

## **Navigating critical challenges associated with immuno-peptidomics-based detection of proteasomal spliced peptide candidates**

Cheryl F. Lichti<sup>1,2\*</sup>, Nathalie Vigneron<sup>3</sup>, Karl R. Clauser<sup>4</sup>, Benoit J. Van den Eynde<sup>3,5</sup>, Michal Bassani-Sternberg<sup>6,7\*</sup>

### Affiliations:

<sup>1</sup> Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA

<sup>2</sup> Bursky Center for Human Immunology and Immunotherapy, Washington University School of Medicine, St. Louis, MO, USA

<sup>3</sup> Ludwig Institute for Cancer Research, Brussels, Belgium; de Duve Institute, Université Catholique de Louvain, Brussels, Belgium

<sup>4</sup> Broad Institute of MIT and Harvard, Cambridge, MA, USA

<sup>5</sup> Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

<sup>6</sup> Ludwig Institute for Cancer Research, Lausanne Branch - University of Lausanne (UNIL), CH-1005, Switzerland

<sup>7</sup> Department of Oncology - Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne CH-1011, Switzerland

\* Corresponding authors: Cheryl F. Lichti [clichti@wustl.edu](mailto:clichti@wustl.edu), Michal Bassani-Sternberg [michal.bassani@chuv.ch](mailto:michal.bassani@chuv.ch)

The authors declare no potential conflicts of interest.

## **Abstract**

Within the tumor immunology community, the topic of proteasomal spliced peptides (PSPs) 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; all are produced by cis-splicing in the proteasome. However, the advent of algorithms to identify candidate PSPs in mass spectrometry data challenged this notion by concluding that the frequency of spliced peptides in the MHC-I ligandome was quite high. Since this time, much debate has centered around the methodologies used in these studies. Several re-analyses of data from these studies have uncovered bioinformatics errors that lead to questions about the validity of the conclusions in the high throughput studies. Further, the biological and technical validation that should be necessary for verifying PSP assignments was often lacking. Recently, it was suggested that our community should unite around a common set of standards for validating candidate PSPs. In this review we highlight 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 proteasomal splicing results.

## **I. Introduction to PSPs**

### **A. Initial discovery of antigenic peptides recognized by CTL and produced by peptide splicing in the proteasome**

Peptide splicing was initially described in the early 2000 when Hanada et al. identified the peptide recognized by a CD8+ cytotoxic T lymphocyte (CTL) clone isolated from a renal cell carcinoma patient (1). This CTL clone was able to kill cancer cells and was found to recognize a peptide, NTYAS\_PRFK, which was composed of two non-contiguous 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 fragments 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, while 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 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 with the identification of a second spliced peptide recognized by a tumor-killing CTL clone isolated from a melanoma patient (2). This 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 in EBV-B cells of 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 PSP at the surface of melanoma cells. The role of the proteasome in peptide splicing was initially highlighted by using proteasome inhibitors and then further confirmed by producing the spliced peptide RTK\_QLYPEW 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 (Figure 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- $\alpha$ -acetylation of QLYPEW completely prevented production of the spliced peptide when incubated with peptide RTKAWNR and the proteasome. Likewise, spliced peptide NTYAS\_PRFK initially described by Hanada et al. was subsequently shown to be produced in the proteasome by transpeptidation (3).

The third example of peptide splicing came a few years later when a CTL clone, isolated from the recipient of MHC-matched allogeneic hematopoietic cell transplantation, was shown to recognize a peptide encompassing the single nucleotide polymorphism A996G in the gene coding the SP110 nuclear protein (4). The CTL recognized cell lines displaying a SP110<sub>A996</sub> allele but not those displaying only the SP110<sub>G996</sub> allele. Interestingly, 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<sub>299</sub>, 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 example of spliced peptide was recognized by a clone of tumor-infiltrating lymphocytes (TILs) isolated from a melanoma patient (5). Interestingly, these TILs were used for adoptive T cell transfer and shown to induce dramatic tumor rejection in this melanoma patient (6). This highlights the biological and clinical relevance of PSPs. This peptide is composed of two non-contiguous fragments of the tyrosinase protein, which are 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 melanoma patient (7). In contrast to other examples of spliced peptides, which were composed of 3 to 6 amino-acid fragments, the peptide recognized by this TIL contained an N-terminal splicing partner of 8 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 3-amino acid long for the splicing reaction to take place. 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, the question is whether it also occurs in a biologically meaningful fashion within cells. 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 contain full-length FGF-5 constructs bearing mutations at critical residues of the fragments NTYAS and 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 (Figure 1). 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 physiological 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.

## **B. Mass spectrometry and first large-scale spliced peptide identifications by immuno-peptidomics**

In these initial studies, identification of PSPs always 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). Alternatively, identification of novel PSPs was attempted by systematically analyzing proteasome digests by MS 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 a 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 the identification of spliced KRAS<sub>G12V</sub> 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, by immunizing mice harboring the human TCR $\alpha\beta$  coding genes with peptide KL\_VVGAVGV, Willimsky et al. cloned a TCR able to recognize this peptide (11). This TCR was transduced into PBMC, which could then recognize

the target cells pulsed with the KL\_VVGAVGV peptide but not HLA-A2+ tumor cells endogenously expressing the mutant KRAS<sub>G12V</sub> or overexpressing a mutant KRAS<sub>G12V</sub> construct, supporting a lack of natural presentation of the spliced peptide in cells. This confirmed that in vitro-in silico 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 et al. used an immunopeptidomics approach to identify HLA ligands produced by peptide splicing (12). To do so, they created a custom search database which included all potential short peptides created by the splicing of non-contiguous protein fragments. To restrict the size of the database, they only considered spliced peptides produced from the cis-splicing of peptide fragments separated by 25 amino acids or less. Nevertheless, this resulted in an enormous reference database that was 100x larger than the typical human reference proteome. Using this approach, they initially estimated the contribution of spliced peptides to 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 et al. (14). Predictably, this raised many controversies, as the properties of the reported spliced HLA peptides were drastically different than those of the linear ligands detected within the same samples (12). Using identical data but different computational approaches, others have estimated that percentage to be at most 2 to 6% (15) or even less than 0.1% (16).

Two years later, Faridi et al. 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

8 monoallelic datasets of their own, but also in 9 previously published monoallelic datasets from Abelin et al. (18). These findings were controversial for several reasons, both at the biological and bioinformatics levels. Importantly, as previously noted, 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 paper brought major controversy about the correct identification of many of these peptides at the MS level (20,21).

To understand one part of the controversy, it is instructive to consider the likelihood of finding a random, hypothetical pair of trans-splicing donor proteins in the human reference proteome. Let's consider a 9mer PSP, formed by splicing 4 and 5 amino acids-long fragments 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 (Figure 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 co-processing at the proteasome and combined with lack of molecular validation, it seems as though it would be more sensible to categorize those left-over, high-confidence *de novo*-derived peptide sequences as peptides of UNKnown origin (pUNKs) than to infer random trans-splicing.

In recent years it has been shown that, beyond canonical peptides altered by phenomena such as single nucleotide variations, somatic mutations, and post-translational modifications (PTMs), HLA bound peptides can be derived from sources other than protein-coding regions (22). These 'non-canonical' 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, have been reported to generate non-canonical tumor antigens (23-28). Ouspenskaia et al. further demonstrated that, while many of the peptide sequences reported as spliced peptides by Faridi et al. were correct MS/MS interpretations, they could be accounted for as linear peptides derived from novel human open reading frames (ORFs) whose translation was supported by Riboseq (28). Such sources were annotated as non-coding and thus absent from the reference proteome considered by Faridi et al. While mis-identification of peptides in MS analyses can be rooted in the use of incomplete or non-personalized references, including all possible variants of canonical and non-canonical 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 mass spectrometry-based validation of peptide sequences as described in the following sections. These approaches are equally valid for peptides derived from canonical and non-canonical sources.

## **II. Bioinformatics and spliced peptide assignments - methods for evaluating overall dataset quality**

Given the controversy in possible spliced peptides as assigned by custom bioinformatics algorithms, it is instructive to discuss the evaluation of peptide-spectrum matches (PSMs) assigned by search engines and custom algorithms. In all cases, PSMs can be evaluated based on the same basic criteria: precursor mass error distributions, MS/MS spectral quality, chromatographic retention time predictions 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. Since 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 simple methods can sensibly be used as relative quality metrics to raise red flags about the overall integrity of peptide identifications in published datasets reporting PSPs.

### **A. 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) can be an orthogonal parameter for validation of peptide identification (20,32). Such simple analysis may be sufficient to eliminate some identifications that are likely to be wrong.

### **B. 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  $1e6$ . 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 standard deviations of the standard mass error (33). Mass errors that fall outside this range usually indicate an incorrect assignment. 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 Figure 3A, with a stacked bar histogram illustrating mass error distributions at the PSM level for an HLA-A\*01:01 datafile from Faridi et al. (17). The dotted lines, indicating 3x standard deviation, 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 Figure 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.

### C. 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 (CID) and high-energy collision-induced dissociation (HCD) fragmentation of peptides, amide bonds fragment in a predictable manner to generate 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 unmodified others (see 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 (MLC) of 80 (15) or average (ALC) score of ~80 (17) has been used as a threshold for exporting high quality sequences. In Figure 4, the LC scores 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 ELC 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 supplemental table 4 from Faridi et al. (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 re-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, while the 1835 linear peptides reported for the same allele are rarely isobaric. This suggests that the authors lost track of the sequence-to-spectrum



associations and reported multiple sequences per spectrum as if they were independent observations rather than as sequence ambiguity with only one possible correct peptide assignment. Similarly, Ouspenskaia et al. (28) 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 et al. for the 9 monoallelic datasets originally published by Abelin et al. (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 et al. (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 et al.) to 26% (-14%: (15320-2145)/(53665-2145) or -10%: (15320-1532)/(53665-1532)).

#### **D. 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, recently developed deep learning methods 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 identification accuracy and sensitivity (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 et al. (35), for the subset of data that is derived from Bassani-Sternberg et al. (14). When comparing spectrum similarity score distributions for measured vs predicted spectra for the sets of non-spliced (canonical) peptides vs PSPs reported by Liepe et al. , most of the proposed spliced peptides have much lower spectral similarity to the Prosit predictions compared to 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 re-analysis concluded that 1067 of the 1,230 (87%) PSPs were not conclusively supported by the mass spectrometry 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 to that of a canonical peptide. A similar re-analysis of the 3994 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.

## **E. 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 (DTT) 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 mM) in the lysis buffer both inhibits cysteine proteases and increases (~5x) 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 et al., for the subset of data that is derived from Bassani-Sternberg et al. (14), many spectra assigned as PSPs containing adjacent glycine and cysteine residues were shown to be more consistent with carbamidomethylcysteine-containing (see Table 1) canonical sequences (39). Within the subset of data presented in Faridi et al. that is derived from the mono-allelic immunopeptidomics resource published by Abelin et al., 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 et al., 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 et al. using Abelin et al. data is likely related to misidentification (see Figure 4).

## **F. HLA anchor motifs and binding specificity**

Following processing by the proteasome, PSPs follow the same 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 9-14mers) and hundreds of HLA-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 et al. 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 mono-allelic cell lines (including the alleles studied by Liepe et al.

and Faridi et al.) showed that PSPs from both Liepe et al. and Faridi et al. 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 et al. Similarly, poor results were obtained for peptides described by Faridi et al.: 84% linear vs. 36% cis- and 37% trans-spliced.

### **III. Experimental MS based validation of peptides**

Many modifications (such as oxidation, acetylation, phosphorylation) that occur biologically and/or as sample handling artifacts, may lead to similar errors in interpretation of MS/MS spectra (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 co-elution provide the most definite validation of correct peptide identification (24,27). This method, however, remains an expensive method with low throughput. While 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 case of non-labeled peptides, truncated byproducts and incomplete coupling of hard-to-synthesize peptide sequences may also lead to artifacts. For example, Fritsche et al. reported recently their attempt to validate a PSP derived from mutated KRAS<sub>2-35</sub> G12V precursor sequence (TEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPT), and found frequent synthesis byproducts that contains only one (32%) or two valine (6%) residues (42). Such a sequence precursor was used for supporting in vitro proteasomal splicing of the mutated KRAS by Mishto et al. (10) leading to the generation of the spliced KLVVGAVGV peptide. As no quality control of the synthesized precursor was provided by Mishto et al. 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 (42).

#### **IV. Biological and molecular validation**

*In vitro* studies have clearly 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 dramatic contrast with the notion that around 25% of the peptidome corresponds to PSPs. Because the latter estimation is based on mass spectrometry, which can lead to ambiguous identifications, 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 more importantly the natural presentation of these peptides specifically 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 pioneering studies discovered the first PSPs, immuno-peptidomics and the targeted validation methods using spiked-in heavy labeled standard peptides were not developed enough, hence, these first PSPs were not validated by these analytical 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 which recognizes the peptide of interest. This can be done by 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 co-cultured 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<sup>51</sup>-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 an excess of unlabeled antigen-presenting cells pulsed with that specific peptide should inhibit lysis of the tumor cells by the T cells.

A recent study reported the detection of PSP-specific T cells after *in vitro* stimulation of lymphocytes from melanoma patients with candidate PSP peptides and suggested such peptides as relevant immunotherapy targets (PMID: 32938616). 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 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 particularly important for PSPs to confirm that the peptide can be processed from the full-length protein or that the peptide indeed originates from that protein. An easy way to test this is to show that the isolated CTL also recognizes cells transfected with plasmids encoding the antigen-coding protein and the relevant HLA class I molecule. Additionally, 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 knock-out of the gene(s) encoding the antigenic peptide should also abolish T cell recognition.

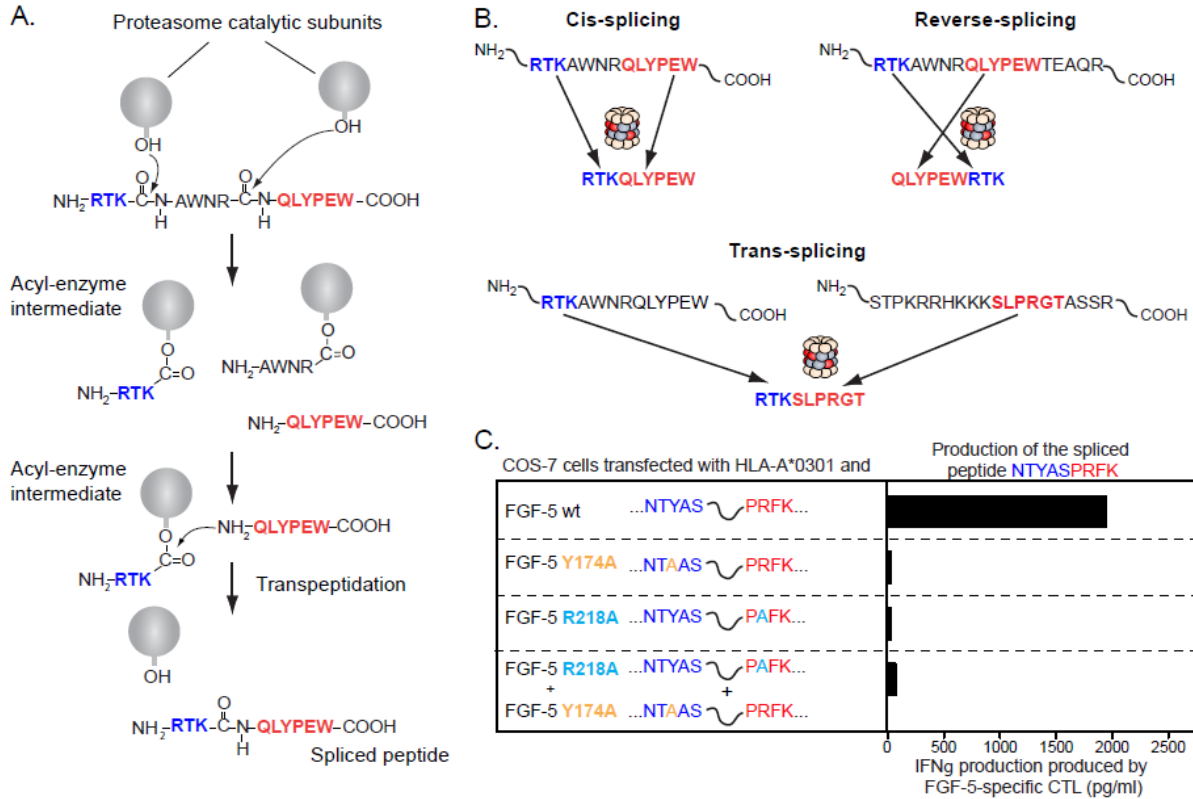
If no CTL can be made available or if large-scale identifications are required, tumor lines knocked-out for genes encoding either of the splicing partners or mutated at key residues of the peptide splicing partner sequences could be obtained and their peptidome compared to that of the original tumor 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 WT eluate and its absence in the knock-out or mutated samples.

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

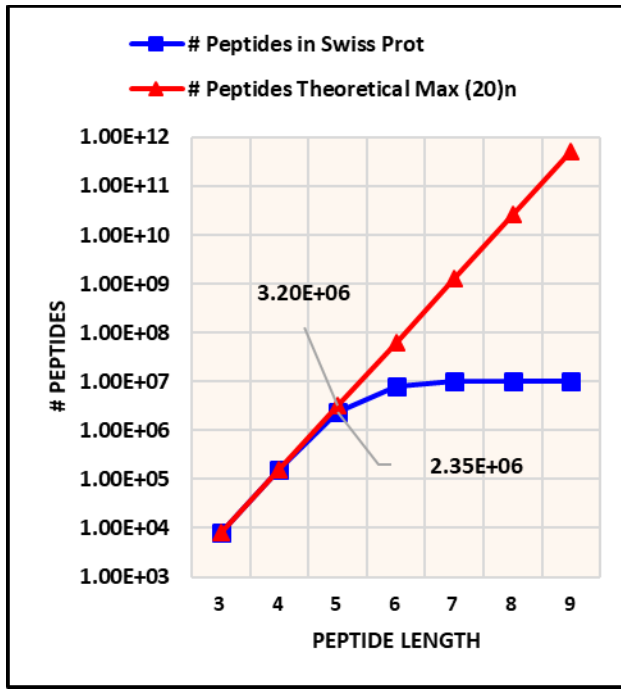
## **V. Editorial and publication considerations**

As scientists all of us tend to adhere to the adage 'Extraordinary claims require extraordinary evidence' (ECREE, attributed to Carl Sagan). While some recent papers might suppose that PSPs are ordinary (12,17), the journal editors of those papers have, to the surprise of many in the proteomics field, not required those claims to be fully supported by the now ordinary evidence typically accompanying papers published in proteomics and immunopeptidomics: the raw mass spectra deposited in a public data repository accompanied by exact description of the search parameters, the reference database, and tables of peptide identifications including a spectrum identifier so that a peptide spectrum match can be verified by a motivated reader (48). 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 as mentioned above. It is critical to provide this data to confirm reports of any novel peptide assignments, whether they arise from novel open reading frames, proteasomal splicing, or alternative RNA splicing. Additionally, these datasets prove to be invaluable tools in validating new immunopeptidomics workflows, as has been done in several recent studies. Thus, we hope that both reviewers and editors will request such data when any paper is submitted.

## Figures and Figure Legends

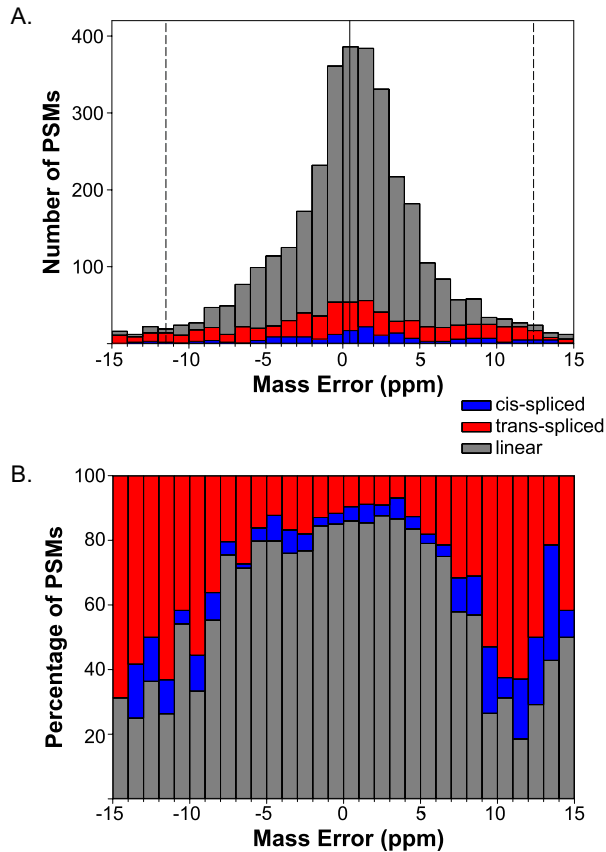


**Figure 1. Peptide splicing in the proteasome occurs by transpeptidation.** (A) Splicing of the gp100-derived peptide RTKQLYPEW is shown. Upon formation of an acyl-enzyme intermediate involving the N-terminal splicing partner RTK, the amino group of the C-terminal splicing partner QLYPEW produces a nucleophilic attack on the acyl-enzyme intermediate to create a new peptide RTKQLYPEW 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 physiological conditions (data originally published in (3)). To study the physiological 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 IFN $\gamma$  production by the FGF-5-specific CTL-C2. The amount of IFN $\gamma$  produced by the CTL after incubation with the transfected cells is measured by ELISA.



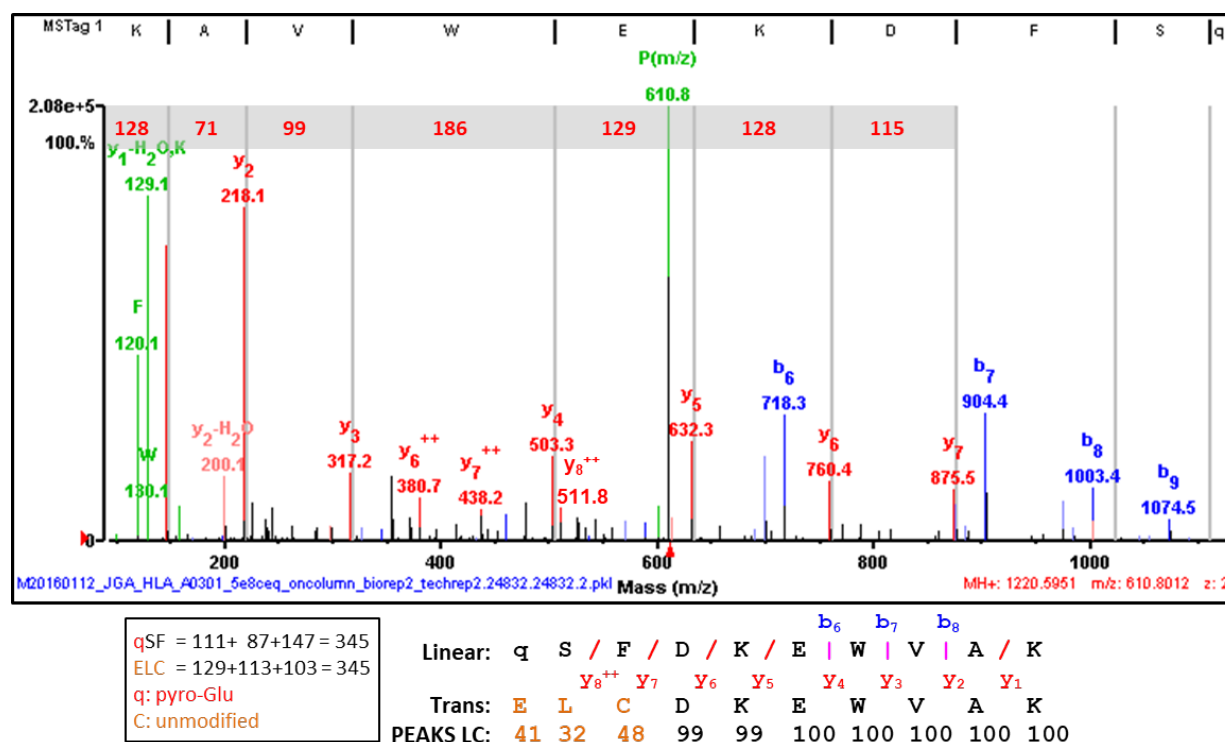
**Figure 2. Probability of finding n-mers.** The number of unique sequences of length  $n$  found in Swiss-Prot (20,381 proteins, 3/31/2021) plotted along with the theoretical maximum number of sequences of length  $n$  ( $20^n$ ). 73.4% of all possible sequences of length 5 and 99.9% of all sequences 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.

Figure 3



**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 et al. The dotted lines illustrate  $\pm 3$  x standard deviations 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.





**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 et al. (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 bars highlight 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 et al. data analysis workflow (17) but was in Sarkizova et al. (37). With ELC\_DKEWVAK 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 QExactive Plus that generated the spectrum. While qSF and ELC have nearly identical mass, qSF is supported better by the F immonium ion at 120 m/z, and the y<sub>8</sub><sup>++</sup> ion at 511.8 m/z.

**Table 1. A representative list of isobaric amino acid residues, modified amino acid residues, and chemical and post-translational modifications.** Isobaric amino acids/groups of amino acids/modified residues can confound spectral assignments during database searching, especially when incomplete fragmentation occurs. Chemical and post-translational modifications can also confound assignments and should be considered when confirming PSMs for PSPs or other noncanonical peptides. A brief list of these species, including chemical modifications that commonly lead to erroneous PSP assignments, is included.

**I. Isobaric amino acid(s)/modified amino acid(s)**

<b>Da</b>	<b>Amino acid(s)/modifications</b>
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 (Gln)
42.011	N-terminal acetylation
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

21.982 Sodium adduct (Asp, Glu)

-18.011 Dehydration (Ser, Thr)

0.984 Deamidation (Asn)

47.985 Cysteine oxidation to cysteic acid

79.966 Phosphorylation (Ser, Thr, Tyr)

## References:

1. Hanada K, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* **2004**;427(6971):252-6 doi 10.1038/nature02240.
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(5670):587-90 doi 10.1126/science.1095522.
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. *Journal of immunology* **2010**;184(6):3016-24 doi 10.4049/jimmunol.0901277.
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(5792):1444-7 doi 10.1126/science.1130660.
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. *Proceedings of the National Academy of Sciences of the United States of America* **2011**;108(29):E323-31 doi 10.1073/pnas.1101892108.
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 research* **1994**;54(12):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. *Journal of immunology* **2014**;192(4):1962-71 doi 10.4049/jimmunol.1302032.
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 computational biology* **2010**;6(6):e1000830 doi 10.1371/journal.pcbi.1000830.
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 doi 10.1038/srep24032.
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. *Frontiers in immunology* **2019**;10:2572 doi 10.3389/fimmu.2019.02572.
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 doi 10.7554/eLife.62019.
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(6310):354-8 doi 10.1126/science.aaf4384.
13. Liepe J, Sidney J, Lorenz FKM, Sette A, Mishto M. Mapping the MHC Class I-Spliced Immunopeptidome of Cancer Cells. *Cancer immunology research* **2019**;7(1):62-76 doi 10.1158/2326-6066.CIR-18-0424.
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. *Molecular & cellular proteomics : MCP* **2015**;14(3):658-73 doi 10.1074/mcp.M114.042812.

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. *Molecular & cellular proteomics : MCP* **2018**;17(12):2347-57 doi 10.1074/mcp.RA118.000877.
16. Erhard F, Dolken L, Schilling B, Schlosser A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer immunology research* **2020**;8(8):1018-26 doi 10.1158/2326-6066.CIR-19-0886.
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(28) doi 10.1126/sciimmunol.aar3947.
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(2):315-26 doi 10.1016/j.immuni.2017.02.007.
19. Admon A. Are There Indeed Spliced Peptides in the Immunopeptidome? *Molecular & cellular proteomics : MCP* **2021**;20:100099 doi 10.1016/j.mcpro.2021.100099.
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(38) doi 10.1126/sciimmunol.aaw1622.
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(38) doi 10.1126/sciimmunol.aaw8457.
22. Laumont CM, Perreault C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cellular and molecular life sciences : CMLS* **2018**;75(4):607-21 doi 10.1007/s00018-017-2628-4.
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(37):59704-13 doi 10.18632/oncotarget.10923.
24. 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. *Science translational medicine* **2018**;10(470) doi 10.1126/scitranslmed.aau5516.
25. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, *et al.* Intron retention is a source of neoepitopes in cancer. *Nature biotechnology* **2018** doi 10.1038/nbt.4239.
26. 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(6482):1140-6 doi 10.1126/science.aay0262.
27. 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. *Nature communications* **2020**;11(1):1293 doi 10.1038/s41467-020-14968-9.
28. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, *et al.* Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. **2020**:2020.02.12.945840 doi 10.1101/2020.02.12.945840 %J bioRxiv.
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. *Nature methods* **2019**;16(6):509-18 doi 10.1038/s41592-019-0426-7.
30. Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature communications* **2020**;11(1):1759 doi 10.1038/s41467-020-15456-w.

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(17):6265-9 doi 10.1021/ac060251b.
32. Rolfs Z, Solntsev SK, Shortreed MR, Frey BL, Smith LM. Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion. *Journal of proteome research* **2018** doi 10.1021/acs.jproteome.8b00651.
33. Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. *Molecular & cellular proteomics : MCP* **2007**;6(3):377-81 doi 10.1074/mcp.M600380-MCP200.
34. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* **1999**;6(3-4):327-42 doi 10.1089/106652799318300.
35. Wilhelm M, Zolg DP, Graber M, Gessulat S, Schmidt T, Schnatbaum K, *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nature communications* **2021**;12(1):3346 doi 10.1038/s41467-021-23713-9.
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(4):766-79 e17 doi 10.1016/j.immuni.2019.08.012.
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. *Nature biotechnology* **2020**;38(2):199-209 doi 10.1038/s41587-019-0322-9.
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. *Molecular & cellular proteomics : MCP* **2018**;17(3):533-48 doi 10.1074/mcp.TIR117.000383.
39. Lichti CF. Identification of spliced peptides in pancreatic islets uncovers errors leading to false assignments. *Proteomics* **2021**;21(7-8):e2000176 doi 10.1002/pmic.202000176.
40. Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* **2013**;29(1):8-14 doi 10.1093/bioinformatics/bts621.
41. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic acids research* **2017** doi 10.1093/nar/gkx248.
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. *Molecular & cellular proteomics : MCP* **2021**;20:100110 doi 10.1016/j.mcpro.2021.100110.
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. *Frontiers in immunology* **2021**;12:523906 doi 10.3389/fimmu.2021.523906.
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. *Frontiers in immunology* **2021**;12:679836 doi 10.3389/fimmu.2021.679836.
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. *Molecular & cellular proteomics : MCP* **2012**;11(10):1008-23 doi 10.1074/mcp.M112.020164.
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 immunity* **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 immunology, immunotherapy : CII* **2006**;55(7):867-72 doi 10.1007/s00262-005-0053-2.

48. 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 doi 10.1002/pmic.201800110.