MCP TECHNOLOGICAL INNOVATION AND RESOURCES Special Issue: Immunopeptidomics

Pitfalls in HLA Ligandomics—How to Catch a Li(e)gand

Authors

Jens Fritsche, Daniel J. Kowalewski, Linus Backert, Frederik Gwinner, Sonja Dorner, Martin Priemer, Chih-Chiang Tsou, Franziska Hoffgaard, Michael Römer, Heiko Schuster, Oliver Schoor, and Toni Weinschenk

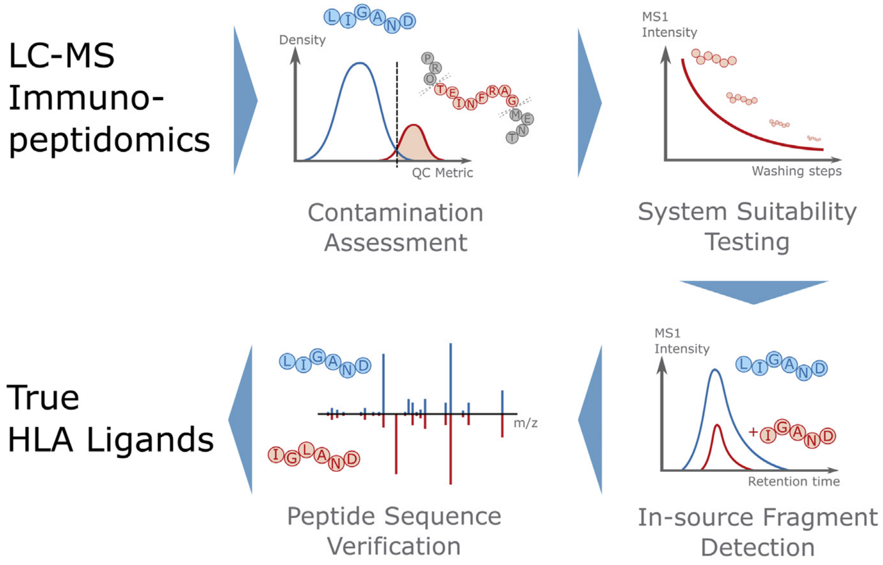
Correspondence

toni.weinschenk@immatics.com

In Brief

Accurate characterization of the HLA ligandome by mass spectrometry holds the key to unlock target-specific cancer immunotherapies such as adoptive cell therapies or bispecific T cell engaging receptors. Quality control at each step of the immunopeptidomics pipeline is relevant to differentiate between false and true HLA ligands. Here, we present computational and experimental methods as part of our XPRESIDENT technology platform that allow identification of true HLA ligands—an essential prerequisite for successful therapeutic development.

Graphical Abstract



Highlights

- Best practices to identify true HLA ligands as targets for cancer immunotherapies.
- Quality control in mass spectrometry to improve neoantigen/crypto-target discovery.
- Computational methods to assess fragment contamination in immunopeptidomics.
- Experimental methods for LC system suitability testing and HT sequence verification.

2021, Mol Cell Proteomics *20*, 100110

https://doi.org/10.1016/j.mcpro.2021.100110

Pitfalls in HLA Ligandomics—How to Catch a Li(e)gand

Jens Fritsche[1,‡], Daniel J. Kowalewski[1,‡], Linus Backert[1], Frederik Gwinner[1], Sonja Dorner[1], Martin Priemer[1], Chih-Chiang Tsou[2], Franziska Hoffgaard[1], Michael Römer[1], Heiko Schuster[1], Oliver Schoor[1], and Toni Weinschenk[1,2,*]

Knowledge about the peptide repertoire presented by human leukocyte antigens (HLA) holds the key to unlock target-specific cancer immunotherapies such as adoptive cell therapies or bispecific T cell engaging receptors. Therefore, comprehensive and accurate characterization of HLA peptidomes by mass spectrometry (immunopeptidomics) across tissues and disease states is essential. With growing numbers of immunopeptidomics datasets and the scope of peptide identification strategies reaching beyond the canonical proteome, the likelihood for erroneous peptide identification as well as false annotation of HLA-independent peptides as HLA ligands is increasing. Such "fake ligands" can lead to selection of nonexistent targets for immunotherapeutic development and need to be recognized as such as early as possible in the preclinical pipeline. Here we present computational and experimental methods that enable the identification of "fake ligands" that might be introduced at different steps of the immunopeptidomics workflow. The statistics presented herein allow discrimination of true HLA ligands from coisolated HLA-independent proteolytic fragments. In addition, we describe necessary steps to ensure system suitability of the chromatographic system. Furthermore, we illustrate an algorithm for detection of source fragmentation events that are introduced by electrospray ionization during mass spectrometry. For confirmation of peptide sequences, we present an experimental pipeline that enables high-throughput sequence verification through similarity of fragmentation pattern and coelution of synthetic isotope-labeled internal standards. Based on these methods, we show the overall high quality of existing datasets but point out limitations and pitfalls critical for individual peptides and how they can be uncovered in order to identify true ligands.

Immunotherapy has opened new ways to treat cancer. Knowledge about targets specific for tumor tissue is essential for successful treatment development (1). The targets relevant for immunotherapy are peptides presented by human leukocyte antigen (HLA) molecules. Therefore, efforts in determining HLA peptidomes (immunopeptidomics) have increased over years since its first application in 1991 (2). Liquid chromatography coupled to mass spectrometry (LC-MS) has become the method of choice for in-depth analysis of the immunopeptidome (3). The acquired data was used as starting point for the development of predictive models ranging from position-specific scoring matrices (PSSM) (4) to deep-learning approaches (5). Nevertheless, mass spectrometry remains the crucial factor for confirming presentation of peptides by HLA for a given tissue (6, 7).

Due to this growing importance, ensuring quality control throughout the entire immunopeptidomics workflow from HLA peptide isolation over LC-MS to sequence identification and HLA annotation is essential. Because of various pitfalls, peptide sequences reported from immunopeptidomics experiments can contain peptides that were never actually bound to HLA. The most prominent reason is that peptides are falsely identified. Immunopeptidomics is particularly prone to false discovery for multiple reasons: While the search space in proteomics contains about 330,000 tryptic peptides (8), immunopeptidomics database searching covers approximately 60 million theoretical class I peptides encoded by the proteome with lengths between 8 and 12 amino acids. While stringent control for false discovery rate (FDR) will reduce the problem, it will also substantially limit the sensitivity of identifying veritable HLA ligands. In addition, while proteomics operates under the hypothesis that the top-ranking peptide–spectrum match (PSM) is correct, this may not be true for immunopeptidomics where the much larger search space results in more mass ambiguities. Thus, low FDR alone does not sufficiently control for false positives. This problem becomes even more severe if proteogenomics approaches are applied. Biased inflation of the search space, *e.g.*, inclusion of hypothetical mutation events in neoantigen searches can lead to confirmation bias.

From the [1]Immatics Biotechnologies GmbH, Tübingen, Germany; and [2]Immatics US, Inc, Houston, Texas, USA
‡These authors contributed equally to this work.
*For correspondence: Toni Weinschenk, toni.weinschenk@immatics.com.

https://doi.org/10.1016/j.mcpro.2021.100110

Another reason for identifying peptides not derived from HLA is in-source fragmentation of true HLA ligands. This phenomenon has also been described for shotgun proteomics (9) and is an artifact of electrospray ionization generating b- and y-ions in the MS1 scan that might be selected for acquisition of MS/MS spectra that may be mistaken for HLA ligands. N-terminal source fragments are usually not identified since the water loss at the C-terminus is not a routinely applied dynamic modification in database searching. Still such events could lead to false-positive identification. C-terminal source fragments are more problematic since they generate a truncated sequence containing proper N- and C-terminus and will not be recognizable as an artifact directly. If these fragments display an HLA-binding motif by chance, they can easily be misannotated as HLA ligands.

HLA peptidomics is a fundamentally peptide-centric field, which requires attributing high relevance even to low-abundance peptide identifications and single PSMs. For this reason, highly stringent system suitability testing and quality controls are required to ensure high data quality. One notable aspect of this is the implementation of effective LC-cleanup protocols and monitoring strategies to avoid carryover of analyte between subsequent samples. Recent developments in LC may help mitigate carryover and reduce system flush times by using larger column diameters, higher flow rates (10), or by implementing disposable trap columns (11). However, these techniques were designed for proteomics and may not be ideally suited for the extremely limited sample amounts typically encountered in HLA peptidomics. For the prevalent technique of nano-LC in HLA peptidomics, users need to identify their system-specific extent and sources of carryover and design cleanup protocols to address these in the most effective order (*e.g.*, autosampler *versus* column flush cycles and equilibration).

Furthermore, peptides not related to antigen presentation can occur as result of proteolytic cleavage by endogenous proteases and peptidases, which has also been shown for proteomics (9). These enzymes originate from the sample analyzed and can be lysosomal endo- and exopeptidases (12) or peptidases specific for the analyzed tissue, for instance, carboxypeptidases for pancreas, aminopeptidases for intestine, and pepsin for stomach tissues. Previously described approaches for detection of such contaminations used the protein coverage as metric to exclude problematic proteins (13).

Here we present statistical and experimental methods that help to avoid these common pitfalls in immunopeptidomics and illustrate showcases that highlight the importance of addressing these issues.

EXPERIMENTAL PROCEDURES

Experimental Design and Statistical Rationale

For the statistical method development, one dataset was used to model the observed peptide properties while validation was performed on three datasets:

Modelling Dataset ZH2018—The statistical modeling of peptide properties was based on a population-scale immunopeptidomics dataset generated by the target discovery platform XPRESIDENT as described in Zhang *et al*. (14). In total, 1514 human tissue samples were measured in five technical replicates resulting in 7825 label-free LC-MS runs. This dataset was also used to analyze tissue-specific differences of proteolytic contamination. Fresh frozen tissues of 35 different organs and 23 tumor types were included with at least five donors per group and a median group size of 16 donors. Additionally, LC-MS data from 73 cell lines were analyzed.

Validation Dataset EC500 (Expert Review)—The dataset was generated by manual assignment of contaminant peptides by two immunopeptidomics experts. In total, 500 peptides were randomly selected from ZH2018. To ensure sufficient representation of contaminant peptides in the benchmark, a ratio of 1:5 between peptides with and without evidence for contamination was enforced.

Validation Dataset AB2017 (External Data)—This dataset is based on peptides reported in Abelin *et al*. (15) that were isolated from 16 monoallelic cell lines derived from the human B lymphoblastoid cell line 721.221. The dataset also included a negative control consisting of immunoprecipitations with beads lacking an HLA-specific antibody as well as immunoprecipitations of untransduced cells.

Validation Dataset GlyT98G (Mock Immunoprecipitation)—The dataset was generated by isolating peptides from the glioblastoma cell line T98G with a glycine-coupled column lacking an antibody recognizing HLA molecules, thus reflecting non-HLA-specific precipitation. The sample was analyzed by DDA-MS in three technical replicates.

Peptide Synthesis

Peptides were synthesized in 0.5 μmol scale with a filter tip-based approach on a Syro II synthesizer (Multisyntech) using solid-phase standard Fmoc-chemistry. For stable isotope-labeled (SIL-) peptides, C13N15-labeled Fmoc-amino acids were purchased from Eurisotop and loaded onto Tritylchloride-Polystyrene resins by Intavis.

Peptide ID Validation

For experimental control of false-positive identifications, we acquired two LC-MS runs, one as direct infusion of a synthesized version of the identified peptide and one with an SIL internal standard peptide spiked into a retention vial of the original sample. To control for synthetic contaminations, blank acquisitions of the synthetic standards were performed. To show the feasibility of this approach, we acquired data for peptides derived from the muscle isoform of pyruvate kinase (PKM) and the Kirsten rat sarcoma viral oncogene (KRAS).

Peptide Isolation

HLA peptides were isolated as described in ZH2018. The control precipitations for the validation dataset GlyT98G were analogously generated by running the cell lysate of 250 million T98G cells over CnBr-sepharose columns coupled with glycine instead of HLA-specific antibody. The HLA peptidome matrix used for validation of peptide identity of $\mathrm{PKM}_{28/3\text{-}12}$ *versus* $\mathrm{PKM}_{2\text{-}12}$ was generated by immunoprecipitation of 25 million cells of the lymphoblastoid cell line LCL11 using the HLA-DR-specific monoclonal antibody L243 (Department of Immunology, University of Tübingen, Germany).

Mass Spectrometry

LC-MS analysis of HLA peptide extracts was performed on a nanoACQUITY UPLC system (Waters) online coupled to an Orbitrap

Fusion mass spectrometer (Thermo Fisher). A trapping setup using Waters 25 cm × 75 μm BEH C18 analytical columns was used employing a stepped gradient ranging from 1 to 34.5 acetonitrile over 70 min for DDA-MS runs and 120 min for targeted MS runs. MS acquisition in data-dependent mode (DDA) was performed using a top speed method with a maximum cycle time of 3 s. MS1 scan range was set to 200–1500 m/z with an AGC target of 1e5 and 120k resolution. For FT runs, MS2 scans were acquired with an isolation width of 2 m/z for the topN precursors with an m/z range of 280–720 m/z and charge states of 2+ and 3+ at 30k resolution and an AGC target of 5e4 in the Orbitrap. Precursor fragmentation was performed by collision-induced dissociation (CID) at 35% normalized collision energy (NCE) or higher collisional dissociation (HCD) at 27% NCE, respectively. For ion trap (IT) runs, identical precursor selection and isolation were applied with CID at 35% NCE performed at 1e4 AGC target in the IT and scan speed set to normal. Dynamic exclusion was set to 13 s, corresponding to a "3-point-per-peak" acquisition scheme.

Targeted MS measurements in this manuscript correspond to tier 3 analyses as described by the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) Program. The peptide validation of PKM was performed using scheduled parallel reaction monitoring (sPRM) with CID FTMS2 acquisition at 35% normalized collision energy and 30k resolution targeting the 1+ and 2+ precursors of the native peptide and the two differentially labeled SIL internal standard peptides (20 fmol each).

Internal standard triggered PRM (IS-PRM) for large-scale coelution experiments and control of isotopic purity was performed by spiking 100 or 250 fmol of SIL-internal standard peptides and target mass list driven data-dependent triggering of CID FTMS2 scans (60k resolution) on the SIL target peptides with dependent offset scans on the corresponding unlabeled precursor m/z. Isolation windows were set to 2 m/z for all SIL-peptides except for SIL-Alanine, which was isolated at 1.1 m/z to avoid coisolation of labeled and unlabeled isotopologues. Additionally, chromatograms were checked specifically for differentially labeled transitions (y-ions for C-terminally labeled SIL-peptides), to preclude false-positive detection of coisolated unlabeled species.

For quality control of synthetic peptides (KRASG12V$_{2-35}$, SIL-PKM$_{2-12}$, and SIL-PKM$_{28/3-12}$), direct infusion MS was performed using 1 μl of 1pmol/μl synthetic peptide in 50% methanol/5% formic acid on an Orbitrap Velos mass spectrometer (Thermo Fisher, Waltham). MS1 scans were acquired with a scan range of 350–2000 m/z at 100k resolution. Peptide identity was confirmed based on CID- and HCD MS2 scans.

All data were acquired in profile mode.

Software

Targeted MS data analysis and MS1 filtering of DDA data were performed using Skyline (v20.2). Statistics and plots were generated in R v3.6.1 (16). Sequence logos were generated using the "Two Sample Logos" software v1.23 (17) using 9mer peptides sampled at random from the ENSEMBL reference proteome as negative dataset to reflect the background frequencies of amino acids in human samples.

Search Parameters and Acceptance Criteria

MS/MS spectra were first converted into mzXML format using msconvert.exe from ProteoWizard package (v3.0.20128.317991700) and then searched by X! Tandem (v2013.06.15.1), Comet (v2016012), and MSGF+ (v7102) against Ensembl 77 human protein sequences (99,436 entries) with addition of same number of reverse sequences as decoys. The MS/MS database search was done using the following parameters: peptide length of 6–15 AAs, mass range of 600–1500 Da, nonspecific enzyme cleavage, and oxidation of methionine as variable modification. By default, X! Tandem also includes N-terminal protein acetylation. Precursor and the fragment ion mass tolerance were set to 10 ppm and 15 ppm respectively for FT MS/MS spectra, and 10 ppm and 600 ppm respectively for IT MS/MS spectra. The search results from the search engines were individually analyzed by PeptideProphet *via* the Trans-Proteomic Pipeline (TPP) (v5.0), and then the results were further combined using iProphet, which estimates a probability score for each PSM with assistance of decoy hit scores. FDR was estimated by target-decoy approach based on iProphet probabilities, and all PSMs were filtered by 5% run-level FDR threshold. For each identified PSM, the precursor MS1 feature was extracted using the MS1 feature detection algorithm described in (18). In order to align retention times among all the LC-MS runs, we empirically selected 1000 peptides, which are commonly identified in most of runs and established globally aligned retention time (gRT) for each of them in the range between 0 and 100. Each run was then aligned to the gRT scale [0–100] by polynomial regression.

RESULTS

Proteolytic Fragments

Proteolytic degradation by endogenous proteases and peptidases introduces peptides to immunopeptidomics that are not derived from HLA. This usually affects highly abundant proteins and creates characteristic peptide ladders, which results in high coverage of the protein by peptides of various lengths (Fig. 1*A*).

To allow detection of such proteins, we defined a *protein coverage ratio* for each protein (*P*) as average number of class I peptides (*p*) per amino acid:

$$protein\ coverage\ ratio\ for\ protein\ P = \frac{1}{L(P)} \sum_{p \in P} L(p);$$

$$L(x) = Number\ of\ amino\ acids\ per\ protein\ or\ peptide\ x$$

While protein coverage in proteomics is defined as fraction of amino acids of a protein covered by at least one peptide, our proposed metric is closer to the definition of coverage used in transcriptomics, for instance, fragments-per-kilobase-per-million. Since the proteomics definition of coverage is bound between 0 and 100%, it will not reflect repeated coverage of the protein by distinct peptides and therefore result in censoring at 100%. In contrast, the proposed metric incorporates that knowledge and results in a ratio distribution approximately following a log-normal distribution. Inspecting the distribution of protein coverage ratios across the ZH2018 dataset showed a subpopulation of genes with high protein coverage ratio (Fig. 1*B*) that originated from housekeeping genes such as hemoglobin and actin (Table 1) indicating proteolytic cleavage contaminations.

To distinguish real HLA ligands from proteolytic contaminations, we deconvoluted the two distributions (ligand and contamination distribution) using an expectation-

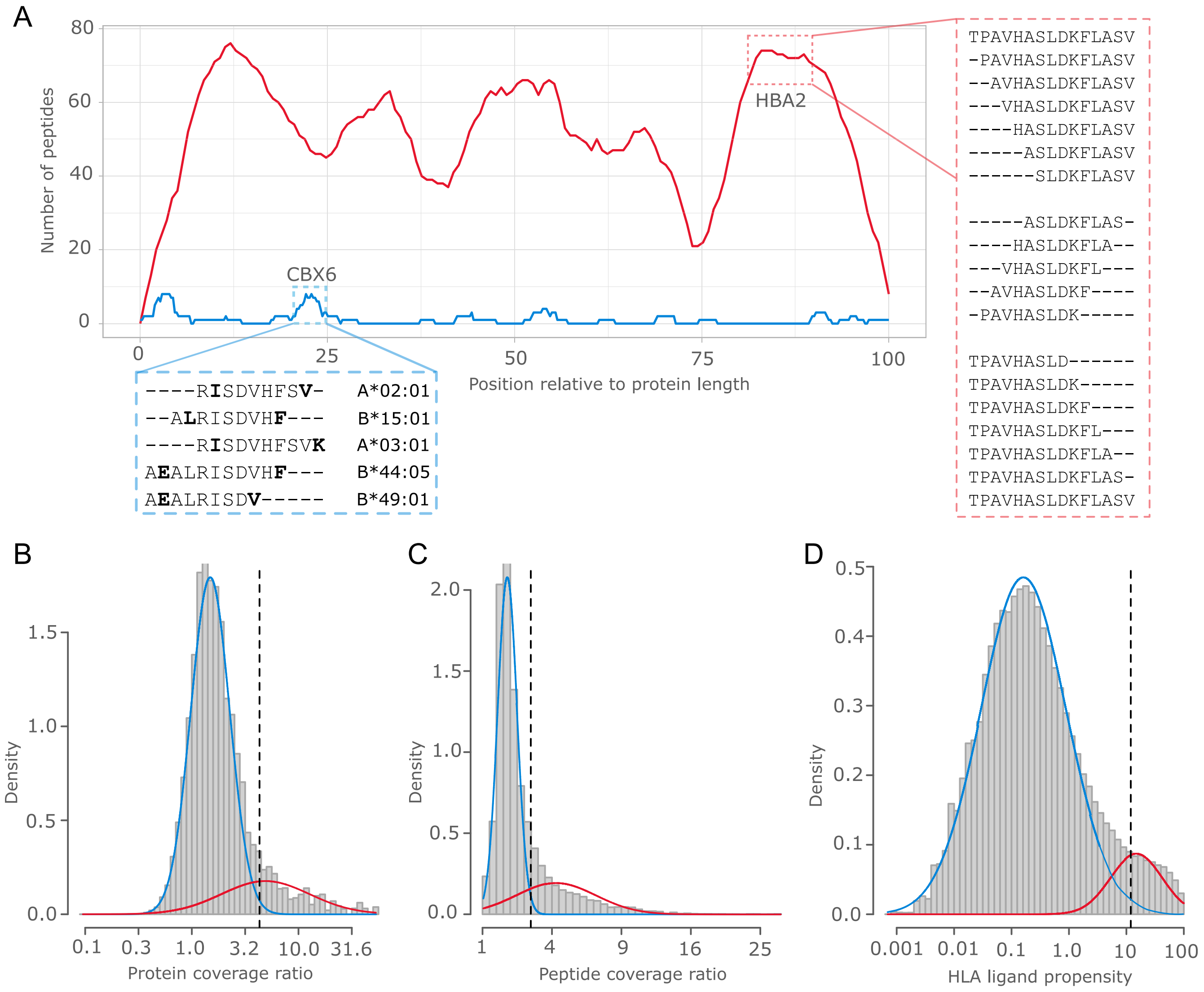


FIG. 1. **Protein coverage by class I HLA ligands.** *A*, comparison of coverage for a representative gene (chromobox homolog 6, CBX6) showing presentation hotspot (anchor amino acids in *bold*) and potential contaminant (hemoglobin alpha 2, HBA2) with characteristic peptide ladders due to proteolytic degradation. Contaminations are colored in *red* while true HLA ligands are colored *blue*. *B*, distribution of protein coverage ratios in ZH2018. Estimated Gaussian mixture model of contamination distribution shown in *red* and ligand distribution in *blue*. The cutoff for assessment of contaminations at 1% false discovery rate is visualized as *dashed line*. *C*, distribution for peptide coverage ratio and *D*, HLA ligand propensity with analogical modeling as before.

maximization (EM) algorithm. The EM algorithm was repeated 5000 times to avoid local minima of the optimization. After model fitting, we determined a cutoff for the protein coverage ratio that distinguishes HLA ligands from proteolytic contaminations with an FDR of 1%. The threshold was determined as 4.312, which means that for such a protein on average 4.312 distinct peptides cover each amino acid. Proteins with a coverage ratio above the threshold are more likely from the contamination distribution and only 1% of proteins from the ligand distribution remain. All peptides derived from proteins above this coverage ratio will be flagged as potential contaminations.

Although such a protein-centric approach is appropriate for proteomics, immunopeptidomics analyses have a peptide-centric focus, thus requiring additional metrics to further disambiguate veritable HLA ligands from unspecific peptides. Thus, in addition to the protein coverage ratio, we calculated the *peptide coverage ratio* for each target peptide (p) as the number of HLA class I peptides (q) that are assigned to the same protein and overlap in their position with the target

Table 1

Top ten genes with highest protein coverage ratio in ZH2018

Gene	Gene description	Protein coverage ratio
HBA2	hemoglobin, alpha 2	52.78
HBA1	hemoglobin, alpha 1	52.78
HBB	hemoglobin, beta	43.97
ACTB	actin, beta	39.27
ACTG1	actin, gamma 1	39.06
ACTA2	actin, alpha 2, smooth muscle, aorta	35.09
ACTG2	actin, gamma 2, smooth muscle, enteric	34.84
ACTC1	actin, alpha, cardiac muscle 1	34.83

peptide. Based on the assumption that proteolysis will generate sample-specific peptide sequence ladders, target and overlapping peptide were required to be identified in the same sample. For the final ratio, sample-specific ratios were averaged across all samples (S).

peptide coverage ratio for peptide p

$$= \frac{1}{card(S)} \sum_{s \in S} \left(\frac{1}{L(p)} \sum_{q \text{ overlaps with } p \text{ in } s} L(q) \right)$$

$$card(x) = Cardinality\ of\ sample\ set$$

While the peptide coverage ratio should work well for identification of contaminations that are detected frequently, low-abundant contaminations with fewer detections pose a problem. To extend applicability to this set of peptides, a third metric was used that was based on the propensity of being an HLA ligand. This property is best reflected by models trained on HLA ligand data rather than on *in vitro* binding data alone. Thus, we used the NetMHCpan-4.0 ligand (EL) rank score (20) for assessment of HLA ligand propensity. This score was calculated as the minimal score across all six HLA allotypes (H) of each sample and further averaged across all samples (S) that present the peptide.

HLA ligand propensity of peptide p

$$= \frac{1}{card(S)} \sum_{s \in S} min_{h \in H(s)} \mathrm{Rank}(p, h)$$

Again, distribution and threshold were determined as described above resulting in a threshold of 11.924 at 1% FDR (Fig. 1*D*). This means that all peptides with an average minimal NetMHCpan rank of 11.924 or higher were marked as potential contaminations.

Taking all three metrics into account, the *proteolytic contamination count* (PCC) was computed by counting the number of passing metrics.

$$Proteolytic\ Contamination\ Count\ \mathrm{PCC} = \left\{ \begin{array}{ll} 1 & if\ protein\ coverage\ ratio > 4.312 \\ 0 & otherwise \end{array} \right\} + \left\{ \begin{array}{ll} 1 & if\ peptide\ coverage\ ratio > 2.874 \\ 0 & otherwise \end{array} \right\} + \left\{ \begin{array}{ll} 1 & if\ HLA\ ligand\ propensity > 11.924 \\ 0 & otherwise \end{array} \right\}$$

The sample-specificity constraint also reduces the effect of nested HLA ligands derived from presentation hotspots. Nested peptides might occur naturally due to consecutive anchor amino acids. For instance, the leucine at second and third amino acid positions of the A*02:01 ligand S**L**LDGFLATV allows for the presentation of the shorter peptide L**L**DGFLATV (14). More importantly, different HLA allotypes can generate overlapping peptides in presentation hotspots (19) that could be interpreted as potential contamination when considering all peptides across the dataset for coverage determination (Fig. 1*A*).

The peptide coverage ratio distribution from ZH2018 was comparable to the protein coverage ratio distribution and a threshold was estimated using mixture modeling as before (Fig. 1*C*). The resulting threshold value for 1% FDR was 2.874, which means that on average each amino acid is covered by 2.874 overlapping peptides averaged across all samples.

To evaluate an appropriate cutoff for the PCC score, we used the EC500 benchmark dataset annotated by two immunopeptidomics experts and reflecting a ratio of 1:5 between peptides with and without PCC score (Table S1). These annotations allowed to estimate performance metrics shown in Table 2. A threshold of 1 (PCC ≥ 1) provided a balanced sensitivity (85.7%) and specificity (93.3%), while a threshold of 2 (PCC ≥ 2) allowed to increase specificity to 99.5% at 51.2% sensitivity. For discovery of HLA peptide targets, the specificity is the primary concern since this metric indicates how many true ligands could be identified. Sensitivity on the other hand describes how many proteolytic fragments could be identified, which is relevant to increase data quality but only a secondary objective compared with the risk of losing relevant targets. Thus, the threshold of PCC ≥ 2 was set for further investigations.

To experimentally evaluate the sensitivity of the proposed contamination prediction, we inspected the GlyT98G dataset

Table 2
Confusion matrices and performance metrics at different proteolytic contamination count scores

Peptide subsets and performance metrics	PCC ≥ 0	PCC ≥ 1	PCC ≥ 2	PCC ≥ 3
Contamination marked as contamination (True Positive)	84	72	43	10
Ligand marked as contamination (False Positive)	416	28	2	0
Ligand marked as ligand (True negative)	0	388	414	416
Contamination marked as ligand (False negative)	0	12	41	74
Percentage of detected contaminations (Sensitivity)	100.00%	85.70%	51.20%	11.90%
Percentage of detected ligands (Specificity)	0.00%	93.30%	99.50%	100.00%

Table includes confusion matrices showing the number of peptides in each group (true/false positive and true/false negatives) as well as performance metrics (sensitivity and specificity) for contamination assignment at different proteolytic contamination count (PCC) scores using expert review benchmark EC500.

containing eluting peptides from a T98G cell line using a mock immunoprecipitation for which 59.6% of all peptides were derived from keratins, immunoglobulins, or albumin. Determining the PCC score for each eluted peptide (Table S2) showed that 62.8% were marked as proteolytic contamination, which is well in line with the 51.2% sensitivity determined by the EC500 benchmark.

We further benchmarked the method using the external AB2017 dataset. The peptides reported from the original publication were annotated with the PCC score (Table S3). The dataset included 16 monoallelic cell lines and negative controls that were generated by immunoprecipitation with beads lacking an HLA-specific antibody or by use of untransduced cells. For 79% of these peptides (n = 22,017), we were able to provide annotation. Inspecting the number of proteolytic fragments for the negative control dataset showed a fraction of 46.7%, which is in a similar range to the 62.8% derived from the GlyT98G dataset. For the monoallelic cell line samples, the overall fraction of contaminant peptides was 2.7% with values between 0.2% for A*24:02 and 14.5% for A*68:02 (Fig. 2*A*).

To investigate if sample input has an effect on the number of proteolytic fragments, we isolated peptides from one tumor tissue divided in aliquots of increasing size. This experiment showed a constant baseline level of proteolytic fragments independent of sample input (Fig. 2*B*).

To elucidate if certain tissues were more affected by proteolysis than others, we determined the number of proteolytic contaminant peptides from pan-class I specific antibody (W6/32) preparations presented in ZH2018 (Fig. 2*C*). For cell lines as well as liquid cancers, we observed around 50 contaminant peptides (1%), which is in line with the findings in AB2017. For primary solid tissues, the number was twice as high with about 80 fragments for cancer and healthy tissues. Elevated levels were observed for digestive organs such as the stomach, esophagus, and digestive glands, with 180 and 470 contaminations on average for cancer and normal tissues, respectively. Non-digestive normal samples, in particular derived from granulocytes, also showed an elevated number of proteolytic fragments with a median of 440 contaminations.

Chromatographic Carryover

In order to establish effective LC-cleanup protocols and monitoring strategies that avoid carryover of analyte between subsequent samples, we inspected the extent and sources of carryover in our chromatographic system by acquiring consecutive blank injections (Fig. 3*A*). Typically, the most abundant peptides, such as prevalent housekeepers, were most prone to carryover in subsequent experiments as demonstrated by the high percentile ranks of MS1 peak areas shown in Figure 3, *B* and *C*. Despite the fact that this may only marginally affect relative quantification among samples of a specific HLA allotype (*e.g.*, immunoprecipitations with the A*02-specific mAb BB7.2), carryover must nevertheless be rigorously minimized to avoid contamination among samples expressing different HLA types as well as carryover of sample specific peptides, *e.g.*, tumor-associated peptides carried over into normal samples. These qualitative and quantitative data from blank injections after different flushing regimens were used to guide protocol design (Fig. 3*D*). The experimental checks for carryover were routinely conducted as part of system suitability testing (SST) prior to sample acquisition using identical LC-MS methods as succeeding analytical runs.

To illustrate the detrimental effects that the observed carryover might have, we performed a simulation using monoallelic cell line data of HLA-A*24:02 and B*44:03 from AB2017. Following the carryover characteristics that we observed in our system (Fig. 3*A*), we computationally added the 6% most abundant peptides from the A*24:02 to the B*44:03 data, thus assuming consecutive acquisition of A*24:02 and B*44:03 cell lines with omission of adequate LC cleanup protocols in between. We compared the sequence logos of the original B*44:03 dataset (Fig. 3*E*) with the simulated dataset (Fig. 3*F*). While for B*44:03 glutamate is the preferred anchor amino acid in the second position, the simulated dataset shows reduced information content with tyrosine being wrongly enriched as a secondary anchor due to the carryover from A*24:02. Thus, without controlling for carryover, spurious training data might be generated and

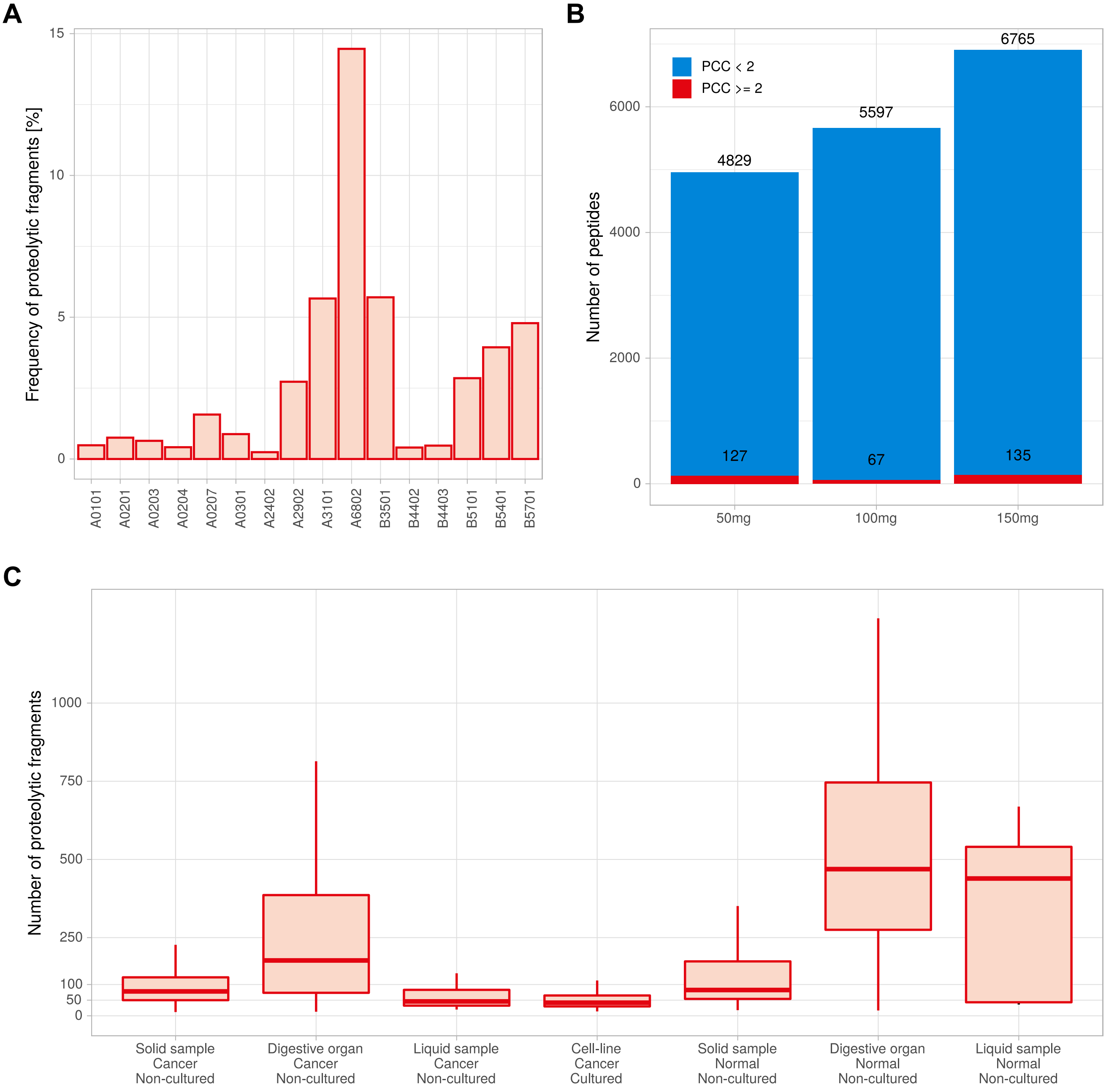
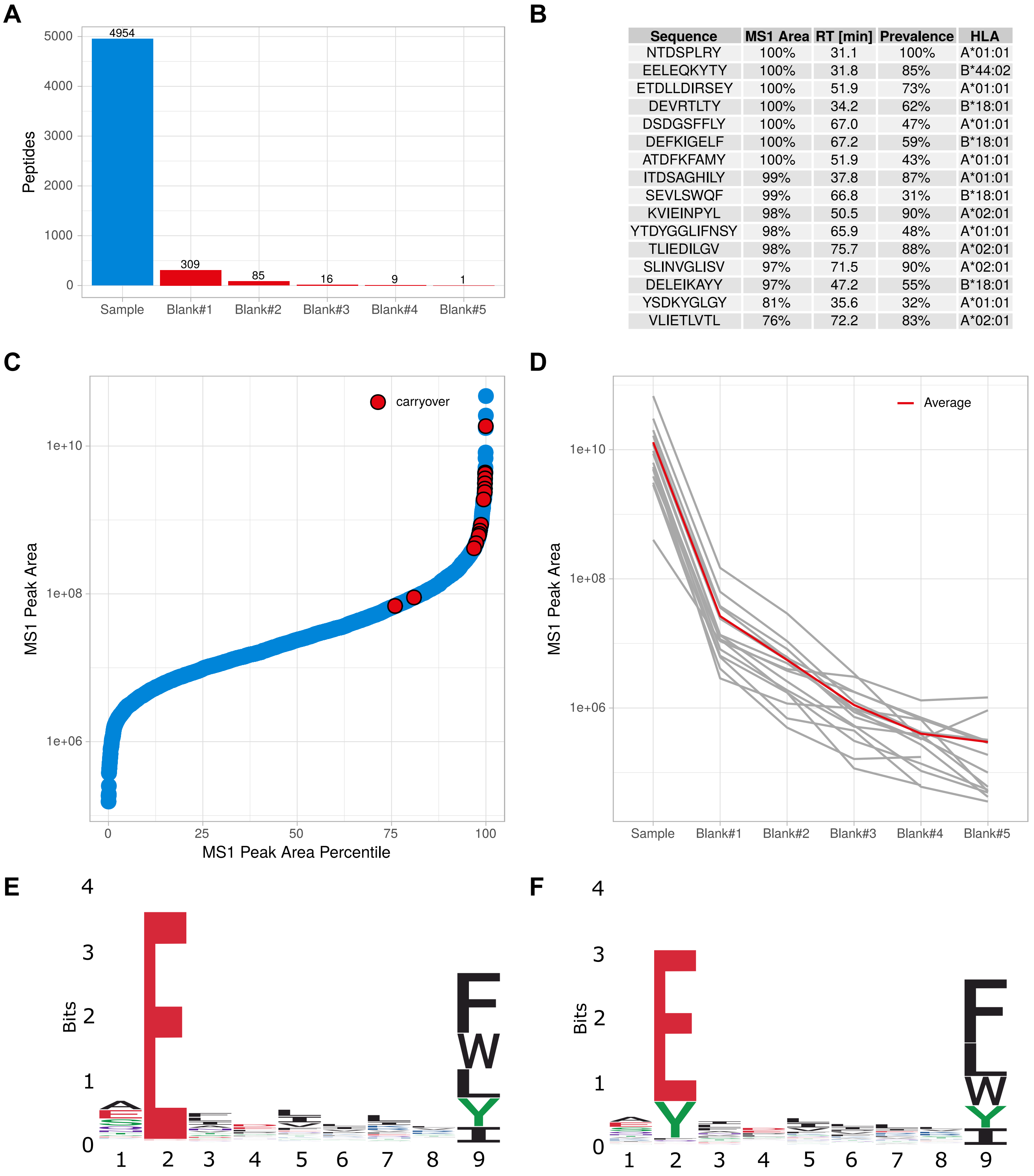


FIG. 2. **Contamination analyses-based proteolytic contamination count PCC ≥ 2.** *A*, proteolytic fragments observed in monoallelic cell lines of AB2017 for different HLA alleles with low frequency of contamination on average but substantial increase in HLA-A*68:02. *B*, proteolytic fragments as a function of sample input (tumor tissue weight) for lung cancer adenocarcinoma tissue showing low constant baseline levels of contaminations. *C*, number of proteolytic fragments in population-scale immunopeptidome dataset ZH2018 aggregated according to tissue highlighting three groups of degree of contamination: low (cell lines and uncultured liquid cancers), medium (solid tissue), and high (digestive organs and healthy liquid tissue).

misleading conclusions would be drawn from this data with regard to HLA-binding characteristics. While the observed effect shown here serves as an example, carryover is system-specific and needs to be addressed dependent on the individual system in use.

In-Source Fragmentation

An example of potential misinterpretation of immunopeptidomics data due to in-source fragmentation during electrospray ionization is the peptide TP53$_{189\text{-}197}$ (APPQHLIRV)



Sequence	MS1 Area	RT [min]	Prevalence	HLA
NTDSPLRY	100%	31.1	100%	A*01:01
EELEQKYTY	100%	31.8	85%	B*44:02
ETDLLDIRSEY	100%	51.9	73%	A*01:01
DEVRTLTY	100%	34.2	62%	B*18:01
DSDGSFFLY	100%	67.0	47%	A*01:01
DEFKIGELF	100%	67.2	59%	B*18:01
ATDFKFAMY	100%	51.9	43%	A*01:01
ITDSAGHILY	99%	37.8	87%	A*01:01
SEVLSWQF	99%	66.8	31%	B*18:01
KVIEINPYL	98%	50.5	90%	A*02:01
YTDYGGLIFNSY	98%	65.9	48%	A*01:01
TLIEDILGV	98%	75.7	88%	A*02:01
SLINVGLISV	97%	71.5	90%	A*02:01
DELEIKAYY	97%	47.2	55%	B*18:01
YSDKYGLGY	81%	35.6	32%	A*01:01
VLIETLVTL	76%	72.2	83%	A*02:01

FIG. 3. **Peptide carryover observed after a single nanoLC-MS run of a representative pan-HLA class I (W6/32) peptidomics sample.** *A*, carryover peptide IDs identified in database search of blank runs performed subsequently to the sample analysis. *B*, for 16 carryover peptides retained until Blank #3 different characteristics are shown including the percentile of MS1 peak area, retention time, prevalence in ZH2018 samples, and HLA allotype. *C*, MS1 peak areas and their percentile for baseline sample (*blue*) and the 16 carryover peptides (*red*). *D*, reduction of MS1 signal intensity for carryover peptides across repeated blank runs shown for all 16 peptides and fitted (*red line*) using an exponential decay model resulting in $\log_{10}\text{Area}(t) = 5.21 + 4.60e^{-0.72t}$. *E*, sequence logo of HLA-B*44:03 for peptides from AB2017. *F*, sequence logo for simulated data assuming carryover from HLA-A*24:02 into B*44:03. The 6% most abundant peptides from the A*24:02 data were computationally admixed to B*44:03, simulating consecutive LC-MS analysis without intermittent LC cleanup.

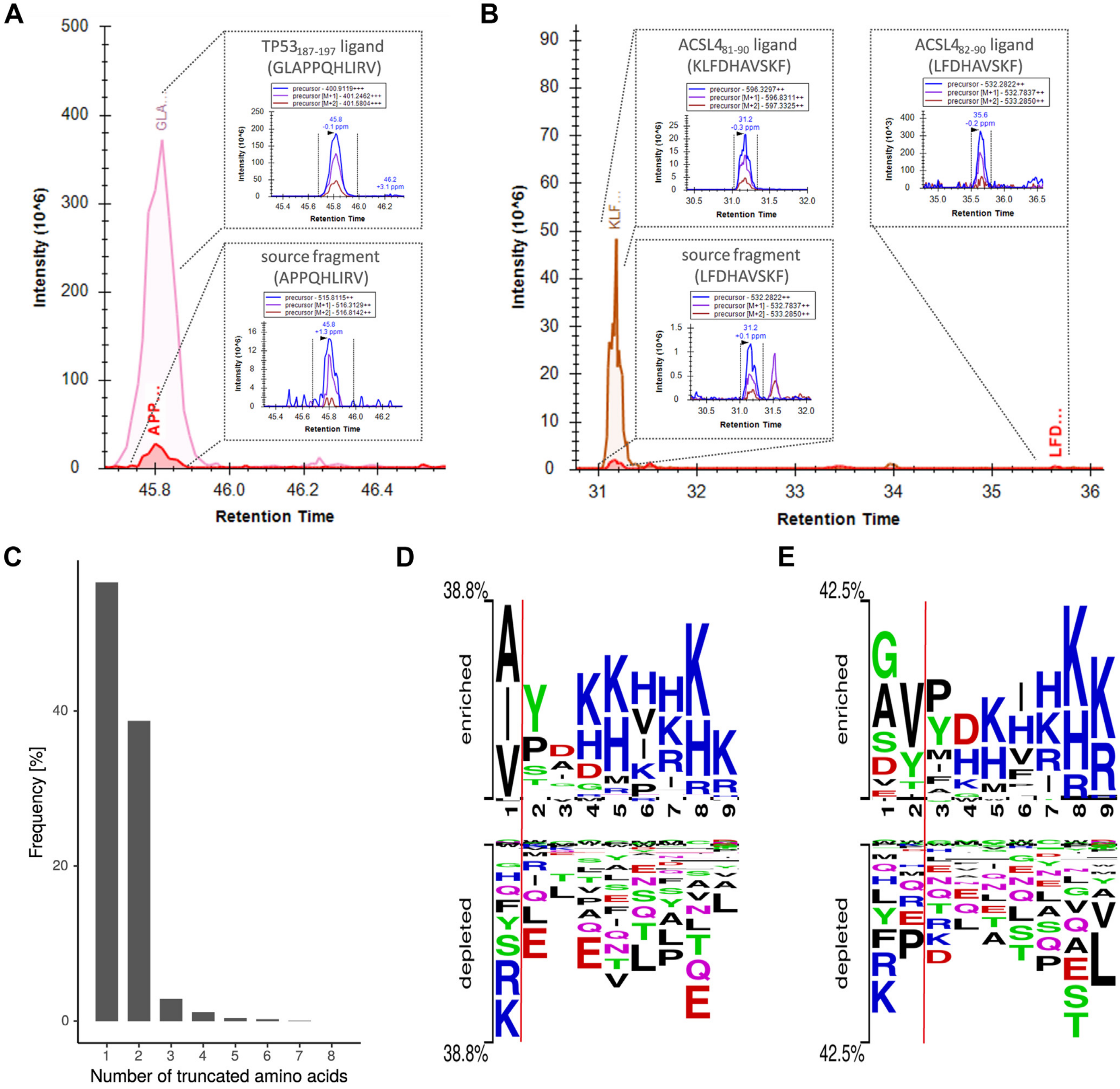


Fig. 4. **In-source fragmentation.** *A*, example of a source fragmentation of the TP53-derived A*02:01 ligand $TP53_{187-197}$ (GLAPPQHLIRV) resulting in concomitant detection of the truncated C-terminal source fragment APPQHLIRV on an HLA-A*02 specific immunoprecipitation using the antibody BB7.2. As the fragment fulfills the consensus motif for B*07 and B*35, it may be misinterpreted as a veritable HLA ligand. *B*, example of a peptide that can occur both as a source fragment and as a veritable HLA ligand. Extracted MS1 ion chromatograms are shown for a pan-HLA class I immunoprecipitation using the antibody W6/32 on a sample expressing two restricting allotypes (A*03:01 & C*04:01). The short variant LFDHAVSKF was detected as a source fragment of the A*03:01 ligand at experimental RT 31.2 and as a veritable C*04:01 ligand at experimental RT 35.6. *A* and *B*, the main panel summarizes precursor ion intensities of the two peptide species over retention time. The cutouts provide the underlying MS1 extracted ion chromatograms including precursor m/z. *C*, frequency distribution of N-terminal amino acid loss by in-source fragmentation. *D*, sequence logo of peptides showing loss of one N-terminal amino acid by source fragmentation. *E*, sequence logo of peptides with loss of two amino acids.

derived from tumor protein p53 (Fig. 4*A*). This peptide is a source fragment of the HLA-A*02:01 ligand $TP53_{187-197}$ (GLAPPQHLIRV), which has been described as potential tumor target (21). The proline in position 2 of $TP53_{189-197}$ suggests that it belongs to the B07 supertype (22), which is also reflected by the prediction score for members of this supertype (0.43 NetMHCpan4.0 EL rank for HLA-B*07:02 and 0.51 for HLA-B*35:03). Accordingly, $TP53_{189-197}$ has been suggested as an HLA-B*07 target in a study using predictive approaches and binding affinity assays (23). We identified the fragment on the glioblastoma cell line T98G (Fig. 4*A*), which is positive for HLA-A*02:01 and B*35:03. While this could be interpreted as proof of peptide presentation, the peptide was only detected in the immunoprecipitation that used the HLA-A*02-specific antibody BB7.2 and not the concomitantly used pan-class I-specific antibody W6/32, thus providing strong evidence for A*02:01 restriction. This finding is also supported by the IEDB assignment as HLA-A*02:01 based on Jensen *et al*. (24).

Fragmentation of peptides is easily detectable by searching for peptide pairs where one peptide sequence (source fragment) is a substring of a longer sequence (precursor sequence) and both peptides are coeluting. We determined the absolute retention time difference (deltaRT) between the apexes of the extracted ion chromatograms of the monoisotopic peaks of both peptides. To account for slight variations in apex estimation, peptides with deltaRT below 0.1 were considered as coeluting.

While this run-wise detection of cases of source fragmentation is relatively straightforward, single occurrences of this event do not necessarily disqualify a peptide from being a ligand. Since most gradients are optimized for high and uniform identification rates, elution of nested sets of HLA ligands might occur in close temporal proximity leading to small deltaRT and thus potentially to false assignment of source fragmentation. In other cases, the same peptide sequence might be a source fragment as well as a genuine HLA ligand, depending on sample HLA type.

To address such challenges, we determined the *source fragmentation fraction* (SFF) for a peptide *p* as the proportion of LC-MS acquisitions in which *p* was detected as source fragment of any precursor peptide *q* relative to all experiments R where *p* was identified.

$$\textit{Source Fragmentation Fraction of peptide } p = \frac{1}{card(R)} \sum_{r \in R} \begin{cases} 1 & if\ |RT(p,r) - RT(q,r)| < 0.1 \\ 0 & otherwise \end{cases}$$

If the peptide was annotated as a source fragment in less than 26.4% of all identifications, *i.e.*, SFF<0.264, the peptide was not considered as a contamination in general. This threshold was derived as described above by mixture modeling and 1% FDR threshold estimation.

One example of a peptide that can occur as a veritable HLA ligand as well as a source fragment is ACSL4$_{82-90}$ (LFDHAVSKF) from acyl-CoA synthetase long-chain family member 4. The source fragment is derived from the ligand ACSL4$_{81-90}$ (KLFDHAVSKF) that can be presented by HLA-A*03, HLA-A*32, and HLA-B*15. ACSL4$_{82-90}$ was found as source fragment in the ZH2018 dataset for two donors at an average globally aligned retention time (gRT) of 43 (Table S4). However, ACSL4$_{82-90}$ was in total identified on 53 donors. The fact that these donors are predominantly C*04 positive and the elution at a different gRT of 47 indicated that in these cases the peptide is not a source fragment but rather a C*04 ligand. This was reflected by a source fragmentation score below threshold (SFF = 0.002). We even identified one sample positive for HLA-C*04 and HLA-A*03, which presented both versions of ACSL4$_{82-90}$, the C*04 ligand at 47 gRT equivalent to an experimental retention time (eRT) of 35.6 min and the source fragment of the A*03 ligand at a gRT of 43 (eRT at 31.2 min, see Fig. 4*B*). Thus, in general the peptide cannot be considered a contamination.

In order to validate our method using an external dataset, we used the SFF score to annotate all the identified peptides from the AB2017 dataset (Table S3) and found 45 identifications flagged as source fragmentations representing 0.2% of all IDs. Since in-source fragmentation is dependent on the particular instrument and its settings, we looked for peptides coeluting in the original raw data with peptides with N-terminal extension of one or two amino acids. This returned 43 peptides, and 31 among them were flagged by the SFF score. This overlap of 72.1% between the datasets underlined the generalizability of the described approach to other datasets.

Inspecting the characteristics of source fragmented peptides in ZH2018 revealed that the loss of one or two amino acids was the most prevalent case (Fig. 4*C*). The sequence logo of these peptides showed an enrichment of amino acids with positively charged side chains (Arg, Lys, His) for the C-terminal part of the peptide with fragmentation sites enriched most significantly for proline or tyrosine residues (Fig. 4, *D* and *E*). This finding is in line with the established “proline effect” in CID, which arises due to a particularly low threshold energy for the cleavage of the amide bond N-terminal to proline and is enhanced by a high proton affinity of the C-terminal proline-containing fragment and a low proton affinity of the N-terminal fragment (25). Accordingly, the N-terminal fragments observed in our data show enrichment of amino acids with low proton affinity (Gly, Ala, Val), whereas high proton affinity amino acids (Arg, Lys) are depleted.

False-Positive Identifications

Although false-positive identification might be addressed computationally, final experimental validation of the peptide sequence is mandatory for moving such peptides forward into clinical pipelines, especially if search spaces beyond the standard reference proteome are considered (26). Spectral confirmation by acquisition of MS/MS of a synthetic peptide counterpart and subsequent comparison with the eluted MS/MS spectra provides a first level of evidence. Yet, due to mass ambiguities, the likelihood of high spectral similarity for spectra from different peptides increases the more similar the peptides are.

One example for this is the peptide EKPHSEAGTAF, a putative proteasomally spliced peptide, which we aimed to verify on a lymphoblastoid cell line after initial identification in a

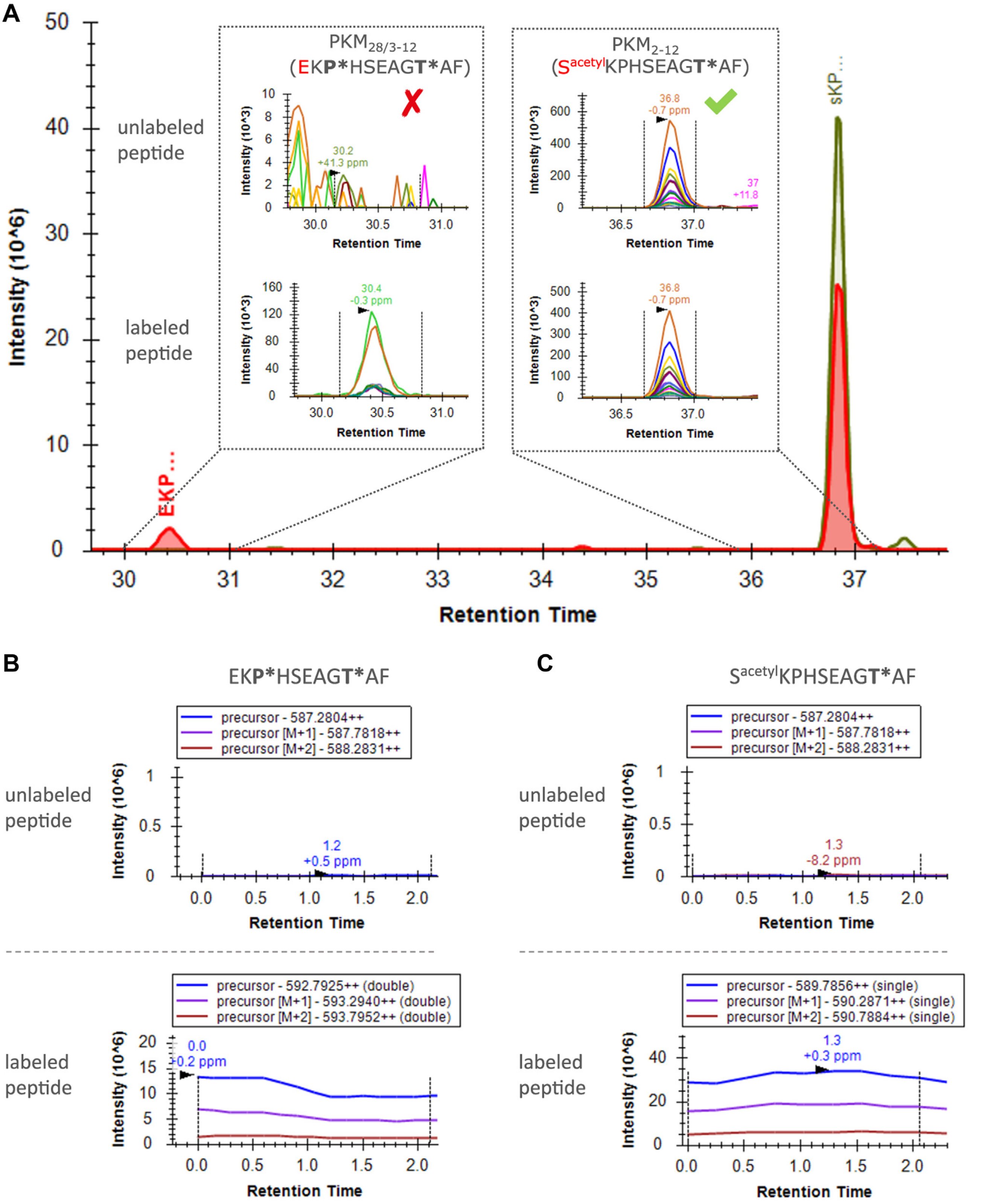


FIG. 5. **Peptide sequence verification.** Experimental disambiguation of conflicting peptide sequence annotations by spike-in of stable isotope-labeled internal standard (SIL-) peptides. *A*, differentially labeled SIL counterparts for two possible spectral annotations were spiked into an HLA peptidome sample previously found positive for the spectrum in DDA-MS and analyzed by targeted MS (PRM). For the putative proteasomally spliced sequence for which the IS elutes at RT 30.5 min, no endogenous unlabeled signal was detected. For the alternative sequence

customized database search containing proteasomally spliced peptides reported by Liepe *et al*. (27). While spectral similarity was supportive, a subsequent coelution experiment using a SIL internal standard peptide did not confirm coelution of the natural and labeled isotopologues. However, an endogenous signal was present at a different retention time. Since glutamate and acetylated serine have identical mass, the S-acetylated canonical peptide S^{acetyl}KPHSEAGTAF (PKM$_{2\text{-}12}$) was also a candidate for this signal. Of note, this peptide was previously described in HLA peptidomics analyses of U937 (histolytic lymphoma cell line) (28), HCC1143, and fibroblast cell lines (28), while the acetylation site was confirmed in proteomics analyses by Jacome *et al*. (29). Synthesis of differentially labeled SIL peptides and coelution confirmed that, in our data, only the S-acetylated canonical peptide PKM$_{2\text{-}12}$ was detected (Fig. 5*A*). To control for isotopic purity of the labeled internal standard, quality control analyses were performed by direct infusion MS (Fig. 5, *B* and *C*). While in this particular example omission of a variable modification from the search space resulted in confirmation bias, any ultralarge protein database search strategy needs to carefully address this issue since FDR-controlled database search is hampered by inclusion of all conceivable mechanisms of peptide biogenesis and modification, *e.g.*, posttranslational modifications, sequence artifacts, single-nucleotide polymorphisms, cryptic peptides originating from small open reading frames or other sources not covered by the reference proteome. While computational approaches have been developed to reduce false positives (30), final confirmation can only be obtained through experimental validation based on fragment spectrum identity and matching retention time.

In order to streamline peptide ID confirmation, we established a two-step pipeline consisting of in-house synthesis of SIL-labeled peptides and automated spectral comparison followed by coelution using internal standard triggered parallel reaction monitoring LC-MS (IS-PRM) (31). In the first step synthetic reference spectra acquired for all relevant fragmentation modes and collision energies were shifted *in silico* to compensate for the mass offset introduced by SIL labeling and then compared with the eluted spectra considering spectra with a spectral correlation of above 0.865 as positive match. Peptides passing this initial filter for spectral similarity were subjected to IS-PRM coelution LC-MS using SIL peptides spiked into the sample for final confirmation of peptide identity. While IS-PRM enables the evaluation of >100 peptides in a single run without requiring time-consuming method setup and RT scheduling, the initial step of spectral validation is an effective means to limit the number of peptides in coelution to the most relevant candidates. This is particularly relevant in case the initial search was performed on a large search space and/or using relaxed FDR filtering.

Synthetic peptides are essential for experimental validation; however, their utilization as internal standards in LC-MS or as substrates for validation assays, *e.g.*, *in vitro* proteasomal digestion can introduce another pitfall in the form of synthetic artifacts. For SIL internal standard peptides isotopic impurities can pose a significant problem, which may lead to false-positive peptide ID validation if no diligent inspection of negative controls is performed. Highly pure commercially available SIL amino acids and peptides are typically specified at >99% isotopic enrichment. However, the high sensitivity and dynamic range of targeted mass spectrometry (32) can lead to detection of trace impurities of unlabeled isotopomers derived from the internal standard and misinterpretation as endogenous signal. Control runs should be performed using adequate non-HLA peptidome matrix as carrier to achieve similar signal intensities of the internal standard as in analytical runs while precluding the endogenous presence of matrix-derived unlabeled peptide. In case of detection of isotopic impurities in control runs (Fig. 6*A*), peptide validation by coelution may still be achieved by quantitative evaluation of the endogenous peptide signals normalized to internal standard and background subtraction. However, in our experience trace contamination of SIL peptides with unlabeled isotopologue was specific for single lots of SIL amino acids and was remedied by exchanging the affected lot (Fig. 6*B*). Follow-up analysis of one isotopically impure synthetic peptide product traced the isotopic impurity back to the original SIL-AA lot as opposed to later contamination during coupling to resin by a different provider: Isotopomers of the sequence were synthesized using either SIL lysine coupled to resin by a second service provider for C-terminal labeling (Fig. 6*C*, left panel) or the free SIL lysine directly obtained from the vendor for internal labeling (Fig. 6*C*, right panel). The presence of isotopic impurity in both synthetic products suggests that the trace impurity was already present on the originally obtained SIL lysine lot.

For synthetic substrate peptides used in *in vitro* assays, truncated synthesis by-products derived from incomplete couplings may also lead to misinterpretation of artifacts. One potential example is the peptide KLVVGAVGV suggested as result of proteasomal splicing of KRAS G12V (33). During our attempt to reproduce the experiment by synthesizing the precursor peptide KRAS$_{2\text{-}35}$ G12V used for *in vitro* proteasomal digestion in Mishto *et al*. (33), we observed truncated

annotation corresponding to an S-acetylated canonical peptide, coelution of SIL internal standard and endogenous signal was observed, serving as peptide identity confirmation. The lower panel summarizes MS2 fragment ion intensities of the different peptide species over retention time. The cutouts provide the underlying MS2 extracted ion chromatograms. *B* and *C*, quality control of spiked SIL-peptides for trace isotopic impurity in direct infusion MS1 QC runs. Labeled peptide was detected at signal intensities >1e7 (arbitrary units) with full isotopic envelopes displayed in *blue* (M), *purple* (M + 1), *brown* (M + 2), whereas no discernible signal was detected for the unlabeled isotopologues.

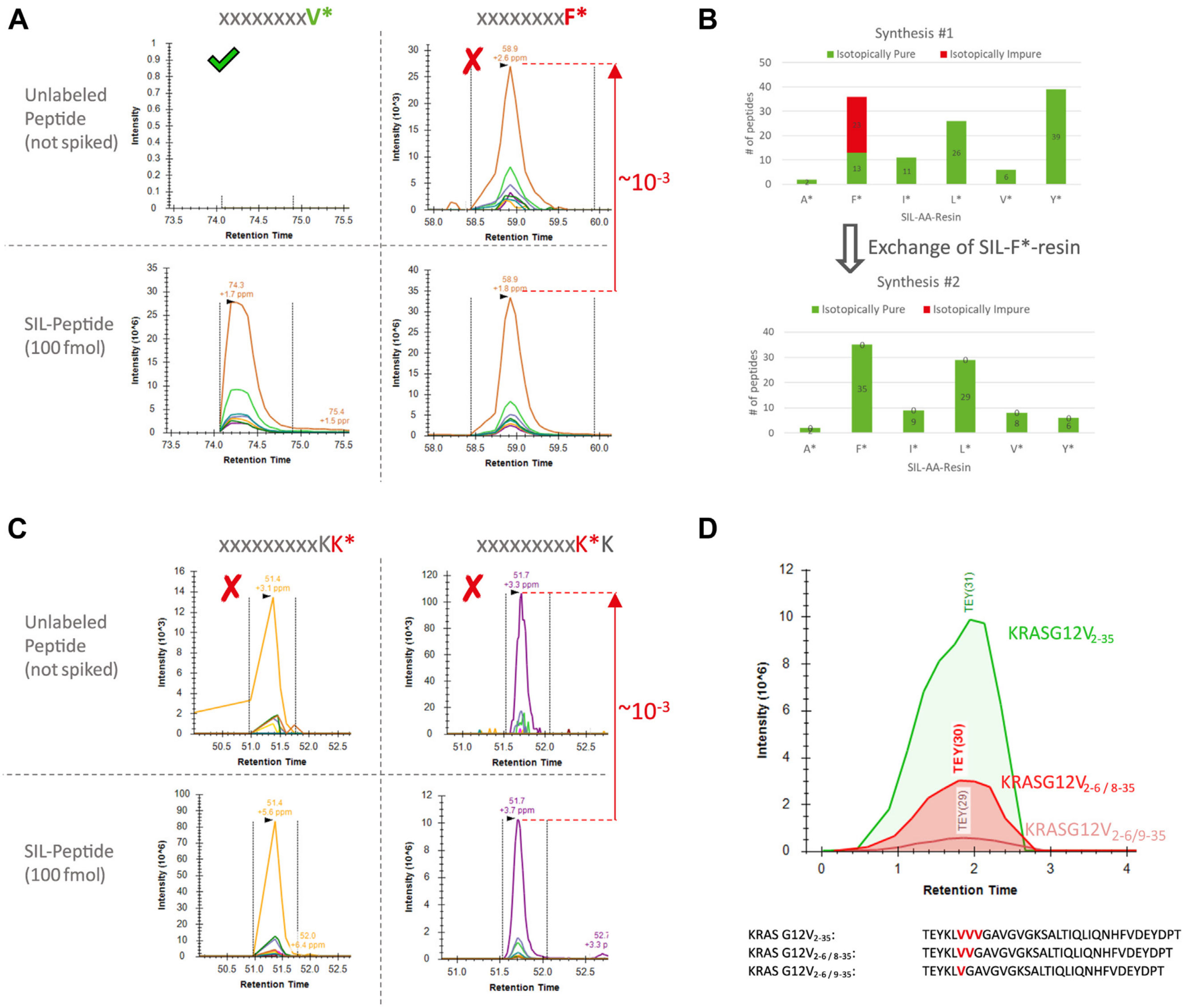


FIG. 6. **Synthetic artifacts.** *A–C*, isotopic impurities of synthetic stable isotope-labeled (SIL-) peptides. *A*, MS2 extracted ion chromatograms of two SIL peptides analyzed by targeted MS. While the SIL peptide with C-terminal labeled valine (V*, $^{13}C_5{}^{15}N$, >99%) does not show any trace signal in the unlabeled channel (*left panel*), the peptide with C-terminal labeled phenylalanine (F* $^{13}C_9{}^{15}N$, >99%) shows unlabeled signal at about 1000-fold lower intensity than in the labeled channel. *B*, comparison of two batches of SIL-peptide synthesis performed with different lots of labeled phenylalanine. The first synthesis batch (*upper panel*) shows prevalent detection of isotopic impurities for SIL-F* peptides. Synthesis of a new batch of SIL peptides with a new lot of SIL-F* showed no isotopic impurities (*lower panel*). *C*, tracing the origin of isotopic impurity for a SIL-lysine labeled peptide. Two isotopomers of the same sequence were synthesized using either C-terminal labeling (*left panel*, SIL-K* $^{13}C_6{}^{15}N_2$, >99%, loaded to resin at service provider) or internal labeling with the same SIL-K* lot directly obtained from the vendor (*right panel*). *D*, truncated by-products of KRASG12V$_{2-35}$ peptide synthesis. Shown are MS1 extracted ion chromatograms of quality control direct infusion MS for the three most abundant synthetic products. The two most prevalent by-products show incomplete coupling in an internal -VVV- sequence previously described as the site of proteasomal splicing.

versions of KRASG12V$_{2-35}$ that contained one (KRASG12V$_{2-6/9-35}$) or two (KRASG12V$_{2-6/8-35}$) instead of three consecutive valines (TEYKL**VVV**GAVGVGKSALTIQLIQNHFVDEYDPT) as the most abundant synthesis by-products (32% and 6% of the target product intensity, respectively) (Fig. 6*D*).

This observation can be explained by the decrease in coupling efficiency with growing length of the peptide (34) but also by the fact that residues with beta-branched side chains such as isoleucine, valine, or threonine induce incomplete coupling due to steric hindrance (35). Furthermore, poly-valine stretches in particular are described to result in deletion peptides (36). Thus, using this nonpurified synthesis product for *in vitro* proteasomal digests would not allow to discriminate whether the proposed peptide is derived from the incomplete coupling artifact or from a splicing mechanism.

DISCUSSION

With the increased importance of immunopeptidomics for target discovery and as training data for HLA ligand prediction, the relevance for diligent quality control grows. Here we presented common pitfalls along the immunopeptidomics pipeline that can introduce false ligands and described how we addressed this computationally and experimentally in our target discovery platform XPRESIDENT. Sample preparation might introduce proteolytic fragments mimicking HLA ligands, which can be assessed by *in silico* methods. Chromatographic procedures pose the risk of peptide carryover between samples, which can be monitored by blank runs. In addition, electrospray ionization MS generates in-source fragments that can be identified computationally. And finally, the peptide sequencing is prone to false-positive identifications that can only be uncovered by experimental sequence validation through synthetic standards.

The proposed statistical method for determining proteolytic contaminations is applicable to any larger immunopeptidomics dataset and aims to capture the hallmarks of this particular type of contamination enabling a generalizable approach for filtering. Although it has to be pointed out that the thresholds described here should be recalibrated for other datasets, this method provides a valuable metric for assessing the quality of the HLA preparation. The method incorporates three metrics that allow to differentiate peptides of proteolytic origin from other contaminations. This can guide adjustments in experimental protocols, for instance, if the underlying cause is insufficient protease inhibition. While the most predictive metric of the three is HLA-binding propensity, the method can also be applied if no ligand prediction model exists for the HLA dataset under investigation. Novel prediction models for well-studied HLAs may also benefit from this approach since any contamination filter solely relying on existing peptide binding models will limit the ability of the new model to capture previously undescribed binding characteristics. To evaluate the method, it was applied to publicly available data of 16 monoallelic cell lines. This data showed a very low number of proteolytic fragments underlining the overall good quality of the data. One exception was found for the HLA-A*68:02 transfected cell line, which had a significantly higher number of proteolytic contaminations. While this will not pose problems for determination of sequence motifs or simple prediction methods such as PSSMs, other methods might incorporate the signal of the proteolytic peptide species, in particular deep-learning networks that allow to capture less frequent patterns in the data. Yair-Sabag *et al*. (37) described HLA-B*27:05 peptides with P2-lysine anchor residues with a prevalence of 1%. Determination of such rare peptide subpopulations will be severely affected by the presence of contamination, thus removal of these peptides as described herein is highly relevant. Furthermore, due to the fact that monoallelic datasets are usually used as benchmarks, the fraction of contamination will result in substantial underestimation of prediction performance. This effect is even more pronounced in cases were test sets are depleted for peptides already used for training. While keeping test and training sets disjunct is best practice in machine learning, in case of monoallelic cell lines such filtering will remove most of the relevant allotypical peptides enriching the test set for contaminations.

Inspection of the number of proteolytic fragments indicated a tissue-dependent but otherwise constant baseline of contaminations. This means that the impact of contamination grows with reduction of the presented peptide repertoire, for instance, due to lower HLA expression or through use of allotype-specific antibodies in immunoprecipitation. The tissues that showed the largest number of contaminations were digestive organs and blood cells, in particular granulocytes. Both cases were expected due to the secreted peptidases and the expressed granular enzymes, respectively. Thus, for these samples the measurement sensitivity or downstream analyses such as normalization could be affected. Search strategies might try to overcome the limitations of the no-enzyme search protocol used for immunopeptidomics by focusing only on the patient-specific HLA allotypes (38). Yet, in such an approach contaminations might not be identified and yield false-positive identifications.

Investigation of in-source fragmentation showed that peptides are usually truncated by one or two amino acids if many positively charged residues exist in the C-terminal end. While the fraction of these peptides is generally very low, annotation of these particular events is nevertheless essential if immunopeptidomics data is used for target discovery. The HLA-A*02:01 ligand GVYDGREHTV is a known cancer target derived from MAGE-A4 (39). If the peptide is presented on cell lines with high HLA expression, the likelihood of in-source fragmentation increases and the fragments VYDGREHTV and YDGREHTV can be observed. The first fragment is considered a strong binder based on NetMHCpan (rank=0.389) and therefore could be mistaken as novel HLA-A*24:02 cancer target. *In vitro* validation experiments using T2 peptide loading would likely confirm binding and T-cell recognition as immunogenicity should theoretically be high if the peptide is never presented by HLA-A*24:02 *in vivo*. Thus, invalidation of the target would most likely only happen during efficacy screenings when a lead molecule has already been developed.

Although the described computational methods are essential in preclinical target discovery and selection, the final sequence confirmation before moving forward with peptides into target validation must be performed experimentally. Providing such validation as high-throughput method allows to work with lower stringency in target selection providing the means to accurately identify lower abundant or cryptic peptide species of noncanonical origin that may serve as shared targets for development of adoptive cell therapies, vaccines, or bispecific T cell engaging receptors. We described the necessary procedure and pointed out cases of misidentification. Any frequencies reported on novel peptide repertoires need to go through such a validation scheme in order to provide reliable estimates because even with standard FDR estimation at 1%, actual FDRs can be as high as 29% (14).

Immunopeptidomics is a field of central relevance for the development of target-specific immunotherapies and has experienced a rapid growth of publicly available datasets coinciding with an expanding scope of proposed target classes. We herein outline potential pitfalls that should be kept in mind when acquiring and analyzing such data. The computational and experimental approaches described in this manuscript allow differentiation between false and true HLA ligands. Identification of true ligands for therapeutic development is a critical step toward increasing the chances of clinical success.

DATA AVAILABILITY

The mass spectrometry data created as part of this study has been deposited at PeptideAtlas with the dataset identifier PASS01640.

Supplemental data—This article contains supplemental data.

Acknowledgments—This work was supported by the Cancer Prevention Research Institute of Texas (DP150029). Thanks to the biobanking team at Immatics for sample acquisition, the Target Research lab team for preparation of tissues and peptides, the mass spec team for peptide analysis and validation, the CMC team for peptide synthesis, and the bioinformatics team for data analysis.

Autthor contributions—J. F., D. J. K., M. P., H. S., and T. W. conceptualization; F. H. and M. R. data curation; L. B. and C.-C. T. formal analysis; M. P. investigation; L. B. methodology; S. D. and M. P. project administration; S. D. resources; L. B., C.-C. T., F. H., and M. R. software; J. F., D. J. K., H. S., O. S., and T. W. supervision; F. G. validation; F. G. visualization; J. F. and D. J. K. writing-original draft; L. B., F. G., S. D., C.-C.T., F. H., M. R., and T. W. writing-review and editing.

Conflict of interest—All authors are employees and shareholders of Immatics.

Abbreviations—The abbreviations used are: CID, collision-induced dissociation; FDR, false discovery rate; gRT, globally aligned retention time; HCD, higher collisional energy dissociation; HLA, human leukocyte antigen; IS-PRM, internal standard triggered PRM; PCC, proteolytic contamination count; PRM, parallel reaction monitoring; PSM, peptide–spectrum match; PSSM, position-specific scoring matrix; SFF, source fragmentation fraction; SIL, stable isotope label; SST, system suitability testing.

Received December 17, 2020, and in revised from, May 10, 2021
Published, MCPRO Papers in Press, June 12, 2021, https://doi.org/10.1016/j.mcpro.2021.100110

REFERENCES

1. Haen, S. P., Löffler, M. W., Rammensee, H. G., and Brossart, P. (2020) Towards new horizons: Characterization, classification and implications of the tumour antigenic repertoire. *Nat. Rev. Clin. Oncol.* **17**, 595–610
2. Falk, K., Rötzschke, O., Stevanovié, S., Jung, G., and Rammensee, H.-G. (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **351**, 290–296
3. Hunt, D., Henderson, R., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A., Appella, E., and Engelhard, V. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263
4. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanović, S. (1999) SYFPEITHI: Database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213–219
5. O'Donnell, T., and Rubinsteyn, A. (2020) High-throughput MHC I ligand prediction using MHCflurry. *Methods Mol. Biol.* **2120**, 113–127
6. Kemps, P. G., Zondag, T. C., Steenwijk, E. C., Andriessen, Q., Borst, J., Vloemans, S., Roelen, D. L., Voortman, L. M., Verdijk, R. M., van Noesel, C. J. M., Cleven, A. H. G., Hawkins, C., Lang, V., de Ru, A. H., Janssen, G. M. C., *et al*. (2019) Apparent lack of BRAF (V600E) derived HLA class I presented neoantigens Hampers neoplastic cell targeting by CD8(+) T cells in Langerhans cell histiocytosis. *Front. Immunol.* **10**, 3045
7. Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., Modrusan, Z., Mellman, I., Lill, J. R., and Delamarre, L. (2014) Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576
8. Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H. C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., *et al*. (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262
9. Kim, J. S., Monroe, M. E., Camp, D. G., Smith, R. D., and Qian, W. J. (2013) In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *J. Proteome Res.* **12**, 910–916
10. Bian, Y., Zheng, R., Bayer, F. P., Wong, C., Chang, Y. C., Meng, C., Zolg, D. P., Reinecke, M., Zecha, J., Wiechmann, S., Heinzlmeir, S., Scherr, J., Hemmer, B., Baynham, M., Gingras, A. C., *et al*. (2020) Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC-MS/MS. *Nat. Commun.* **11**, 157
11. Bache, N., Geyer, P. E., Bekker-Jensen, D. B., Hoerning, O., Falkenby, L., Treit, P. V., Doll, S., Paron, I., Müller, J. B., Meier, F., Olsen, J. V., Vorm, O., and Mann, M. (2018) A novel LC system Embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell Proteomics* **17**, 2284–2296
12. Bohley, P., and Seglen, P. O. (1992) Proteases and proteolysis in the lysosome. *Experientia* **48**, 151–157
13. Backert, L. (2018). In: *Applied Immunoinformatics: HLA Peptidome Analysis for Cancer Immunotherapy, Ph.D. thesis*, Eberhard Karls Universität
14. Zhang, M., Fritsche, J., Roszik, J., Williams, L. J., Peng, X., Chiu, Y., Tsou, C. C., Hoffgaard, F., Goldfinger, V., Schoor, O., Talukder, A., Forget, M. A., Haymaker, C., Bernatchez, C., Han, L., *et al*. (2018) RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat. Commun.* **9**, 3919
15. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass

spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326

16. R Core Team. (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria
17. Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006) Two sample logo: A graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536–1537
18. Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., and Nesvizhskii, A. I. (2015) DIA-umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264. 7 p following 264
19. Muller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017) Hotspots' of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front. Immunol.* **8**, 1367
20. 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
21. Nijman, H. W., Van der Burg, S. H., Vierboom, M. P., Houbiers, J. G., Kast, W. M., and Melief, C. J. (1994) p53, a potential target for tumor-directed T cells. *Immunol. Lett.* **40**, 171–178
22. Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008) HLA class I supertypes: A revised and updated classification. *BMC. Immunol.* **9**, 1
23. Gnjatic, S., Bressac-de Paillerets, B., Guillet, J. G., and Choppin, J. (1995) Mapping and ranking of potential cytotoxic T epitopes in the p53 protein: Effect of mutations and polymorphism on peptide binding to purified and refolded HLA molecules. *Eur. J. Immunol.* **25**, 1638–1642
24. Jensen, S. M., Potts, G. K., Ready, D. B., and Patterson, M. J. (2018) Specific MHC-I peptides are induced using PROTACs. *Front. Immunol.* **9**, 2697
25. Bleiholder, C., Suhai, S., Harrison, A. G., and Paizs, B. (2011) Towards understanding the tandem mass spectra of protonated oligopeptides. 2: The proline effect in collision-induced dissociation of protonated Ala-Ala-Xxx-Pro-Ala (Xxx = Ala, Ser, Leu, Val, Phe, and Trp). *J. Am. Soc. Mass Spectrom.* **22**, 1032–1039
26. Nesvizhskii, A. I. (2014) Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125
27. 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
28. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015) Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell Proteomics* **14**, 658–673
29. Vaca Jacome, A. S., Rabilloud, T., Schaeffer-Reiss, C., Rompais, M., Ayoub, D., Lane, L., Bairoch, A., Van Dorsselaer, A., and Carapito, C. (2015) N-terminome analysis of the human mitochondrial proteome. *Proteomics* **15**, 2519–2524
30. Erhard, F., Dölken, L., Schilling, B., and Schlosser, A. (2020) Identification of the cryptic HLA-I immunopeptidome. *Cancer Immunol. Res.* **8**, 1018–1026
31. Gallien, S., Kim, S. Y., and Domon, B. (2015) Large-scale targeted proteomics using internal standard triggered-parallel reaction monitoring (IS-PRM). *Mol. Cell Proteomics* **14**, 1630–1644
32. Liebler, D. C., and Zimmerman, L. J. (2013) Targeted quantitation of proteins by mass spectrometry. *Biochemistry* **52**, 3797–3806
33. Mishto, M., Mansurkhodzhaev, A., Ying, G., Bitra, A., Cordfunke, R. A., Henze, S., Paul, D., Sidney, J., Urlaub, H., Neefjes, J., Sette, A., Zajonc, D. M., and Liepe, J. (2019) An in silico-in vitro pipeline identifying an HLA-A(*)02:01(+) KRAS G12V(+) spliced epitope candidate for a broad tumor-immune response in cancer patients. *Front. Immunol.* **10**, 2572
34. Young, J. D., Huang, A. S., Ariel, N., Bruins, J. B., Ng, D., and Stevens, R. L. (1990 Jul-Aug) Coupling efficiencies of amino acids in the solid phase synthesis of peptides. *Pept. Res.* **3**, 194–200
35. Milton, R. C.d. L., Milton, S. C. F., and Adams, P. A. (1990) Prediction of difficult sequences in solid-phase peptide synthesis. *J. Am. Chem. Soc.* **112**, 6039–6046
36. Larsen, B. D., and Holm, A. (1994) Incomplete Fmoc deprotection in solid-phase synthesis of peptides. *Int. J. Pept. Protein Res.* **43**, 1–9
37. Yair-Sabag, S., Tedeschi, V., Vitulano, C., Barnea, E., Glaser, F., Melamed Kadosh, D., Taurog, J. D., Fiorillo, M. T., Sorrentino, R., and Admon, A. (2018) The peptide repertoire of HLA-B27 may include ligands with lysine at P2 anchor position. *Proteomics* **18**, e1700249
38. Murphy, J. P., Konda, P., Kowalewski, D. J., Schuster, H., Clements, D., Kim, Y., Cohen, A. M., Sharif, T., Nielsen, M., Stevanovic, S., Lee, P. W., and Gujar, S. (2017) MHC-I ligand discovery using targeted database searches of mass spectrometry data: Implications for T-cell immunotherapies. *J. Proteome Res.* **16**, 1806–1816
39. Duffour, M. T., Chaux, P., Lurquin, C., Cornelis, G., Boon, T., and van der Bruggen, P. (1999) A MAGE-A4 peptide presented by HLA-A2 is recognized by cytolytic T lymphocytes. *Eur. J. Immunol.* **29**, 3329–3337