; FDR, false discovery rate; GATK, Genome Analysis Toolkit; HLA, human leukocyte antigen; iRT, indexed retention time; MHC, major histocompatibility complex; ns-SNV, nonsynonymous single nucleotide variant; PSM, peptide-spectrum match; PTM, post-translationally modified.

Received March 1, 2021, and in revised from, August 10, 2021
Published, MCPRO Papers in Press, September 10, 2021, https://doi.org/10.1016/j.mcpro.2021.100143

REFERENCES

1. Schumacher, T. N., and Schreiber, R. D. (2015) Neoantigens in cancer immunotherapy. *Science* **348**, 69–74
2. Bassani-Sternberg, M., and Coukos, G. (2016) Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr. Opin. Immunol.* **41**, 9–17
3. Schmidt, M., and Lill, J. R. (2019) MHC class I presented antigens from malignancies: A perspective on analytical characterization & immunogenicity. *J. Proteomics* **191**, 48–57
4. Scherrer, K. (2018) Primary transcripts: From the discovery of RNA processing to current concepts of gene expression - review. *Exp. Cell Res.* **373**, 1–33
5. Starck, S. R., and Shastri, N. (2016) Nowhere to hide: Unconventional translation yields cryptic peptides for immune surveillance. *Immunol. Rev.* **272**, 8–16
6. Laumont, C. M., and Perreault, C. (2018) Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cell. Mol. Life Sci.* **75**, 607–621
7. Granados, D. P., Laumont, C. M., Thibault, P., and Perreault, C. (2015) The nature of self for T cells-a systems-level perspective. *Curr. Opin. Immunol.* **34**, 1–8
8. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711
9. Nesvizhskii, A. I. (2014) Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125
10. Liepe, J., Marino, F., Sidney, J., Jeko, A., Bunting, D. E., Sette, A., Kloetzel, P. M., Stumpf, M. P., Heck, A. J., and Mishto, M. (2016) A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* **354**, 354–358
11. 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., and Purcell, A. W. (2018) A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Sci. Immunol.* **3**, eaar3947
12. Alfaro, J. A., Sinha, A., Kislinger, T., and Boutros, P. C. (2014) Onco-proteogenomics: Cancer proteomics joins forces with genomics. *Nat. Methods* **11**, 1107–1113
13. Freudenmann, L. K., Marcu, A., and Stevanović, S. (2018) Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology* **154**, 331–345
14. 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., *et al*. (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516
15. 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., *et al*. (2020) Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293
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. Cesnik, A. J., Miller, R. M., Ibrahim, K., Lu, L., Millikin, R. J., Shortreed, M. R., Frey, B. L., and Smith, L. M. (2020) Spritz: A proteogenomic database engine. *J. Proteome Res.* **20**, 1826–1834
18. Komor, M. A., Pham, T. V., Hiemstra, A. C., Piersma, S. R., Bolijn, A. S., Schelfhorst, T., Delis-van Diemen, P. M., Tijssen, M., Sebra, R. P., Ashby, M., Meijer, G. A., Jimenez, C. R., and Fijneman, R. J. A. (2017) Identification of differentially expressed splice variants by the proteogenomic pipeline splicify. *Mol. Cell. Proteomics* **16**, 1850–1863
19. Park, H., Bae, J., Kim, H., Kim, S., Kim, H., Mun, D. G., Joh, Y., Lee, W., Chae, S., Lee, S., Kim, H. K., Hwang, D., Lee, S. W., and Paek, E. (2014) Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. *Proteomics* **14**, 2742–2749
20. Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M. D., Clauser, K. R., Tabb, D. L., Mertins, P., Slebos, R., Erdmann-Gilmore, P., Li, S., Gunawardena, H. P., *et al*. (2016) An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell. Proteomics* **15**, 1060–1071
21. Zickmann, F., and Renard, B. Y. (2015) MSProGene: Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics* **31**, i106–i115
22. Wang, X., and Zhang, B. (2013) customProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235–3237
23. Wen, B., Xu, S., Zhou, R., Zhang, B., Wang, X., Liu, X., Xu, X., and Liu, S. (2016) PGA: An R/bioconductor package for identification of novel peptides using a customized database derived from RNA-seq. *BMC Bioinformatics* **17**, 244
24. Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., and Smith, L. M. (2014) Using galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **15**, 703
25. Li, Y., Wang, G., Tan, X., Ouyang, J., Zhang, M., Song, X., Liu, Q., Leng, Q., Chen, L., and Xie, L. (2020) ProGeo-neo: A customized proteogenomic workflow for neoantigen prediction and selection. *BMC Med. Genomics* **13**, 52
26. Cifani, P., Dhabaria, A., Chen, Z., Yoshimi, A., Kawaler, E., Abdel-Wahab, O., Poirier, J. T., and Kentsis, A. (2018) ProteomeGenerator: A framework for comprehensive proteomics based on de novo transcriptome assembly and high-accuracy peptide mass spectral matching. *J. Proteome Res.* **17**, 3681–3692
27. [dataset] Pandey, K., Mifsud, N. A., Lim Kam Sian, T. C. C., Ayala, R., Ternette, N., Ramarathinam, S. H., and Purcell, A. W. (2020) Immunopeptidome of an acute myeloid leukemia cell line THP1. *PRIDE*. https://doi.org/10.6019/PXD015039
28. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048
29. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
30. Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–2329
31. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript

assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515
32. Zhu, T., Liang, C., Meng, Z., Guo, S., and Zhang, R. (2017) GFF3sort: A novel tool to sort GFF3 files for tabix indexing. *BMC Bioinformatics* **18**, 482
33. Pertea, G., and Pertea, M. (2020) GFF utilities: GffRead and GffCompare [version 1; peer review: 3 approved]. *F1000Res.* **9,** ISCB Comm J-304
34. Wang, L., Wang, S., and Li, W. (2012) RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185
35. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013) From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33
36. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003) PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342
37. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012) PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**. M111.010587
38. Hulsen, T., de Vlieg, J., and Alkema, W. (2008) BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488
39. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787
40. Andreatta, M., and Nielsen, M. (2016) Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. *Bioinformatics* **32**, 511–517
41. Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017
42. Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J., and Rinner, O. (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121
43. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368
44. Pandey, K., Mifsud, N. A., Lim Kam Sian, T. C. C., Ayala, R., Ternette, N., Ramarathinam, S. H., and Purcell, A. W. (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
45. Partridge, T., Nicastri, A., Kliszczak, A. E., Yindom, L.-M., Kessler, B. M., Ternette, N., and Borrow, P. (2018) Discrimination between human leukocyte antigen class I-bound and co-purified HIV-derived peptides in immunopeptidomics workflows. *Front. Immunol.* **9**, 912
46. Sturm, T., Sautter, B., Wörner, T. P., Stevanović, S., Rammensee, H. G., Planz, O., Heck, A. J. R., and Aebersold, R. (2021) Mild acid elution and MHC immunoaffinity chromatography reveal similar albeit not identical profiles of the HLA class I immunopeptidome. *J. Proteome Res.* **20**, 289–304
47. Olsen, L. R., Tongchusak, S., Lin, H., Reinherz, E. L., Brusic, V., and Zhang, G. L. (2017) TANTIGEN: A comprehensive database of tumor T cell antigens. *Cancer Immunol. Immunother.* **66**, 731–735
48. Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018) The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705
49. Behrens, K., Maul, K., Tekin, N., Kriebitzsch, N., Indenbirken, D., Prassolov, V., Müller, U., Serve, H., Cammenga, J., and Stocking, C. (2017) RUNX1 cooperates with FLT3-ITD to induce leukemia. *J. Exp. Med.* **214**, 737–752
50. Zhao, Q., Laverdure, J.-P., Lanoix, J., Durette, C., Côté, C., Bonneil, É., Laumont, C. M., Gendron, P., Vincent, K., Courcelles, M., Lemieux, S., Millar, D. G., Ohashi, P. S., Thibault, P., and Perreault, C. (2020) Proteogenomics uncovers a vast repertoire of shared tumor-specific antigens in ovarian cancer. *Cancer Immunol. Res.* **8**, 544–555
51. Ehx, G., Larouche, J. D., Durette, C., Laverdure, J. P., Hesnard, L., Vincent, K., Hardy, M. P., Thériault, C., Rulleau, C., Lanoix, J., Bonneil, E., Feghaly, A., Apavaloaei, A., Noronha, N., Laumont, C. M., *et al*. (2021) Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity* **54**, 737–752.e710
52. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020) Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468
53. Yewdell, J. W. (2003) Hide and seek in the peptidome. *Science* **301**, 1334–1335
54. Sendoel, A., Dunn, J. G., Rodriguez, E. H., Naik, S., Gomez, N. C., Hurwitz, B., Levorse, J., Dill, B. D., Schramek, D., Molina, H., Weissman, J. S., and Fuchs, E. (2017) Translation from unconventional 5′ start sites drives tumour initiation. *Nature* **541**, 494–499
55. Dvinge, H., and Bradley, R. K. (2015) Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* **7**, 45
56. Jung, H., Lee, D., Lee, J., Park, D., Kim, Y. J., Park, W. Y., Hong, D., Park, P. J., and Lee, E. (2015) Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248
57. Smart, A. C., Margolis, C. A., Pimentel, H., He, M. X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K.-K., and Van Allen, E. M. (2018) Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.* **36**, 1056–1058
58. Cherry, S., and Lynch, K. W. (2020) Alternative splicing and cancer: Insights, opportunities, and challenges from an expanding view of the transcriptome. *Genes Dev.* **34**, 1005–1016
59. Wang, J., Zhu, S., Meng, N., He, Y., Lu, R., and Yan, G.-R. (2019) ncRNA-encoded peptides or proteins and cancer. *Mol. Ther.* **27**, 1718–1725
60. Wang, R. F., Parkhurst, M. R., Kawakami, Y., Robbins, P. F., and Rosenberg, S. A. (1996) Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J. Exp. Med.* **183**, 1131–1140
61. Weinzierl, A. O., Maurer, D., Altenberend, F., Schneiderhan-Marra, N., Klingel, K., Schoor, O., Wernet, D., Joos, T., Rammensee, H. G., and Stevanovic, S. (2008) A cryptic vascular endothelial growth factor T-cell epitope: Identification and characterization by mass spectrometry and T-cell assays. *Cancer Res.* **68**, 2447–2454
62. Saulquin, X., Scotet, E., Trautmann, L., Peyrat, M.-A., Halary, F., Bonneville, M., and Houssaint, E. (2002) +1 Frameshifting as a novel mechanism to generate a cryptic cytotoxic T lymphocyte epitope derived from human interleukin 10. *J. Exp. Med.* **195**, 353–358
63. Gao, Q., Liang, W. W., Foltz, S. M., Mutharasu, G., Jayasinghe, R. G., Cao, S., Liao, W. W., Reynolds, S. M., Wyczalkowski, M. A., Yao, L., Yu, L., Sun, S. Q., Chen, K., Lazar, A. J., Fields, R. C., *et al*. (2018) Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238.e223
64. Parker, R., Tailor, A., Peng, X., Nicastri, A., Zerweck, J., Reimer, U., Wenschuh, H., Schnatbaum, K., and Ternette, N. (2021) The choice of search engine affects sequencing depth and HLA class I allele-specific peptide repertoires. *Mol. Cell. Proteomics* **20**, 100124
65. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yılmaz, Ş., *et al*. (2018) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450