Navigating Critical Challenges Associated with Immunopeptidomics-Based Detection of Proteasomal Spliced Peptide Candidates

Cheryl F. Lichti[1,2], Nathalie Vigneron[3,4], Karl R. Clauser[5], Benoit J. Van den Eynde[3,4,6], and Michal Bassani-Sternberg[7,8]

ABSTRACT

Within the tumor immunology community, the topic of proteasomal spliced peptides (PSP) has generated a great deal of controversy. In the earliest reports, careful biological validation led to the conclusion that proteasome-catalyzed peptide splicing was a rare event. To date, six PSPs have been validated biologically. However, the advent of algorithms to identify candidate PSPs in mass spectrometry data challenged this notion, with several studies concluding that the frequency of spliced peptides binding to MHC class I was quite high. Since this time, much debate has centered around the methodologies used in these studies. Several reanalyses of data from these studies have led to questions about the validity of the conclusions. Furthermore, the biological and technical validation that should be necessary for verifying PSP assignments was often lacking. It has been suggested therefore that the research community should unite around a common set of standards for validating candidate PSPs. In this review, we propose and highlight the necessary steps for validation of proteasomal splicing at both the mass spectrometry and biological levels. We hope that these guidelines will serve as a foundation for critical assessment of results from proteasomal splicing studies.

Introduction to Proteasomal Spliced Peptides

Initial discovery of antigenic peptides recognized by CTL and produced by peptide splicing in the proteasome

Peptide splicing was initially described in the early 2000s when Hanada and colleagues identified the peptide recognized by a $CD8^+$ CTL clone isolated from a patient with renal cell carcinoma (1). This CTL clone was able to kill cancer cells and was found to recognize a peptide, NTYAS_PRFK, composed of two noncontiguous fragments of the FGF-5 protein. Production of that peptide required the removal of a 40-amino-acid intervening sequence and the creation of a new peptide bond between the amino acids located at either end. The occurrence of peptide splicing was initially suggested by the fact that, upon transfection of an HLA-$A3^+$ line with a subgenic fragment of FGF-5, presentation of the antigenic peptide could occur if the sequence was truncated in its middle part, but it was lost if the sequence was shortened from the 5′ or 3′ ends. Moreover, mutation-induced modification of the amino acids located at both ends of the construct also impeded production of the antigenic peptide. The existence of the spliced peptide at the surface of HLA-$A3^+$FGF-5^+ cells was finally demonstrated by showing that, after separating the peptide eluate by high-performance liquid chromatography (HPLC), the fractions recognized by the CTL were identical to those recognized when the synthetic spliced peptide was fractionated on the HPLC in the exact same conditions.

A few months later, the mechanism of peptide splicing was elucidated during the identification of a spliced peptide recognized by a tumor-killing CTL clone isolated from a patient with melanoma (2). This second spliced peptide, RTK_QLYPEW, was composed of two fragments originating from the melanoma differentiation protein gp100 and spliced together after removal of a 4-amino-acid intervening sequence. Here also, the peptide was identified using an approach based on a genetic screening and transfection of subgenic fragments encoding the antigen. Electroporation of EBV-B cells with the precursor peptide RTKAWNRQLYPEW bearing alanine substitutions showed that the amino acids composing the final antigenic peptide were located at both ends of that peptide precursor. Here again, HPLC fractions obtained from melanoma peptide eluates confirmed the presence of the proteasomal spliced peptide (PSP) at the surface of melanoma cells. The role of the proteasome in peptide splicing was initially suggested by using proteasome inhibitors. Further work showed that the spliced peptide RTK_QLYPEW could be produced *in vitro* by incubating purified proteasomes with RTKAWNRQLYPEW. By incubating pairs of peptides containing portions of RTKAWNRQLYPEW with purified proteasomes, it was shown that peptide splicing involved a transpeptidation reaction via an acyl-enzyme intermediate between fragment RTK and the proteasome (**Fig. 1**). During the splicing reaction, the ester link of this acyl-enzyme intermediate is subjected to nucleophilic attack by the free amino group of QLYPEW. This model was supported by the fact that N-α-acetylation of QLYPEW completely prevented production of the spliced peptide when incubated with peptide RTKAWNR and the proteasome. Likewise, the spliced peptide NTYAS_PRFK initially described by Hanada and colleagues was subsequently shown to be produced in the proteasome by transpeptidation (3).

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri. [2]Bursky Center for Human Immunology and Immunotherapy, Washington University School of Medicine, St. Louis, Missouri. [3]Ludwig Institute for Cancer Research, Brussels, Belgium. [4]de Duve Institute, Université Catholique de Louvain, Brussels, Belgium. [5]Broad Institute of MIT and Harvard, Cambridge, Massachusetts. [6]Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom. [7]Ludwig Institute for Cancer Research, Lausanne Branch—University of Lausanne (UNIL), Lausanne, Switzerland. [8]Department of Oncology—Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland.

Corresponding Authors: Cheryl F. Lichti, 660 S. Euclid Avenue, Campus Box 8118, St. Louis, MO 63110. Phone: 314-273-1736; E-mail: clichti@wustl.edu; and Michal Bassani-Sternberg, Rue du Bugnon 25A CH-1011 Lausanne, Switzerland. Phone: 41 (21) 3148502; E-mail: michal.bassani@chuv.ch

Cancer Immunol Res 2022;10:275–84

doi: 10.1158/2326-6066.CIR-21-0727

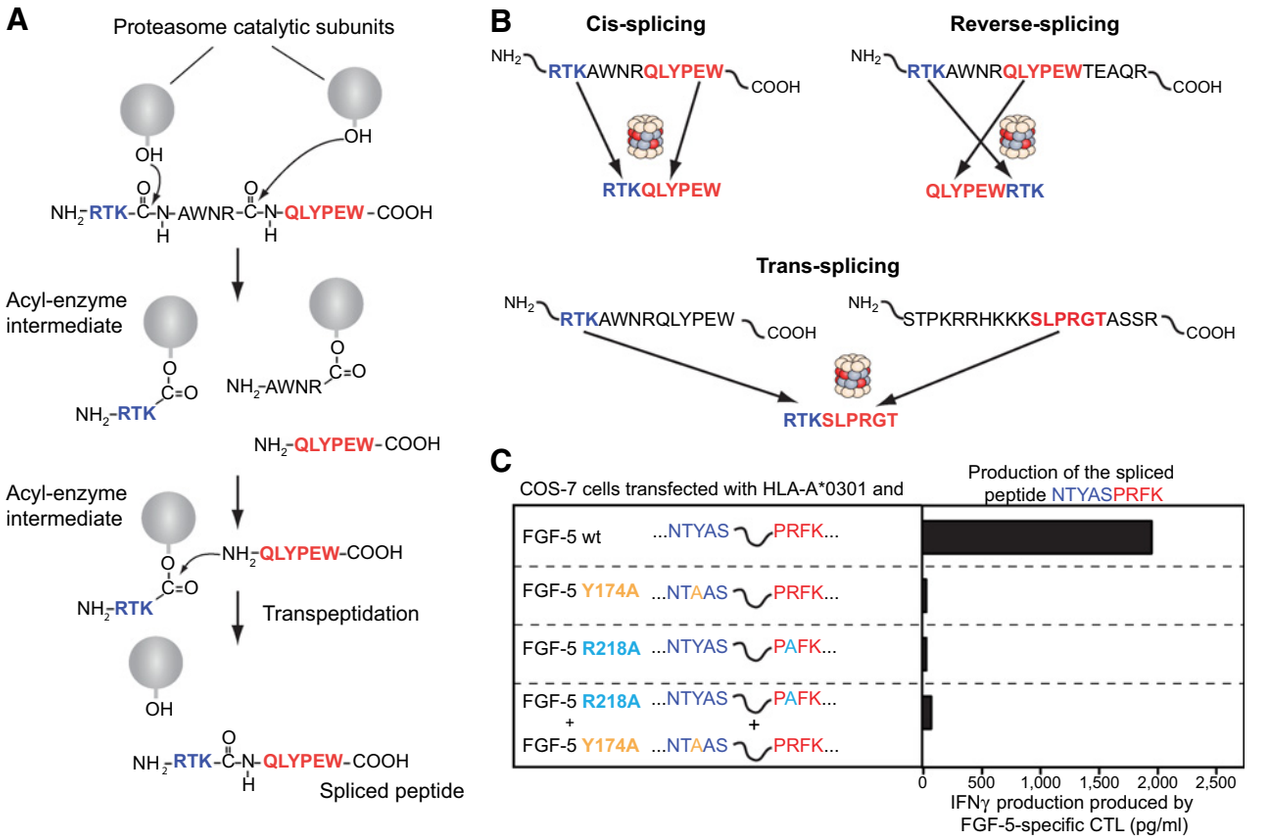

Figure 1.

Peptide splicing in the proteasome occurs by transpeptidation. **A,** Splicing of the gp100-derived peptide RTK_QLYPEW is shown. Upon formation of an acyl-enzyme intermediate involving the N-terminal splicing partner RTK, the amino group of the C-terminal splicing partner, QLYPEW, produces a nucleophilic attack on the acyl-enzyme intermediate to create a new peptide RTK_QLYPEW composed of two fragments originally distant in the protein. **B,** Schema for *cis*-splicing, reverse *cis*-splicing, and *trans*-splicing. **C,** *Trans*-splicing by the proteasome does not occur at a significant level in physiologic conditions (data adapted from ref. 3). To study the physiologic relevance of *trans*-splicing, COS-7 cells were transfected with plasmids encoding HLA-A*0301 and pairs of FGF-5 constructs designed so that production of the antigenic peptide can only occur following splicing of peptide fragments originating from two different proteins. Transfected cells are then tested for their ability to induce IFNγ production by the FGF-5–specific CTL-C2. The amount of IFNγ produced by the CTL after incubation with the transfected cells is measured by ELISA.

The third example of peptide splicing came a few years later when a CTL clone, isolated from the recipient of an MHC-matched allogeneic hematopoietic cell transplant, was shown to recognize a peptide encompassing the SNP A996G in the gene coding the SP110 nuclear protein (4). The CTL recognized cell lines displaying a $SP110_{A996}$ allele but not those displaying only the $SP110_{G996}$ allele. Further analysis revealed that the peptide recognized by the CTL was composed of two fragments derived from SP110 (STPK and SLPRGT), which were assembled in the reverse order to that in which they occur in the parental protein, to form the spliced peptide SLPRGT_STPK. Residue R_{299}, which is encoded by the polymorphism, is located in position 4 of the peptide, and the corresponding peptide SLPGGT_STPK encoded in the recipient cells was not recognized by the CTL. Here again, the peptide could be found in peptide eluates using the CTL and was shown to be produced by transpeptidation in the proteasome.

The fourth spliced peptide was identified as the target of a clone of tumor-infiltrating lymphocytes (TIL) isolated from a patient with melanoma (5). These TILs were used for adoptive T-cell transfer and shown to induce dramatic tumor rejection in the patient (6). This highlights the biological and clinical relevance of PSPs. This peptide is composed of two noncontiguous fragments of the tyrosinase protein spliced in the reverse order to that in which they occur in the parental protein. In addition, it contains two aspartate residues originating from deglycosylation of genetically encoded, N-glycosylated asparagine residues. Such deamidation was previously observed in antigenic peptides derived from glycosylated proteins and results from the action of N-glycanase to remove the sugar moiety after retrotranslocation of the glycoprotein into the cytosol and before degradation by the proteasome. Here, the reverse splicing of deamidated fragments was shown to occur in the proteasome by transpeptidation.

Finally, a fifth spliced peptide RSYVPLAH_R, derived from gp100, was shown to be the target of a TIL clone isolated from a patient with melanoma (7). In contrast to the other examples of spliced peptides, which were composed of three to six amino acid fragments, the peptide recognized by this TIL contained an N-terminal splicing partner of eight amino acids, to which a single arginine residue was added by transpeptidation. *In vitro* proteasome digestion experiments using pairs of peptides showed that the peptide causing the nucleophilic attack on the acyl-enzyme intermediate must be at least three amino acids long for the splicing reaction to take place (7). In the case of

RSYVPLAH_R, a C-terminally extended spliced peptide is first produced and then further trimmed by the proteasome to form the final antigenic peptide.

All these validated peptides are produced by splicing two fragments from the same protein, a process termed "*cis*-splicing." It is unclear at this stage whether the proteasome can also splice fragments from distinct proteins, a process referred to as "*trans*-splicing." Although *trans*-splicing may occur during *in vitro* digestion, whether it also occurs in a biologically meaningful fashion within cells is an open question. *Trans*-splicing was investigated by cellular and molecular assays using the CTL recognizing the spliced peptide NTYAS_PRFK derived from FGF-5. Plasmids were designed to encode full-length FGF-5 constructs bearing mutations at critical residues of either of the fragments NTYAS or PRFK, so that production of the spliced peptide NTYAS_PRFK recognized by the CTL was only possible if splicing occurred from two distinct proteins (3). Cells transfected with these constructs were tested for recognition by the CTL, and results showed that *trans*-splicing hardly occurred in these conditions (**Fig. 1**). Similar results were obtained with spliced peptide RTK_QLYPEW. Considering that the very low signal observed was likely due to overexpression of the transfected constructs in COS-7 cells, it was concluded that *trans*-splicing was unlikely to occur in physiologic conditions. One reason for this is that two distinct proteins might not be able to simultaneously access the catalytic chamber of the proteasome, and the likelihood of having two specific protein substrates degraded repeatedly at the same time is extremely low.

Mass spectrometry and first large-scale spliced peptide identifications by immunopeptidomics

In the studies discussed so far, identification of PSPs relied on the isolation of a CTL clone whose target antigen was subsequently identified as a PSP. Production of PSPs, following incubation of a precursor peptide with purified proteasome, was then verified using mass spectrometry (MS). After the initial identification of PSPs, identification of novel PSPs was attempted by systematic MS analysis of proteasome digests using databases containing all possible spliced products that could be generated from a given linear precursor (8). However, using this approach, only one PSP was fully validated by isolating a T cell from the peripheral blood mononuclear cells (PBMC) of an HLA-A0301$^+$ healthy donor and demonstrating that this T-cell clone was able to recognize and kill gp100$^+$HLA-A0301$^+$ melanoma cells (9). A similar approach was later used for identifying spliced $KRAS_{G12V}$ peptides (10). However, although one of the *in vitro*–generated PSPs identified (KL_VVGAVGV) was shown to bind to HLA-A2, this study showed no evidence that any of the peptides identified were presented naturally by tumor cells. In a parallel study, Willimsky and colleagues cloned a T-cell receptor (TCR) able to recognize the peptide KL_VVGAVGV by immunizing mice harboring the human TCR$\alpha\beta$ coding genes with this peptide (11). This TCR was transduced into PBMC, which could then recognize target cells pulsed with the KL_VVGAVGV peptide but not HLA-A2$^+$ tumor cells endogenously expressing mutant $KRAS_{G12V}$ or overexpressing a mutant $KRAS_{G12V}$ construct, supporting a lack of natural presentation of the spliced peptide in cells. This confirmed that the above combination of *in vitro* and *in silico* approaches for identification of peptides, and in particular PSPs, is not sufficient to demonstrate their relevance in cancer immunotherapy.

With the improvement of MS technologies and supportive bioinformatics tools, the large-scale characterization of naturally presented HLA ligands, a method called immunopeptidomics, has become feasible and straightforward. With this method, HLA complexes are immunoaffinity purified from cells or tissues, and the HLA-bound peptides are isolated and subsequently sequenced by LC/MS-MS. The peptide MS/MS spectra are primarily interpreted using a database search algorithm that matches and scores the similarity of each experimental spectrum against model spectra constructed from the candidate peptide sequences contained in a protein sequence database, typically the human reference proteome, preferably augmented with personalized sequences to account for genetic polymorphisms. This strategy is used, in part, because the fragmentation efficiency of current MS-MS instrumentation is unable to consistently yield spectra from which complete, unambiguous sequences can be interpreted *de novo*. However, leveraging this approach for the detection of PSPs was not straightforward, because it required the scaling up of the look-up reference of candidate sequences to include all anticipated spliced products. Database size inflation subsequently compromises typical false discovery rate (FDR) calculations and can present computational capacity requirements well beyond that of most laboratories.

In 2016, Liepe and colleagues used an immunopeptidomics approach to identify HLA ligands produced by peptide splicing (12). To do so, they created a custom search database that included all potential short peptides produced by the splicing of noncontiguous protein fragments. To restrict the size of the database, they only considered spliced peptides originating from the *cis*-splicing of peptide fragments separated by 25 amino acids or less. Nevertheless, this resulted in an enormous reference database that was 100× larger than the typical human reference proteome. Using this approach, they initially estimated that spliced peptides comprised about 23%–33% of the HLA class I immunopeptidome (12, 13) not only in their own dataset, but also in the previously published dataset of Bassani-Sternberg and colleagues (14). This raised many controversies, as the properties of the reported spliced HLA peptides were drastically different from those of the linear ligands detected within the same samples (12). Using identical data but different computational approaches, others have estimated the percentage of spliced peptides in the HLA class I immunopeptidome to be at most 2%–6% (15) or even less than 0.1% (16).

Two years later, Faridi and colleagues used a hierarchical, *de novo* sequencing–driven data interpretation approach to identify PSPs. First, they searched against a canonical protein database to assign spectra to genomically templated sequences. Then, for unassigned MS-MS spectra with high confidence *de novo* sequence assignments, they searched against a canonical protein database for sequence fragments within the same protein that, after *cis*-splicing, would explain the *de novo* sequence. If no *cis*-splicing event was found, then *trans*-splicing was interrogated (17). They reported that *trans*-splicing could account for more than 25% of the peptides identified not only in eight monoallelic datasets of their own, but also in nine previously published monoallelic datasets from Abelin and colleagues (18). These findings were controversial for several reasons, both at the biologic and bioinformatic levels. One reason, as previously noted, is that *trans*-splicing does not occur readily in cells (3). Moreover, given the abundance and nucleophilicity of water, hydrolysis in the enzyme active site is more likely than splicing, and it has been argued that PSPs are likely extremely rare for this reason (19). The lack of biological evidence supporting the unexpected high occurrence of *trans*-splicing proposed in this study led to intense debate about the identity of many of these peptides at the MS level (20, 21).

To understand one part of the intense debate, it is instructive to consider the likelihood of finding a random, hypothetical pair of *trans*-splicing donor proteins in the human reference proteome. Let us

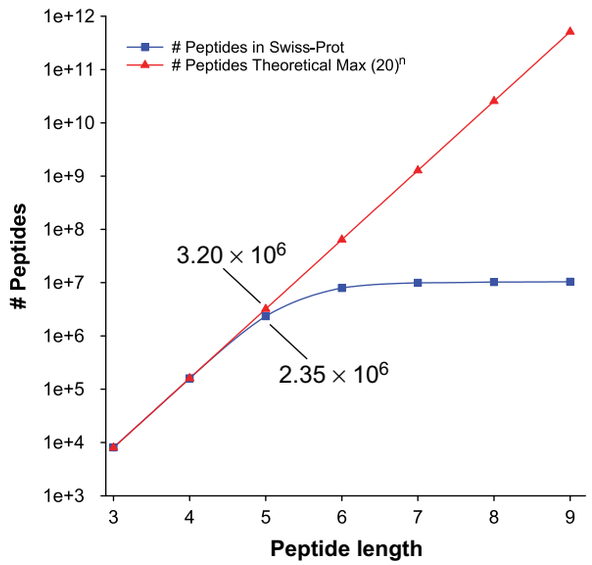

Figure 2.

Probability of finding *n*-mers. The number of unique sequences of length *n* found in Swiss-Prot (20,381 proteins, March 31, 2021) plotted along with the theoretical maximum number of sequences of length *n* (20^n). A total of 73.4% of all possible sequences of length 5 and 99.9% of all sequences of length 4 are present. Hence, with a length 9 *de novo*–derived peptide sequence of unknown origin, it would typically succeed in finding a random hypothetical pair of proteins that could donate a 4-mer and a 5-mer for *trans*-splicing. If one were to more faithfully approximate an MS-MS sequenced 9-mer, where leucine and isoleucine cannot be distinguished, then the probability of success would rise as the number of all possible 5-mers shrinks (19^5). The chance of finding a matching peptide would further increase if peptide fragmentation during MS-MS was incomplete (which is usual), leading to subsequent imperfect *de novo* interpretation of ions for which two or more amino acids were lacking, leading to multiple 9-mer sequences per spectrum to test.

consider a 9-mer PSP, formed by splicing together fragments of 4 and 5 amino acids in length derived from any protein in the human reference proteome. All possible 4-mers (20^4) and 73% of all 5-mers (20^5) can be found at least once in the human reference proteome (**Fig. 2**). Therefore, with such high random chance of success in finding hypothetical source proteins that could donate a 4-mer and a 5-mer for *trans*-splicing juxtaposed to the conceptually rare chance of coprocessing at the proteasome and combined with lack of molecular validation, we propose that these left-over, high-confidence *de novo*–derived peptide sequences should be categorized as peptides of UNKnown origin (pUNK), rather than inferred to be generated by random *trans*-splicing.

It has been shown that, beyond canonical peptides altered by phenomena such as single-nucleotide variation, somatic mutation, and posttranslational modification (PTM), HLA-bound peptides can be derived from sources other than protein-coding regions (22). These "noncanonical" antigens (also called cryptic, alternative, or dark-matter antigens), can originate from alterations at the genomic, epigenomic, transcriptomic, translational, and proteomic levels. For example, alternative splicing, intronic retention, RNA editing, non-canonical translation initiation, and codon read-through, are all reported to generate noncanonical tumor antigens (23–28). Ouspenskaia and colleagues further demonstrated that, although many of the peptide sequences reported as spliced peptides by Faridi and colleagues were correct MS-MS interpretations, they could be accounted for as linear peptides derived from novel human open-reading frames (ORF) whose translation was supported by Riboseq (27). Such sources were annotated as noncoding and thus absent from the reference proteome considered by Faridi and colleagues. Although misidentification of peptides in MS analyses can be rooted in the use of incomplete or nonpersonalized references, including all possible variants of canonical and noncanonical sources and common PTMs would prohibitively inflate the reference database, and as discussed above, compromise typical FDR calculations. Because of these inherent challenges, it is crucial to supplement the standard analysis with bioinformatic and MS-based validation of peptide sequences as described in the following sections. These approaches are equally valid for peptides derived from canonical and noncanonical sources.

Bioinformatics and Spliced Peptide Assignments: Methods for Evaluating Overall Dataset Quality

Given the controversy surrounding the identification of possible PSPs by custom bioinformatics algorithms, it is instructive to discuss the evaluation of peptide-spectrum matches (PSM) assigned by search engines and custom algorithms. In all cases, PSMs can be evaluated on the basis of the same basic criteria: chromatographic retention time predictions, precursor mass error distributions, MS-MS spectral quality, and agreement between search tools. These features should be similar or within the same range for PSPs and linear peptides. However, additional criteria are possible for HLA peptides. Because HLA peptide presentation occurs after proteasomal processing, the overall characteristics of the population of linear peptides and PSPs should be similar, including HLA allele anchor motifs, HLA-binding predictions and the frequency of cysteine-containing peptides. In that vein, several methods can be used as relative quality metrics to raise red flags about the overall integrity of peptide identification in published datasets reporting PSPs.

Chromatographic retention time prediction

During LC/MS-MS, HLA peptides are separated by reversed phase liquid chromatography prior to ionization and MS-MS sequencing. Hydrophilic peptides elute first; gradually, the more hydrophobic peptides elute as the organic solvent concentration increases. Recent applications of deep-learning approaches have improved the accuracy of retention time prediction with (29) and without inclusion of internal standards (30). Correlation between the calculated peptide hydrophobicity index and its measured chromatographic retention time (31) could be an orthogonal parameter for validation of peptide identification (20, 32). Such analysis may be sufficient to eliminate some peptides likely to be wrongly identified as PSPs.

Precursor mass error distributions

When evaluating spectral assignments for proposed PSPs, the precursor (MS1) mass error provides an excellent quality metric. This value, in parts per million (ppm), is determined by subtracting the theoretical precursor mass from the observed precursor mass value, dividing by the theoretical mass, and multiplying by 1×10^6. The magnitude of the mass error is dependent on both instrument model and certain acquisition parameters; generally, it is less than 20 ppm. A parent mass error histogram for all identified peptides in an LC/MS-MS run should show a Gaussian distribution centered near zero. As instrument calibration deteriorates, the

center of this histogram drifts and the distribution becomes broader and less Gaussian.

For a well-calibrated system, 99.7% of peptide assignments fall within three SDs of the standard mass error (33). Mass errors that fall outside this range usually indicate incorrect assignments. Thus, if the parent mass error distribution is substantially different for proposed PSPs than for linear peptides in the same dataset, it is a strong indicator that the spectral assignments for PSPs are not correct.

This point is illustrated in **Fig. 3A**, with a stacked bar histogram illustrating mass error distributions at the PSM level for an HLA-A*01:01 datafile from Faridi and colleagues (17). The dotted lines, indicating 3× SD, can serve as approximate guidelines for PSM evaluation. As can be seen from the stacked bars, PSMs outside the indicated range are dominated by peptides assigned as *trans*-spliced, indicating that these assignments are most likely incorrect. **Figure 3B**, a stacked bar chart that presents the data from **Fig. 3A** by percentage of peptide type within each bin, illustrates this point more clearly. As mass error increases, a higher percentage of the PSMs belong to PSPs and are likely incorrect.

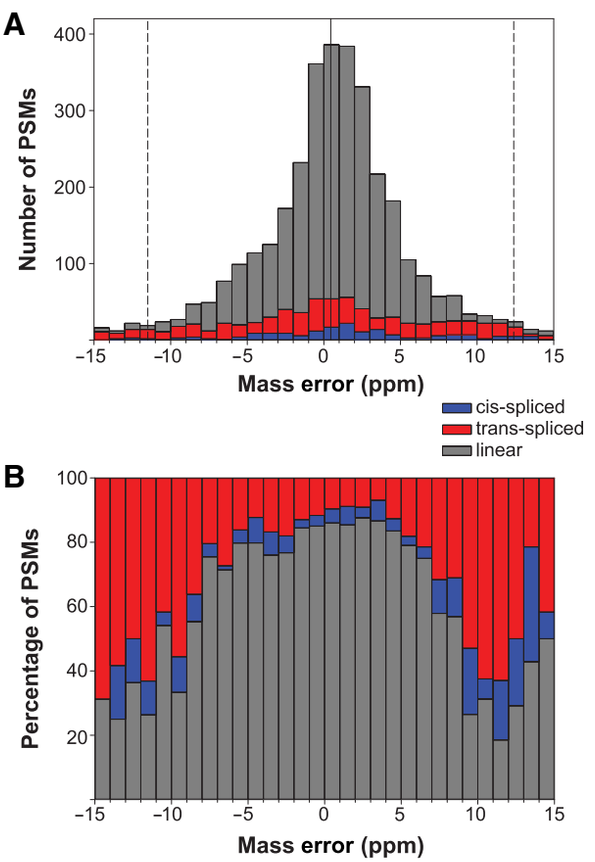

Figure 3.

Precursor mass error distributions. **A,** Stacked bar histogram illustrating mass error distributions at the PSM level for an HLA-A*01:01 datafile from Faridi and colleagues (17). The dotted lines illustrate ± 3× SDs from the mean, calculated from the PSMs for linear peptides. Peptides are grouped into 1 ppm bins. **B,** Stacked bar chart that presents the data from **A**, plotted as the percentage of linear, *cis*-spliced or *trans*-spliced peptides in each histogram bin. This plot highlights the fact that most peptides with high mass error are PSPs and are likely incorrect.

MS-MS spectral fragmentation and *de novo* sequence ambiguity

Once it has been determined that the parent mass errors for proposed PSPs are within a reasonable range, each MS-MS spectrum can be critically evaluated to determine the quality of the match to the proposed peptide sequence. In collision-induced dissociation and high-energy collision-induced dissociation fragmentation of peptides, amide bonds fragment in a predictable manner to generate

Table 1. A representative list of isobaric amino acid residues, modified amino acid residues, and chemical modifications and PTMs.

I. Isobaric amino acid(s)/modified amino acid(s)	
Da	**Amino acid(s)/modifications**
57.0215	Gly, carbamidomethyl group
114.043	Asn, Gly-Gly
115.027	Asp, (formyl)Ser
128.059	Gln, Ala-Gly
129.043	Glu, (acetyl)Ser
158.069	Ala-Ser, Gly-Thr
160.031	(carbamidomethyl)Cys, Cys-Gly
170.106	Ala-Val, Gly-Leu/Ile
171.064	Asn-Gly, Gly-Gly-Gly
185.08	Ala-Asn, Gln-Gly, Ala-Gly-Gly
199.096	Ala-Gln, Ala-Ala-Gly
200.116	Leu/Ile-Ser, Thr-Val
201.075	Asn-Ser, Gly-Gly-Ser
202.078	Cys-Thr, Gly-Met(oxidized)
211.096	Asn-Pro, Gly-Gly-Pro
213.111	Asn-Val, Ala-Ala-Ala, Gly-Gly-Val
215.091	Asn-Thr, Gln-Ser, Ala-Gly-Ser, Gly-Gly-Thr
216.075	Asp-Thr, Glu-Ser
217.052	Asn-Cys, Gly-Cys(carbamidomethyl), Cys-Gly-Gly
218.073	Ala-Met(oxidized), Met-Ser
225.111	Gln-Pro, Ala-Gly-Pro
227.127	Asn-Leu/Ile, Gln-Val, Ala-Gly-Val, Gly-Gly-Leu/Ile
228.086	Asn-Asn, Asn-Gly-Gly
228.111	Asp-Leu/Ile, Glu-Val
229.07	Asn-Asp, Asp-Gly-Gly
229.106	Gln-Thr, Ala-Ala-Ser, Ala-Gly-Thr
II. Important PTMs to include	
15.99	Oxidation (Met)
0.984	Deamidation (Asn, Gln)
42.011	N-terminal acetylation (all)
119.004	Cysteinylation (Cys)
57.022	Cys(carbamidomethyl) (if alkylation with iodoacetamide was performed)
III. Important PTMs to consider	
57.022	N-terminal carbamidomethylation (all)
27.995	N-terminal formylation (all)
21.982	Sodium adduct (Asp, Glu)
−18.011	Dehydration (Ser, Thr)
0.984	Citrullination (Arg)
47.985	Cysteine oxidation to cysteic acid
79.966	Phosphorylation (Ser, Thr, Tyr)

Note: Isobaric amino acids/groups of amino acids/modified residues can confound spectral assignments during database searching, especially when incomplete fragmentation occurs. Chemical modifications and PTMs can also confound assignments and should be considered when confirming PSMs for PSPs or other noncanonical peptides. A partial list of these species, including chemical modifications that commonly lead to erroneous PSP assignments, is included.

two primary series of ions: b ions, which correspond to fragment sequences that include the N-terminus, and y ions, which include the C-terminus. The mass difference between fragment ions in each series is indicative of the amino acids and their order, which can be used to obtain all, or part, of a peptide sequence. In the best-case scenario, a complete set of complementary b and y ions would be present to facilitate spectral assignments. In reality, this rarely happens. Instances often arise when fragment ions are absent. The resulting mass gap will support two or more amino acids, and the order of the amino acids in the sequence cannot be determined from the spectrum. To further complicate things, certain combinations of amino acids are isobaric (have the same mass), and modified forms of certain amino acids can be isobaric with others that are unmodified (**Table 1**).

For *de novo* interpretation, it is common to give a score for each individual amino acid interpreted as well as an overall score. For the widely used *de novo* program PEAKS, the local confidence (LC) score indicates the certainty of individual amino acids. A minimum LC of 80 (15) or average LC (ALC) score of approximately 80 (17) has been used as a threshold for exporting high-quality sequences. In **Fig. 4**, the LC scores for the proposed sequence ELC_DKEWVAK indicate that the DKEWVAK portion of the sequence is strongly supported. In contrast, the ELC portion is so weakly supported that any order of the three amino acids is possible, as would any other combination of amino acids with the same mass. The resulting sequence ambiguity for representing a *de novo* interpretation is accomplished by reporting multiple sequences per spectrum, employing a shorthand via regular expression syntax, or replacing the ambiguous sequence with a mass gap representation [345]DKEWVAK (34).

When one consults Supplementary Table S4 from Faridi and colleagues (17), reporting sequences identified via a *de novo* approach, there is no indication of the sequence ambiguity of the individual peptides reported, despite employing ALC thresholds that would generate significant ambiguity. If one calculates the theoretical precursor mass of each sequence in the table and then sorts by mass, it is apparent that 152/560 (27%) of the PSPs for HLA-A*03:01 exhibit sets with two or more highly similar isobaric sequences, whereas the 1,835 linear peptides reported for the same allele are rarely isobaric. Because the published table omitted spectrum identifiers (see Editorial and Publication Considerations), the isobaric similarity suggests that the sequence-to-spectrum associations were lost and multiple sequences per spectrum were reported as if they were independent observations rather than as ambiguous sequences with only one possible correct

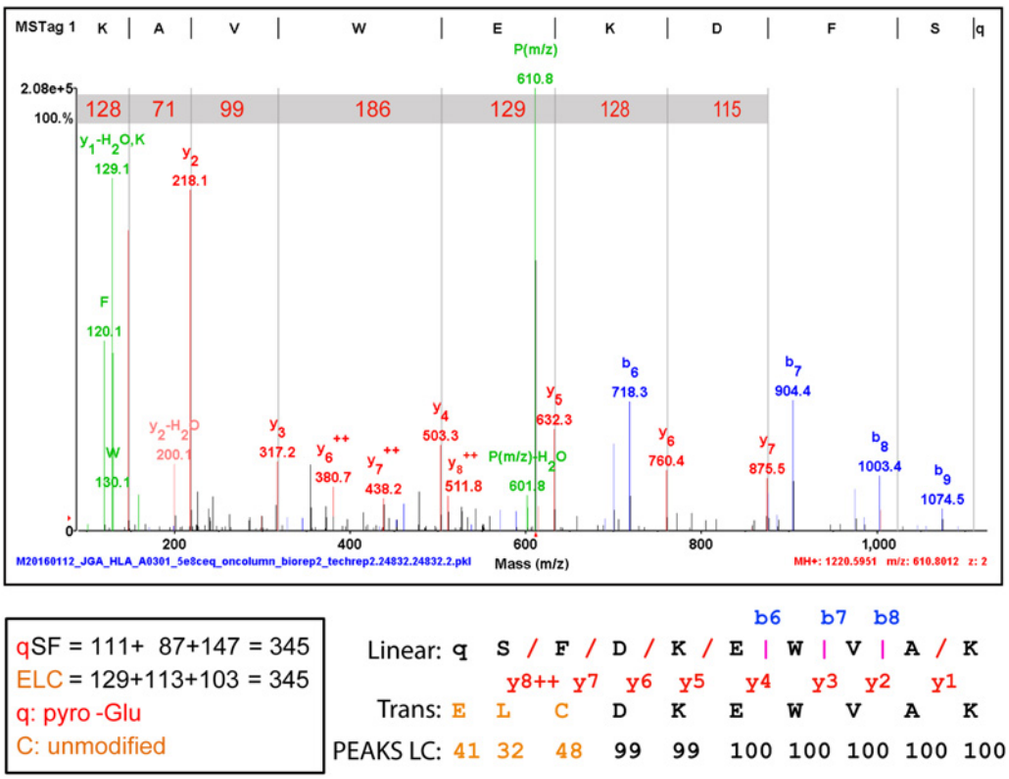

Figure 4.

PSM of proposed *trans*-spliced peptide fits modified linear peptide with shared partial sequence. The *trans*-spliced peptide ELC_DKEWVAK (PASK, 1296–1298 spliced to COX4, 134–140), with C in unexpected sulfhydryl form, reported in Faridi and colleagues (17) fits this MS-MS spectrum without ions confidently supporting the ELC portion of the sequence (amber), as indicated by the much lower PEAKS LC scores for those residues (41-32-48) shown below the sequence. The gray bar highlights the near-complete y ion series (consecutive fragments containing the C-terminus) that allows the partial sequence DKEWVAK to be determined *de novo* in reverse order (above spectrum) via the mass gaps between peaks (red). Instead, the lab that generated the dataset, reported the linear peptide qSFDKEWVAK, derived entirely from the human protein COX4, for this spectrum from allele HLA-A*03:01 with a theoretical precursor m/z of 610.8009 and experimental precursor mass error of +0.4 ppm (37). The common laboratory sample handling modification, pyro-Glutamic acid at peptide N-termini (q) was not accounted for in the Faridi and colleagues data analysis workflow (17) but was in Sarkizova and colleagues (37). With ELCDKEWVAK yielding nearly the same theoretical precursor m/z of 610.8026 and an experimental precursor mass error of −2.3 ppm, both are within the expected mass error tolerance for the Thermo Fisher Scientific Q Exactive Plus that generated the spectrum. Although qSF and ELC have nearly identical mass, qSF is supported better by the F immonium ion at 120 m/z, and the y_8++ ion at 511.8 m/z.

peptide assignment. Similarly, Ouspenskaia and colleagues (27) reported 308 novel unannotated ORF-derived peptides, whose translation is supported by Riboseq, that were mapped to the same MS-MS spectra as 343 spliced peptides reported in Faridi and colleagues for the nine monoallelic datasets originally published by Abelin and colleagues (18). Consequently, if one uses these observations to correct for overcounting due to multiple sequences/spectrum the number of individual spliced peptides reported by Faridi and colleagues (17) would decrease by 10%–14% (35/343 and 76/560) and their corresponding estimate of spliced peptides present in the immunopeptidome would decrease from 29% (15,320/53,665, reported in Faridi and colleagues) to 26% [−14%: (15,320 − 2,145)/(53,665 − 2,145) or −10%: (15,320 − 1,532)/(53,665 − 1,532)].

MS-MS spectral intensity prediction and search engine agreement

Current database search algorithms typically employ simple models for relative intensity of different fragment ion types. While y ions and b ions may be modeled with different intensity, all y ions are usually modeled with the same intensity, without incorporating well-known sequence-specific tendencies such as the generation of a very intense y ion from fragmentation at the N-terminal side of proline. However, there are now deep-learning methods that enable accurate prediction of intensities in MS-MS spectra for HLA-binding peptides (35). Prosit, trained specifically on hundreds of thousands of MS-MS spectra of synthetic HLA-binding peptides, enabled rescoring of PSMs generated by conventional search engine tools, thus improving the accuracy and sensitivity of identification of HLA-binding peptides (35).

Once trained, the Prosit tool was used to assess similarities between the measured MS-MS assigned to potential PSPs and the predicted MS-MS of those same sequences reported in Liepe and colleagues (35), for the subset of data that is derived from Bassani-Sternberg and colleagues (14). When comparing spectrum similarity score distributions for measured versus predicted spectra for the sets of nonspliced (canonical) peptides versus PSPs reported by Liepe and colleagues, most of the proposed spliced peptides have much lower spectral similarity to the Prosit predictions compared with the canonical peptides [spectral angle (SA) = 0.72 vs. SA = 0.87] (35). Using a combination of Prosit scoring and agreement between the search engines Mascot, MaxQuant, and MSFragger, the reanalysis concluded that 1,067 of the 1,230 (87%) PSPs were not conclusively supported by the MS data. Those conclusions were summarized in four categories: 596 (48%) did not remain confident after Prosit rescoring, 90 (7%) were leucine/isoleucine isomers that cannot be distinguished by MS-MS, 315 (26%) had a more confident canonical PSM identified by MaxQuant and/or MSFragger, and in 66 (5%) the proposed spliced peptide had a score comparable with that of a canonical peptide. A similar reanalysis of the 3,994 reported canonical peptides from the same study rejected just 475 (11%) peptides; 179 (4%) did not remain confident after Prosit rescoring and another 296 peptides (7%) had a more confident canonical PSMs identified by MaxQuant and/or MSFragger.

Cysteine-containing peptide observation rate

Cysteine-containing peptides present unique challenges to routine experimental detection. Sample handling and chromatographic separation often create artifacts and compromise observation rates when free sulfhydryl cysteine or disulfide cross-linked cysteines are present. Hence, typical proteomics experiments employ reduction and alkylation with reagents like dithiothreitol and iodoacetamide (IAA) to reduce and then covalently modify the cysteines, and the mass of the stable modification is defined in the search parameters. For detection of HLA-presented peptides, inclusion of a low amount of IAA (10 mmol/L) in the lysis buffer both inhibits cysteine proteases and increases (∼5×) the LC/MS-MS detection of cysteine-containing peptides (36–38) in carbamidomethylated form. Cysteines are highly reactive, and not all become alkylated. Thus, the fixed and variable modifications allowed on cysteine during database searching or *de novo* interpretation are critical to the successful identification of cysteine-containing peptides.

Discrepancies related to the inclusion of cysteine modification in database searches have made significant contributions to erroneous PSP identifications. For example, in Liepe and colleagues, for the subset of data that is derived from Bassani-Sternberg and colleagues (14), many spectra assigned as PSPs containing adjacent glycine and cysteine residues were shown to be more consistent with carbamidomethylcysteine-containing canonical sequences (see **Table 1**; ref. 39). Within the subset of data presented in Faridi and colleagues that is derived from the monoallelic immunopeptidomics resource published by Abelin and colleagues, the median cysteine-containing peptide proportions per allele are 0.2% for linear peptides, 4% for *cis*-spliced, and 15% for *trans*-spliced, resulting from a data analysis workflow where cysteines were only considered to be in free sulfhydryl form. In Abelin and colleagues, cysteines were predominantly observed in cysteinylated form (18). Thus, the simplest interpretation is that the enrichment of cysteine-containing peptides in the cis and *trans*-spliced peptides reported by Faridi and colleagues using Abelin and colleagues data is likely related to misidentification (see **Fig. 4**).

HLA anchor motifs and binding specificity

Following processing by the proteasome, PSPs follow the same antigen presentation pathway as linear peptides; therefore, overall, their HLA-binding specificities should be similar. Peptide sequence motifs for each HLA allele are well characterized, and HLA-I–binding predictors can accurately predict association between any given peptide of defined length (typically 9–14 mers) and hundreds of HLA class I alleles. Unsupervised alignment and clustering of MS-detected immunopeptidomes, using a tool such as Gibbs clustering (40, 41), can reveal the binding motifs. Clustering separately the fraction of potentially spliced peptides and linear peptides should reveal similar motifs, and comparison of the sample-generated motifs with known reference motifs can provide a quick assurance for the overall correct identification. The clustering approach typically performs well when hundreds or thousands of peptides are analyzed, also when the HLA typing of the investigated sample is unknown. Applying this approach to the hundreds of *cis*-spliced peptides and the thousands of linear peptides identified in the same sample by Liepe and colleagues revealed marked differences in the binding motifs (15).

The HLAthena binding prediction program trained on >185,000 MS-MS identified peptides eluted from 95 HLA-A, -B, -C, and -G monoallelic cell lines (including the alleles studied by Liepe and colleagues and Faridi and colleagues) showed that PSPs from both Liepe and colleagues and Faridi and colleagues are distributed over a wide range of predicted binding scores, whereas linear peptides are predicted with very high affinity (37). Most reported PSPs had poor predicted binding. A binding likelihood score threshold of >0.75 passed 81% of canonical linear peptides but only 28% of *cis*-spliced peptides described by Liepe and colleagues. Similar results for binding likelihood score were obtained for the peptides described by Faridi and colleagues: 84% linear vs. 36% *cis*- and 37% *trans*-spliced.

Experimental MS-Based Validation of Peptides

Many modifications (such as oxidation, acetylation, and phosphorylation) that occur biologically and/or as sample handling artifacts, can lead to errors in interpretation of MS-MS spectra (ref. 39; see **Table 1** for a list of modifications). Hence, identification of PSPs, for which no genomic or transcriptomic complementary validation datasets are available, must be supported with additional thorough experimental validation.

Experimental validation of peptide identification by MS-MS can be done by analyzing, with the same LC/MS-MS instrumentation and methods, synthetic counterparts of the identified peptides. Occasionally, the score difference provided by the search engine tools between the best fit and the second-best fit is very low, resulting in high ambiguity. Applying search engine tools with different algorithms may give more supportive evidence. Ideally, both the best score hits and the second hits should be synthesized, and their fragmentation patterns and chromatographic retention time should be compared. The most reliable experimental method to confirm the correct identification of a peptide is by targeted MS analysis performed with synthetic heavy isotopically labeled peptide counterparts spiked into the original immunopeptidomics sample in which the PSP was initially identified. Synonymous MS-MS spectra of the endogenous "light" and the synthetic "heavy" peptide and their coelution provide the most definite validation of correct peptide identification (25, 26). This method, however, remains an expensive method with low throughput. Although it will validate the correct identification of a peptide, it cannot reveal a peptide's mechanism of creation inside a cell.

It is important to evaluate the quality of each synthetic peptide as impurities may lead to critical artifacts. In the case of nonlabeled peptides, truncated byproducts and incomplete coupling of hard-to-synthesize peptide sequences can also lead to artifacts. For example, when Fritsche and colleagues attempted to validate a PSP derived from a mutated $KRAS_{2-35}$ G12V precursor sequence (TEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPT), they found frequent synthesis byproducts containing only one (32%) or two (6%) instead of three consecutive valines (42). Such a sequence precursor was used for supporting *in vitro* proteasomal splicing of the mutated KRAS by Mishto and colleagues (10), leading to the generation of the spliced KLVVGAVGV peptide. As no quality control of the synthesized precursor was provided by Mishto and colleagues, and given this high rate of impurities, the production of this spliced product was debated (11, 43, 44). In general, the computational and experimental approaches we discuss above can be applied for the validation of PSPs generated in *in vitro* digestion assays. In addition, it is important to evaluate the quality of synthetic heavy isotopically labeled peptides prior to spike-in experiments, since trace impurities of "light" counterparts can lead to false positives, as shown and discussed by Fritsche and colleagues (42).

Biological and Molecular Validation

In vitro studies have suggested that peptide splicing by the proteasome is a low efficiency process, as only 1% to 2% of all fragments produced by proteasome-mediated degradation are PSPs (45). This is in stark contrast with the notion that around 25% of the peptidome corresponds to PSPs. Because the latter estimation is based on MS, which can lead to ambiguous identification of HLA-bound PSPs, it is necessary to confirm these assertions by adding additional biological controls to fully validate the existence of PSPs at the surface of tumor cells. Complete characterization of tumor-associated peptides includes validating the nature, the immunogenicity and the natural presentation of these peptides by tumor cells. This is essential not only to clarify the ambiguities raised about the fraction of PSPs in the peptide repertoire, but also to confirm the relevance of PSPs as targets for immunotherapeutic approaches. At the time the first PSPs were discovered, immunopeptidomics and targeted validation methods using spiked-in heavy labeled standard peptides were not developed enough, hence, these first PSPs were not validated by these analytic methods. However, in these pioneering studies several key biological experiments were performed to validate both the existence of these peptides at the surface of tumor cells and their generation through proteasomal splicing.

A key step in the validation of antigenic peptides is to show that tumor cells potentially expressing that antigen can be recognized by a stable human T lymphocyte clone specifically recognizing that peptide (46). To do so, a CTL clone should be isolated that recognizes the peptide of interest. This can be done using the "reverse immunology approach," which is based on the *in vitro* activation of $CD8^+$ T cells from a healthy donor with mature dendritic cells pulsed with the relevant peptide (47). This CTL clone is then further tested for its ability to lyse the target tumor cells or to produce cytokines when cocultured with tumor cells. If clones cannot be isolated and polyclonal T cells are used instead, it is essential to demonstrate that the CTLs that recognize tumor cells in this polyclonal population are the same as those recognizing the peptide. This can be done by performing a "cold target inhibition" experiment, in which lysis of tumor cells by the polyclonal T cells is monitored in the presence of an excess of unlabeled cells pulsed or not with the relevant peptide. Tumor cell lysis can be measured in a standard Cr^{51}-release assay, or in a FACS-based assay with fluorescently labeled tumor cells. If tumor cells naturally express the peptide against which the polyclonal T cells were obtained, the addition of excess unlabeled antigen-presenting cells pulsed with that specific peptide should inhibit lysis of the tumor cells by the T cells.

One study has reported the detection of PSP-specific T cells after *in vitro* stimulation of lymphocytes from patients with melanoma with candidate PSPs and suggested such peptides as relevant immunotherapy targets (48). However, it is important to note that, because the immune system is educated to recognize any non–self-peptide, it is generally easy to observe some degree of immunogenicity against any given non–self-peptide. Hence, the detection or isolation of T cells against a peptide does not mean that this peptide can be naturally presented by tumor cells. Therefore, demonstration of the immunogenicity of a candidate spliced peptide is by itself not a way to validate this peptide. The goal of immunogenicity studies is to isolate specific T cells, which then can be used for the primary validation, which is to show that such T cells can kill cells expressing the parental protein(s) and not cells that do not. Even when a CTL clone is isolated against a peptide and shown to recognize the tumor, cross-recognition of another peptide–HLA complex can still occur (11). It is therefore important to confirm that the PSP can be processed from the full-length protein or that the PSP indeed originates from that protein. A way to test this is to show that the isolated CTL recognizes cells transfected with plasmids encoding the antigen-coding protein and the relevant HLA class I molecule. In addition, mutation of key residues present in either splicing partner should alter T-cell recognition and confirm that this spliced peptide constitutes a valid antigen. Alternatively, full knockout of the gene(s) encoding the antigenic peptide should also abolish T-cell recognition.

If no CTL can be made available or if large-scale identifications are required, tumor lines knocked out for the gene(s) encoding either of

the splicing partners or mutated at key residues of the peptide splicing partner sequences could be obtained and their peptidome compared with that of the original tumor cells. Spike-in experiments using heavy synthetic peptides as described above should be used to confirm the presence of the peptide of interest in the wild-type eluate and its absence in the knockout or mutated samples.

If a particular peptide sequence has passed all the bioinformatics and MS tests described above, yet fails biological validation as a PSP, this peptide is likely to be of yet an unknown source.

Editorial and Publication Considerations

As scientists all of us adhere to the adage "Extraordinary claims require extraordinary evidence" (attributed to Carl Sagan). Although some articles have concluded that PSPs occur frequently (12, 17), those claims were not fully supported by the now ordinary evidence typically accompanying articles published in proteomics and immunopeptidomics: the raw MS data deposited in a public data repository accompanied by exact description of the search parameters, the reference database, and tables of peptide identifications including spectrum identifiers so that PSMs can be verified by a motivated reader (49). Furthermore, quality controls of synthetic peptides, which are crucial to validate or invalidate *in vitro* digestion experiments, must be performed and the accompanying MS datasets should be deposited in a public data repository as mentioned above. It is critical to provide raw and processed data to confirm reports of any novel peptide assignments, whether they arise from novel ORFs, proteasomal splicing, or alternative RNA splicing (14, 18, 37, 50, 51). In addition, these datasets are invaluable tools in validating new immunopeptidomics workflows, as has been done in several studies (27, 35, 52).

We hope that the discussion we put forth in this review article will provide a critical foundation for the generation of well validated and highly reproducible proteasomal splicing data that withstands critical assessment by all in the research community.

Authors' Disclosures

K.R. Clauser reports grants from NIH during the conduct of the study, as well as a patent for methods for identifying neoantigens pending and a patent for HLA single allele lines issued, licensed, and with royalties paid from Neon Therapeutics. B.J. Van den Eynde reports personal fees from iTeos Therapeutics, Amgen, Oncorus, and Vaccitech outside the submitted work. No disclosures were reported by the other authors.

Acknowledgments

C.F. Lichti was supported by the NIH (R01 DK120340) and the Bursky Center for Human Immunology and Immunotherapy Programs (CHiiPs) at Washington University. N. Vigneron and B.J. Van den Eynde were supported by the Ludwig Institute for Cancer Research. K.R. Clauser was supported by National Cancer Institute Clinical Proteomic Tumor Analysis Consortium grants (NIH/NCI U24-CA210986 and NIH/NCI U01 CA214125, to S.A. Carr; NIH/NCI U24CA210979, to D.R. Mani). M. Bassani-Sternberg was supported by the Ludwig Institute for Cancer Research, grant PR00P3_193079 from the Swiss National Science Foundation, grant KFS-4680-02-2019-R from the Swiss Cancer Research Foundation and the Swiss Cancer League, grants from Cancera, Mats Paulsson, and a gift from the Biltema Foundation that was administered by the ISREC Foundation, Lausanne, Switzerland. We thank Jenn Abelin for expert assistance with the clarity of our presentation.

Received August 31, 2021; revised October 26, 2021; accepted January 14, 2022; published first February 1, 2022.

References

1. Hanada K, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. Nature 2004;427:252–6.
2. Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S, et al. An antigenic peptide produced by peptide splicing in the proteasome. Science 2004;304:587–90.
3. Dalet A, Vigneron N, Stroobant V, Hanada K, Van den Eynde BJ. Splicing of distant peptide fragments occurs in the proteasome by transpeptidation and produces the spliced antigenic peptide derived from fibroblast growth factor-5. J Immunol 2010;184:3016–24.
4. Warren EH, Vigneron NJ, Gavin MA, Coulie PG, Stroobant V, Dalet A, et al. An antigen produced by splicing of noncontiguous peptides in the reverse order. Science 2006;313:1444–7.
5. Dalet A, Robbins PF, Stroobant V, Vigneron N, Li YF, El-Gamil M, et al. An antigenic peptide produced by reverse splicing and double asparagine deamidation. Proc Nat Acad Sci U S A 2011;108:E323–31.
6. Robbins PF, el-Gamil M, Kawakami Y, Stevens E, Yannelli JR, Rosenberg SA. Recognition of tyrosinase by tumor-infiltrating lymphocytes from a patient responding to immunotherapy. Cancer Res 1994;54:3124–6.
7. Michaux A, Larrieu P, Stroobant V, Fonteneau JF, Jotereau F, Van den Eynde BJ, et al. A spliced antigenic peptide comprising a single spliced amino acid is produced in the proteasome by reverse splicing of a longer peptide fragment followed by trimming. J Immunol 2014;192:1962–71.
8. Liepe J, Mishto M, Textoris-Taube K, Janek K, Keller C, Henklein P, et al. The 20S proteasome splicing activity discovered by SpliceMet. PLoS Comput Biol 2010;6:e1000830.
9. Ebstein F, Textoris-Taube K, Keller C, Golnik R, Vigneron N, Van den Eynde BJ, et al. Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. Sci Rep 2016;6:24032.
10. Mishto M, Mansurkhodzhaev A, Ying G, Bitra A, Cordfunke RA, Henze S, et al. 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 2019;10:2572.
11. Willimsky G, Beier C, Immisch L, Papafotiou G, Scheuplein V, Goede A, et al. In vitro proteasome processing of neo-splicetopes does not predict their presentation in vivo. Elife 2021;10:e62019.
12. Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A, et al. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science 2016;354:354–8.
13. Liepe J, Sidney J, Lorenz FKM, Sette A, Mishto M. Mapping the MHC class I-spliced immunopeptidome of cancer cells. Cancer Immunol Res 2019;7:62–76.
14. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics 2015;14:658–73.
15. Mylonas R, Beer I, Iseli C, Chong C, Pak HS, Gfeller D, et al. Estimating the contribution of proteasomal spliced peptides to the HLA-I ligandome. Mol Cell Proteomics 2018;17:2347–57.
16. Erhard F, Dolken L, Schilling B, Schlosser A. Identification of the cryptic HLA-I immunopeptidome. Cancer Immunol Res 2020;8:1018–26.
17. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, et al. A subset of HLA-I peptides are not genomically templated: evidence for cis- and trans-spliced peptide ligands. Sci Immunol 2018;3:eaar3947.
18. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity 2017;46:315–26.
19. Admon A. Are there indeed spliced peptides in the immunopeptidome? Mol Cell Proteomics 2021;20:100099.
20. Rolfs Z, Muller M, Shortreed MR, Smith LM, Bassani-Sternberg M. Comment on "A subset of HLA-I peptides are not genomically templated: evidence for cis- and trans-spliced peptide ligands. Sci Immunol 2019;4:eaaw1622.
21. Faridi P, Li C, Ramarathinam SH, Illing PT, Mifsud NA, Ayala R, et al. Response to Comment on "A subset of HLA-I peptides are not genomically templated: evidence for cis- and trans-spliced peptide ligands. Sci Immunol 2019;4:eaaw8457.

22. Laumont CM, Perreault C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. Cellular and molecular life sciences. Cell Mol Life Sci 2018;75:607–21.
23. Charpentier M, Croyal M, Carbonnelle D, Fortun A, Florenceau L, Rabu C, et al. IRES-dependent translation of the long non coding RNA meloe in melanoma cells produces the most immunogenic MELOE antigens. Oncotarget 2016;7: 59704–13.
24. Chen J, Brunner AD, Cogan JZ, Nunez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science 2020;367:1140–6.
25. Chong C, Muller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun 2020;11:1293.
26. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure JP, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med 2018;10:eaau5516.
27. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol 2021 Oct 18 [Epub ahead of print].
28. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol 2018;36: 1056–8.
29. Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 2019;16:509–18.
30. Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun 2020;11:1759.
31. Krokhin OV, Ying S, Cortens JP, Ghosh D, Spicer V, Ens W, et al. Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS. Anal Chem 2006;78:6265–9.
32. Rolfs Z, Solntsev SK, Shortreed MR, Frey BL, Smith LM. Global identification of post-translationally spliced peptides with neo-fusion. J Proteome Res 2019;18: 349–58.
33. Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. Mol Cell Proteomics 2007;6:377–81.
34. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. J Comput Biol 1999;6:327–42.
35. Wilhelm M, Zolg DP, Graber M, Gessulat S, Schmidt T, Schnatbaum K, et al. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. Nat Commun 2021;12:3346.
36. Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, Goulding SP, et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. Immunity 2019;51:766–79.
37. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol 2020;38:199–209.
38. Chong C, Marino F, Pak H, Racle J, Daniel RT, Muller M, et al. High-throughput and sensitive immunopeptidomics platform reveals profound interferongamma-mediated remodeling of the human leukocyte antigen (HLA) ligandome. Mol Cell Proteomics 2018;17:533–48.
39. Lichti CF. Identification of spliced peptides in pancreatic islets uncovers errors leading to false assignments. Proteomics 2021;21:e2000176.
40. Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. Bioinformatics 2013;29:8–14.
41. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res 2017;45:W458–W463.
42. Fritsche J, Kowalewski DJ, Backert L, Gwinner F, Dorner S, Priemer M, et al. Pitfalls in HLA ligandomics-how to catch a Li(e)gand. Mol Cell Proteomics 2021; 20:100110.
43. Beer I. Commentary: 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 2021;12:523906.
44. Mishto M, Rodriguez-Hernandez G, Neefjes J, Urlaub H, Liepe J. Response: Commentary: 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 2021;12:679836.
45. Mishto M, Goede A, Taube KT, Keller C, Janek K, Henklein P, et al. Driving forces of proteasome-catalyzed peptide splicing in yeast and humans. Mol Cell Proteomics 2012;11:1008–23.
46. Vigneron N, Stroobant V, Van den Eynde BJ, van der Bruggen P. Database of T cell-defined human tumor antigens: the 2013 update. Cancer Immun 2013; 13:15.
47. Ottaviani S, Colau D, van der Bruggen P, van der Bruggen P. A new MAGE-4 antigenic peptide recognized by cytolytic T lymphocytes on HLA-A24 carcinoma cells. Cancer Immunol Immunother 2006;55:867–72.
48. Faridi P, Woods K, Ostrouska S, Deceneux C, Aranha R, Duscharla D, et al. Spliced peptides and cytokine-driven changes in the immunopeptidome of melanoma. Cancer Immunol Res 2020;8:1322–34.
49. Lill JR, van Veelen PA, Tenzer S, Admon A, Caron E, Elias J, et al. Minimal information about an immuno-peptidomics experiment (MIAIPE). Proteomics 2018:e1800110.
50. Bassani-Sternberg M, Braunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun 2016;7: 13404.
51. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. J Immunotherapy Cancer 2021; 9:e002071.
52. Löffler MW, Mohr C, Bichmann L, Freudenmann LK, Walzer M, Schroeder CM, et al. Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. Genome Med 2019;11:28.