Spliced HLA bound peptides; a Black-Swan event in Immunology

Pouya Faridi[1], Mohammadreza Dorvash[1], Anthony W. Purcell[1,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, VIC 3800, Australia

*to whom correspondence should be addressed, Anthony.purcell@monash.edu

Abstract

Peptides that bind to and are presented on the cell surface by Human Leukocyte Antigens (HLA) molecules play a critical role in adaptive immunity. For a long time, it was believed all of the HLA bound peptides were generated through simple proteolysis of linear sequences of cellular proteins, and therefore, are templated in the genome and proteome. However, evidence for untemplated peptide ligands of HLA molecules has accumulated over the last two decades, with a recent global analysis of HLA-bound peptides suggesting that a considerable proportion of HLA bound peptides are potentially generated through splicing/fusion of discontinuous peptide segments from one or two distinct proteins. In this review, we will review recent discoveries and debates on the contribution of spliced peptides to the HLA class I immunopeptidome, consider biochemical rules for splicing, and the potential role of these spliced peptides in immune recognition.

Introduction

The term Black Swan originates from the (Western) belief that all swans are white because these were the only ones that had been encountered. However, in 1697 the Dutch explorer Willem de Vlamingh discovered black swans in Australia. This was an unexpected event in (scientific) history and changed zoology. The black swan theory is now a metaphor that describes an event that comes as a surprise, has a significant effect, and is often poorly rationalised after the fact with the benefit of hindsight [1].

Two seminal papers in the early 21st century discovered peptides presented by Human Leukocyte antigen class I (HLA-I) that were not monolithically templated in the genome [2, 3]. This was like a Black-Swan event in immunology as the central dogma in biology is "all the proteins and peptides in a cell" are genomically templated. These peptides were subsequently shown to be generated through a post-translational modification (PTM) called splicing. This PTM generally happens in the proteasome – the central peptide generation machinery in the antigen processing and presentation pathway (Figure 1). The first identified spliced peptide was generated through splicing of segments 172-176 and 217-220 of Fibroblast growth factor 5 (FGF-5) [2], with a second example appearing soon after in the literature that was generated from splicing of segments 40-42 and 47-52 of the melanocyte protein PMEL [3]. The source protein of both of these peptides are cancer-associated antigens, and both of these peptides were immunogenic, which indicates their role in antigen recognition, and ultimately, their potential as targets in cancer immunotherapy. These discoveries were a surprise to the immunology field, and since their identification, they have been reported sporadically. For example, in 2006, the first spliced peptides generated through a reverse splicing mechanism (the peptide contains two non-contiguous segments of a protein that were spliced together in the reverse order to that in which they occur in the parental protein) were described (Sp110 nuclear body protein 296–301/286–289) [4] and in 2011 a reverse spliced peptides which carried chemical PTMs (Tyrosinase 368–373/336–340) were discovered [5]. In 2010, a spliced peptide from FGF-5 was identified, which the distance between two donor segments was 49 amino acids [6]. In the vast majority of cases, both donor segments of spliced peptides were derived from the same protein source (*i.e.*, a cis-splicing event). However, it has also been shown that trans-splicing is also possible, and this event occurs between two donor segments derived from distinct proteins [5].

In 2016, the first spliced peptide presented by Human Leukocyte antigen class II (HLA-II) molecules were reported. Surprisingly, this was a trans-spliced peptide resulting from splicing between two beta cell autoantigens, proinsulin (64-71) and Islet amyloid polypeptide 2 (IAPP2 74-80) [7]. T-cell clones from two Type 1 diabetic (T1D) patients recognised this peptide. This discovery suggested a role of the spliced peptides in autoimmune diseases, although the precise mechanism for generating HLA-II bound spliced peptides, coined hybrid insulin peptides (HIPs) by the authors, is yet to be defined.

Spliced peptide discovery and validation using mass spectrometry techniques

Rapid improvements in speed and sensitivity have positioned mass spectrometry as the method of choice for high-throughput peptide antigen discovery studies [8]. The most precise method for investigating immunopeptidomes (the array of peptides bound to and presented at the cell surface by HLA molecules) is through immunoaffinity purification of native peptide-HLA complexes and subsequent sequencing of the bound peptides by liquid chromatography-tandem mass spectrometry (LC-MS/MS) – a technique coined immunopeptidomics [9]. The data acquired is interrogated using algorithms that typically rely upon a reference proteome for spectral matching to assign peptide sequences to the experimental MS/MS data [9]. As expected, proteomics also follows the central dogma of biology that all the proteins/peptides have an intact template in the genome. Although this general method is successful for identifying genomically templated sequences (i.e., in this case, "linear" or templated peptide antigens), the absence of sequence information for spliced peptides in the predicted proteome precludes the use of this workflow for their identification [10].

The first attempts that used mass spectrometry data to identify spliced peptides were based on generating a theoretical database of spliced peptides and on interrogating this database post immunopeptidomics workflows [11-14]. The size of a database that contains all possible spliced peptides (for instance, for the human proteome) is massive and intolerable for immunopeptidomics/bioinformatics workflows. To address this issue, Liepe and colleagues generated a more restricted database that contained theoretical cis- and reversed cis-spliced peptides with a maximum of 25 amino acid distance between any two segments [15]. Although still computationally intensive, this approach suggested that spliced peptides were far more prevalent than previously suspected, with up to 30% of HLA-bound peptides best described as cis-spliced peptides.

Meanwhile, Delong *et al*. [7] used a focussed approach to identify potential autoantigens in T1D. By constructing a series of hybrid peptides containing insulin and other beta-cell antigen derived segments, they mapped the reactivity of the BDC 2.5 and related diabetogenic T cell clones. Importantly they demonstrated T cell reactivity to similar peptides in humans with T1D. Based on these observations, they have gone on in several subsequent studies to identify these HIPs' presence in beta cells by generating a database containing theoretical cis- and trans-spliced peptides derived from several β-cell secretory proteins. The theoretical database was combined with the reference proteome and used to analyse mass spectrometry data [7].

We developed a workflow based on incorporating *de novo* sequencing and database search [16]. In this workflow, after filtering out all the spectra that had a match to peptide sequences contained within the reference proteome, we used *de novo* sequencing to sequence the remaining high-quality spectra. Candidate sequences for which there was a high confidence assignment were then queried using an in-house developed algorithm (called Hybrid-Finder) to see if there was an explanation for cis and trans-splicing from the inferred peptide sequences(s) derived from *de novo* sequencing of these spectra. Cis- and trans-spliced candidate sequences were incorporated into the human proteome reference and the whole MS data set were searched against this merged database. Using this approach, we also found a high proportion of the *de novo* sequenced spectra were best explained as spliced peptides. Mylonas et al. [17] and Paes et al. [18] used a similar approach (*de novo* sequencing) and Rolfs et al. developed Neo-Fusion algorithm for identification of spliced peptides [19]. Using an *in silico*-*in vitro* approach, Mishto and colleagues identified a new reversed spliced peptide derived from gp100 [20], also a potential HLA-A*02:01 restricted spliced peptides which present KRAS G12V mutation [21].

Spliced peptide generation - mechanisms and features

It has been suggested that splicing is a stochastic process in the cell. However, we have recently found the repeated presentation of the same spliced peptides between biological replicates and even the same spliced peptides being presented by shared

HLA alleles from different individuals [22]. We also observed a similar prevalence of spliced peptide presentation in the presence or absence of IFN-γ treatment [22], which suggests that both constitutive and immunoproteasomes can generate spliced peptides. This indicates that understanding the rules behind the generation of spliced peptides is essential to predict which spliced peptides are likely to be generated and presented from an antigen or restricted to a given HLA allotype. Two approaches are being used for finding the rules governing peptide splicing: (i) *in vitro* proteasome-catalysed peptide splicing (PCPS) [23-26] and (ii) data-oriented approaches that analyse the MS data from HLA-I-eluted spliced peptides [12, 15, 16, 18, 22].

In the *in vitro* PCPS method, several short to intermediate-sized polypeptides are digested by purified proteasome/immunoproteasome preparations, and the resultant peptides analysed by mass spectrometry. These sequenced peptides result from the proteasomal digestion of longer peptide precursors (not intact or ubiquitinated protein antigens) typically by the 20S proteasome and may not necessarily reflect what that will be included in the immunopeptidome. However, these studies do recapitulate known spliced peptide antigens [23, 25] and can help inform more precisely the nature of spliced peptide junctions.

The data-oriented approach analyses peptides eluted from peptide HLA complexes purified from the surface of antigen presenting cells. The composition of these sequences is affected by several steps in the antigen processing and presentation pathway, such as transport into the ER by the transporter associated with antigen processing (TAP) molecule, trimming of these transported peptides by ERAPs, and most importantly, the selective force of the polymorphic HLA-I alleles that allow only a fraction of the transported peptides (spliced as well as linear) to dock into HLA-I binding grooves. Whilst this approach incorporates natural antigen degradation, the more complex input (the entire cellular proteome) can often make defining the precise splice junction difficult due to the many permutations and combinations of peptide segments that may contribute to the spliced peptide sequence [16]. Tus the combination of *in vitro* and *in cellulo/in vivo* approaches offer complementary and often synergistic information.

Table 1 summaries our current knowledge about splicing rules and features. These rules can be classified into three general areas of i) size, length, and concentration of the precursor antigen; ii) physicochemical properties of the precursor polypeptide; iii) the amino acid composition of the spliced entities, especially at the cleavage-ligation site.

Both the *in vitro* PCPS and data-oriented methods determined that antigen length and antigen concentration/abundance can increase the probability of peptide splicing [25, 27]. *In vitro* PCPS revealed that cis-spliced peptides have a narrower length distribution in comparison to linear peptides [23], while the data-oriented method showed an almost similar length distribution among the linear and spliced peptides. This could be because the data-oriented method analyses HLA-I bound peptides after they are trimmed and optimized for the HLA-I binding in the ER, which typically results in peptide ligands of 8 to 12 amino acid residues in length [16]. Paes *et al.* [23] showed that with increasing the size of the intervening sequence, the probability of both forward and reverse *cis*-splicing decreases; without being wholly abolished, though. However, this trend was not observed in data-oriented approaches [15, 22].

The process of splicing (transpeptidation) is depicted in Figure 2. Splicing happens as a parallel reaction to the normal hydrolysis; the acyl-enzyme (Acyl-Ez) intermediate is either hydrolysed by water or is attacked by another short peptide to form a spliced peptide (Figure 2 A). One factor that governs the second reaction's efficiency to go for the transpeptidation is the junction site's amino acid composition. With this regard, one thing is (almost) universally concluded by these studies. While the polarity and the charge of the side chains are of great importance, generally, the bulkier the side chains at positions P1 and P1' are, the lesser the efficiency of the PCPS will be. Using amino acid replacement in a unique naturally occurring spliced sequence, Berker et al. [25] examined the efficiency of *in vitro* PCPS for different amino acids in P1, P2, P1', and P2'. Namely, comparing Lue and Ile in P1, which only differ in the position of one methyl group in their side chain, Lue is almost twice as efficient as Ile in deriving PCPS. Re-evaluation of their results shows that Asp, Lue, and Ser are preferred over their closely chemically-related but bulkier amino acids Glu, Ile, and Thr, respectively (Figure 2 B). With a similar but slightly different approach, Specht et al. [26] performed large-scale *in vitro* PCPS on 55 unique polypeptide precursors and created an *in vitro* PCPS database. They evaluated the frequency of different amino acids in positions P1 to P4 and P1' to P4' of the spliced peptides. Although they show some disagreements with Berker's results, very similar conclusions can be drawn from this database. Asp, Lue, and Ser, for instance, are slightly over-represented compared to Glu, Ile, and Thr in the P1. A comparable trend is observed for the P1' residue, as Lue, for instance, is slightly preferred over Ile in both Berker's[25] and Specht's [26] works. Additionally, in our 2018 study [16], as well as Paes's study[23], an analogous trend for P1-P1' pairs of amino acids is observed, where smaller amino acids are enriched for both sites. Moreover, Figure 2C portrays the concord of deriving conditions for splicing in terms of species abundance, special hindrance at P1-P1', and the species length.

The role of spliced peptides in immune recognition

Several immunogenic spliced antigens have been discovered in the context of cancer, T1D, HIV, and bacterial infections. In total, six spliced peptides derived from cancer antigens have been discovered by mapping their specificities to tumour reactive T-cell clones or lines (Table 2). Five of these peptides are presented by the HLA-A3 subtype, and one peptide is presented by HLA-A24. We have recently used an immunopeptidomics approach and identified another six immunogenic spliced peptides derived from cancer antigens that were all restricted to HLA-A2 [22]. To understand the contribution of spliced peptides in the presentation of cancer associated antigens in the HLA-I immunopeptidome, we used list of spliced and linear peptides that reported in previous studies including three melanoma LM-Mel44[22], Mel22 and Mel35[28], one colon (HCT116) and one breast (HCC1143) carcinoma [12]. A total of 1608 linear and 1214 spliced peptides derived from these antigens have been reported in the immunopeptidomics data from these five cancer lines (Figure 3). Beyond cancer, in a recent study, a panel of cell lines expressing different HLA allotypes were infected with HIV-1 [18], and five cis-spliced peptides derived from Vif (2), Gag (2), and Pol (1) antigens were identified. One of the peptides was subsequently found to be immunogenic; however, cross-reactivity between the spliced peptide and the linear backbone of this peptide was observed in the T-cell studies [18]. Similarly, using a mouse-infection model, three spliced antigens were discovered from *Listeria monocytogenes* [13, 14]. The prevalence and importance of T1D related spliced peptides have been reviewed recently [29]. Table 2 shows discovered immunogenic spliced peptides to date.

The debate on the existence, prevalence, and importance of spliced peptides

The initial discovery of spliced peptides was met with awe and accolades and memorably explained in an excellent new and views article by a pioneer of the immunopeptidomics field Hans-Georg Rammensee [30]. However, despite corroboration and in-depth studies from Van den Eynde and colleagues [3-6, 20, 24, 31], the existence of spliced peptides was comfortably tucked away as an exception to the rule rather than a prevalent feature of the immunopeptidome. The relegation of spliced peptides to an immunological curiosity has been challenged in recent years, suggesting that they form a significant component of the immunopeptidome.

The peptides' controversial nature has been explored by the bioinformatics community, who have offered alternative explanations or dismissed their identification without offering a satisfactory alternative explanation for the spectra. Moreover, the fact that spliced peptides do not have a template in the genome brings up the question of how the immune system is tolerant to a potentially prevalent class of antigenic peptides. At least a part of the central tolerance to spliced peptides could be explained by the fact that the thymoproteasome, a form of the proteasome found in the thymus expressing a unique catalytic subunit, may also perform these splicing events for self-antigens [32].

Rather than add fuel to the fire here, we will dwell on why this controversy exists and the challenges of moving forward. Identification of spliced peptides using immunopeptidomics is in its infancy and using different software, pipelines, and more importantly, data interpretation will result in different observations. The range of prevalence of spliced peptides has been reported from 2-30% for cis [12-16, 18, 19, 22, 28] and 2-32% for trans spliced peptides [16, 19]. Immunopeptidomics data is very different from traditional proteomics data; it relies on peptide-centric identification from high-quality MS/MS data, often in the absence of corroborating peptides from the same protein [10]. As such, this can predispose the analysis to higher than normal false discovery rates and therefore requires more careful validation of the data. One solution to this is to retrospectively synthesise the peptides and show they have similar chromatographic behaviour and identical MS/MS spectra. This, however, is clearly not feasible when thousands of peptides are in question. Typically, when this is done for a subset of peptides, the matches are good, although this could be anticipated for high confidence *de novo* sequenced peptides for which most explanations would find a good MS/MS match. Perhaps the ultimate validation is the demonstration that these peptides are immunogenic; this has been the stalwart of the field pre-controversy and was recently demonstrated by us for spliced peptides derived from tumour antigens [22]. Recent improvements to peptide-centric identifications, including deep learning-based *de novo* sequencing algorithms [33, 34] and the incorporation of spectral prediction engines to spectral searching algorithms, will improve the authenticity of peptide identification in immunopeptidomics. Similarly, the accumulation of independent parameters such as collision cross sections in ion mobility mass spectrometers and incorporation of stable isotope labelled amino acids in cell culture to distinguish closely related amino acids (e.g., the isobaric Isoleucine and leucine residues) may aid future validation experiments.

Putative spliced sequences reported by Liepe and colleagues [15] and us [16] have attracted considerable criticism, with others preferring to describe these species as unanticipated or alternative reading frames, highly chemically post-translationally modified species, or no explanation given at all [19, 28, 35-38]. One of the debates is on other possible explanations for proposed spliced peptides. A proteogenomic study on HLA-I immunopeptidome reported around 5% of the immunopeptidome derived from non-canonical reading frames [39]. We also reported that around 13% of sequences assigned as spliced peptides (~5% of total immunopeptidome) could also be explained by other non-canonical protein sources such as those derived from a six-frame translation of mRNAs based on extensive highly replicated RNASeq data [40]. Recently, by using a ribosome profiling (Ribo-seq) method, it has been found that a small proportion of peptides that have previously been assigned as spliced peptides could also be explained by translation of novel unannotated open reading frames [38]. Irrespective of their source and origin, the validity of such sequences should not be in question.

Some debates in the field arise for inappropriate data analysis or interpretation. *In vitro* studies show cysteine residues are often enriched at the cleavage site of spliced peptides[23]. This could explain the higher prevalence of cysteine in the spliced peptides in comparison to linear peptides. Another study showed that the presence of cysteine was underestimated in peptide HLA binding models [41]. This could (at least partly) explains the reason of in general poorer predicted binding affinity of spliced peptides by current peptide-HLA binding models. This matter is exacerbated when re-analysis of immunopeptidomics data uses an inappropriate cysteine modification (such as carbamidomethylation) [19, 36] essentially ignoring cysteine containing spliced peptides. In a recent study, Erhad and his colleagues [36] used a *de novo* sequencing approach to identify cryptic peptides (peptides without a template in the reference proteome). In addition to assuming cysteine carbamidomethylation as a fixed modification in their analysis (which is not always the case) they also prioritise spectral assignment to cryptic species. In their approach, all top 10 *de novo* sequencing candidates for each spectrum were matched to each category with highest priority as following: Reference proteome> 5'-UTR > Off-Frame > Frameshift > 3'-UTR > ncRNA> Substitution > Intronic > Intergenic > spliced peptides). If a match was found, and all other hits among the ten *de novo* candidates were discarded. For spliced peptide databases, they used an incomplete list of just cis spliced peptides with a maximum distance of 25 amino acids between donor segments. Surprisingly, even they clearly mentioned the bias in their approach and positioned spliced peptides as the lowest priority (which obviously will end up to matching to low-quality spectra which don't have any other explanation). This then led to their somewhat erroneous conclusion that most *de novo* matches to spliced peptides were of poor and their dismissal of the high prevalence of spliced peptides in the immunopeptidome [36]. It is clear that a consensus on the methods to interrogate immunopeptidomics data is required rather than spliced peptides being unfairly dismissed despite strong evidence for their existence and relevance to immunity.

Conclusion

In the last two decades, our understanding of spliced peptides has transitioned from an immunological curiosity to serious consideration in studies of different peptide antigen classes. Several workflows have been developed for the identification of these antigens. Despite all the controversies surrounding the prevalence of these peptides, different technical approaches have confirmed their existence. Even with the lowest estimations of the contribution of spliced peptides to immunopeptidome, they still represent the dominant post-translational modification in the HLA-I immunopeptidome.

We have learnt that the prevalence of spliced peptides is altered across different HLA allotypes and can change under various physiological conditions. Several thousand spliced peptides have been identified in melanoma [15, 16], breast and colon cancer [12], Lymphoblastoid cell lines [15], and other antigen-presenting cells [16]. We know the same spliced peptides can be presented across individual cell lines and could be generated from self-proteins, cancer-specific antigens, and pathogen-derived proteins. However, there are still lots of gaps in our knowledge about spliced peptides. Bioinformatics pipelines are gradually improving, and cumulative data are building a clearer picture of the biochemical rules for spliced peptide generation. More studies are needed to determine which subunit(s) of the proteasome or even other cellular proteases are the primary drivers of splicing. The prevalence of recognition of and tolerance to spliced peptides by T cells is still a big hole in our knowledge.

In conclusion, evolving knowledge in this area indicates spliced peptides significantly diversify the composition of the immunopeptidome and, as a result, increase the possibility of immune recognition of cancer, infection, and self-tissues in autoimmunity.

Table1. Physicochemical rules and patterns for the generation of spliced peptides.

Rule Area	**Rule Description**	**Method [Ref]**
Size, length, and concentration	- The longer the antigen is, and the more abundant the antigen is, the higher the chance of PCPS will be.	Data-oriented[15], *in vitro* PCPS [23, 25]
	- The fragment that is making the nucleophilic attack must at least be 3aa in length.	*in vitro* PCPS [24]
	- Spliced and linear peptides eluted from p-MHC-I complexes have similar size distribution. (8 to 12 aa)	Data-oriented [15, 16]
	- Spliced and linear peptides generated by *in vitro* PCPS show different size distribution (linear $q_{90\%}$=21 aa and spliced $q_{90\%}$=13 aa).	*in vitro* PCPS [23]
	- Spliced and linear peptides generated by *in vitro* PCPS show similar size distribution.	*in vitro* PCPS [26]
	- Forward and reverse *cis*-spliced peptides have similar size distribution (8-15 aa).	*in vitro* PCPS [23]
	- No straightforward rule or pattern for intervening sequence.	Data-oriented [15]
	- Increasing the size of the intervening sequence decreases the chance of forward *cis*-splice more than it decreases the chance of reverse *cis*-spliced.	*in vitro* PCPS [6, 23]
	- Generally, the N-terminal fragment is shorter (2-5 aa), and the C-terminal fragment is usually longer.	*in vitro* PCPS [23]
Physicochemical properties	- Hydrophobicity of the antigen is correlated with its chance to give rise to PCPS.	Data-oriented[15, 16]
Amino acid composition	For forward *cis*-splicing junction ([P1][P1']): - [C/H/I][L] or [F][E] or [G][I] or [I/K][S]	*in vitro* PCPS [23]

Rule Area	Rule Description	Method [Ref]
	For reverse *cis*-splicing junction ([P1][P1’]): - [C/I/L/S][K] or [N][G] or [S][E/H/K/Q]	*in vitro* PCPS [23]
	For [P2 P1][P1' P2' P3']: - P1 cannot be C, H, K, P, R. - P1 is preferably a negatively charged (Asp) or an uncharged polar residue - P1 can be a hydrophobic residue + a small or polar or negatively charged residue at P2 - P1' is K/R (basic aa) - P2' and P3' preferably are hydrophobic residues and are not negatively charged residues. - These ligation rules are mostly valid when the precursor concentration is low. In higher concentrations, the ligation may occur more randomly.	*in vitro* PCPS [25]
	For [P1][P1']: - P1 is preferably G/A/S. - P1' is preferably G/A/S or I/L/V - If P1 is Pro, P1' is preferably I/L	Data-oriented [16]
	The MHC-I anchor positions of the peptides in the immunopeptidome are highly correlation between spliced peptides, almost regardless of the HLA allele.	Data-oriented [16]

Table 2. Immunogenic spliced peptides identified in cancer, infection and autoimmune diseases

Peptide sequence	**Splicing type**	**Source protein**	**Disease**	**MHC type**	**Antigen**	**Reference**
NTYASPRFK	Cis	FGF-5	Cancer	HLA-A3	Human	[2, 5, 6]
RTKQLYPEW	Cis	gp100	Cancer	HLA-A32	Human	[3, 5]
QLYPEWRTK	Cis	gp100	Cancer	HLA-A3	Human	[20]
RSYVPLAHR	Cis	gp100	Cancer	HLA-A3	Human	[42]
IYMDGTADFSF	Cis	Tyrosinase	Cancer	HLA-A24	Human	[5]
SLPRGTSTPK	Cis	SP110	Cancer	HLA-A3	Human	[4]
KLLILELHV	Cis	RNF213	Cancer	HLA-A2	Human	[22]
LILGLLTKV	Cis	MAGEC2	Cancer	HLA-A2	Human	[22]
LLLEALEQL	Cis	NCOA1	Cancer	HLA-A2	Human	[22]
LLSLLIPAL	Cis	VPS18	Cancer	HLA-A2	Human	[22]
LLSLLLPAI	Cis	NEK10	Cancer	HLA-A2	Human	[22]
ILSLILPAL	Cis	ATP1A1	Cancer	HLA-A2	Human	[22]
ISYAFYKL	Cis	PlcB	Bacterial infection	H2Kb	Human	[13]
FSDQLIHLY	Cis	Vif	Viral infection	HLA-A1	HIV	[18]
IHYAFYKL	Cis	PlcB	Bacterial infection	H2Kb	Listeria	[13]
SAYGRQVYL	Cis	LLO	Bacterial infection	H2Kb	Listeria	[14]
GQVELGGGGIVEQCC	Cis	INS	Autoimmune disease	HLA-DQ8?	Human	[43]
VALKLQVFL	Cis	IAPP	Autoimmune disease	HLA-A2	Human	[44]
LQTLALNAARDP	Trans	INS1/IAPP	Autoimmune disease	H2-IAg7	Mouse	[45]
LQTLALWSRMD	Trans	INS1/ CHGA	Autoimmune disease	H2-IAg7	Mouse	[46]
GQVELGGGNAVEVLK	Trans	INS/ IAPP2	Autoimmune disease	HLA-DQA1*03:01/DQB1*03:02; HLA-DRB1*04:01	Human	[7, 43]
GQVELGGGSSPETLI	Trans	INS/NP-Y	Autoimmune disease	HLA-DQ8	Human	[7, 43]
GQVELGGGTPIESHQ	Trans	INS/ IAPP	Autoimmune disease	HLA-DRB1*04:01	Human	[43]

References

1. Taleb NN. The black swan : the impact of the highly improbable. 2nd Edn. New York: Random House Trade Paperbacks, 2010.
2. Hanada K-I, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. Nature 2004; **427**:252-6.
3. Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S, van der Bruggen P, Boon T, Van den Eynde BJ. An antigenic peptide produced by peptide splicing in the proteasome. Science 2004; **304**:587-90.
4. Warren EH, Vigneron NJ, Gavin MA, Coulie PG, Stroobant V, Dalet A, Tykodi SS, Xuereb SM, Mito JK, Riddell SR, Van den Eynde BJ. 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, Hanada K-i, Yang JC, Rosenberg SA, Van den Eynde BJ. An antigenic peptide produced by reverse splicing and double asparagine deamidation. Proceedings of the National Academy of Sciences 2011; **108**:E323–E31.
6. 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.
7. Delong T, Wiles TA, Baker RL, Bradley B, Barbour G, Reisdorph R, Armstrong M, Powell RL, Reisdorph N, Kumar N, Elso CM, DeNicola M, Bottino R, Powers AC, Harlan DM, Kent SC, Mannering SI, Haskins K. Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion. Science 2016; **351**:711-4.
8. Vizcaino JA, Kubiniok P, Kovalchik KA, Ma Q, Duquette JD, Mongrain I, Deutsch EW, Peters B, Sette A, Sirois I, Caron E. The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. Mol Cell Proteomics 2020; **19**:31-49.
9. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. Nat Protoc 2019; **14**:1687-707.
10. Faridi P, Purcell AW, Croft NP. In Immunopeptidomics We Need a Sniper Instead of a Shotgun. Proteomics 2018; **18**:e1700464.
11. Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A, Kloetzel PM, Stumpf MP, Heck AJ, Mishto M. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science 2016; **354**:354-8.
12. 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.
13. Platteel ACM, Liepe J, Textoris-Taube K, Keller C, Henklein P, Schalkwijk HH, Cardoso R, Kloetzel PM, Mishto M, Sijts AJAM. Multi-level Strategy for Identifying Proteasome-Catalyzed Spliced Epitopes Targeted by CD8(+) T Cells during Bacterial Infection. Cell Reports 2017; **20**:1242-53.
14. Platteel ACM, Mishto M, Textoris-Taube K, Keller C, Liepe J, Busch DH, Kloetzel PM, Sijts AJAM. CD8(+) T cells of Listeria monocytogenes-infected mice recognize both linear and spliced proteasome products. European Journal of Immunology 2016; **46**:1109-18.
15. Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A, Kloetzel PM, Stumpf MPH, Heck AJR, Mishto M. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science 2016; **354**:354-8.
16. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, Ayala R, Song J, Gearing LJ, Hertzog PJ, Ternette N, Rossjohn J, Croft NP, Purcell AW. A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. Science Immunology 2018; **3**.
17. Mylonas R, Beer I, Iseli C, Chong C, Pak HS, Gfeller D, Coukos G, Xenarios L, Muller M, Bassani-Sternberg M. Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-1 Ligandome. Molecular & Cellular Proteomics 2018; **17**:2347-57.

18. Paes W, Leonov G, Partridge T, Chikata T, Murakoshi H, Frangou A, Brackenridge S, Nicastri A, Smith AG, Learn GH, Li Y, Parker R, Oka S, Pellegrino P, Williams I, Haynes BF, McMichael AJ, Shaw GM, Hahn BH, Takiguchi M, Ternette N, Borrow P. Contribution of proteasome-catalyzed peptide cis-splicing to viral targeting by CD8(+) T cells in HIV-1 infection. Proc Natl Acad Sci U S A 2019; **116**:24748-59.
19. Rolfs Z, Solntsev SK, Shortreed MR, Frey BL, Smith LM. Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion. Journal of Proteome Research 2019; **18**:349-58.
20. Ebstein F, Textoris-Taube K, Keller C, Golnik R, Vigneron N, Van den Eynde BJ, Schuler-Thurner B, Schadendorf D, Lorenz FKM, Uckert W, Urban S, Lehmann A, Albrecht-Koepke N, Janek K, Henklein P, Niewienda A, Kloetzel PM, Mishto M. Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. Sci Rep-Uk 2016; **6**.
21. Mishto M, Mansurkhodzhaev A, Ying G, Bitra A, Cordfunke RA, Henze S, Paul D, Sidney J, Urlaub H, Neefjes J, Sette A, Zajonc DM, Liepe J. 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.
22. Faridi P, Woods K, Ostrouska S, Deceneux C, Aranha R, Duscharla D, Wong SQ, Chen W, Ramarathinam S, Lim Kam Sian TCC, Croft NP, Li C, Ayala R, Cebon J, Purcell AW, Schittenhelm RB, Behren A. Spliced peptides and cytokine driven changes in the immunopeptidome of melanoma. Cancer Immunol Res 2020; **8**:1322-34.
23. Paes W, Leonov G, Partridge T, Nicastri A, Ternette N, Borrow P. Elucidation of the Signatures of Proteasome-Catalyzed Peptide Splicing. Front Immunol 2020; **11**.
24. Michaux A, Larrieu P, Stroobant V, Fonteneau JF, Jotereau F, Van den Eynde BJ, Moreau-Aubry A, Vigneron N. 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.
25. Berkers CR, de Jong A, Schuurman KG, Linnemann C, Meiring HD, Janssen L, Neefjes JJ, Schumacher TNM, Rodenko B, Ovaa H. Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. The Journal of Immunology 2015; **195**:4085-95.
26. Specht G, Roetschke HP, Mansurkhodzhaev A, Henklein P, Textoris-Taube K, Urlaub H, Mishto M, Liepe J. Large database for the analysis and prediction of spliced and non-spliced peptide generation by proteasomes. Scientific Data 2020; **7**:146.
27. Liepe J, Ovaa H, Mishto M. Why do proteases mess up with antigen presentation by re-shuffling antigen sequences? Current Opinion in Immunology 2018; **52**:81-6.
28. Mylonas R, Beer I, Iseli C, Chong C, Pak HS, Gfeller D, Coukos G, Xenarios I, Müller M, Bassani-Sternberg M. Estimating the contribution of proteasomal spliced peptides to the HLA-I ligandome. Molecular and Cellular Proteomics 2018; **17**:2347-57.
29. Wiles TA, Delong T. HIPs and HIP-reactive T cells. Clin Exp Immunol 2019.
30. Rammensee H-G. Protein surgery. Nature 2004; **427**:203-4.
31. Van den Eynde BJ, Gaugler B, Probst-Kepper M, Michaux L, Devuyst O, Lorge F, Weynants P, Boon T. A New Antigen Recognized by Cytolytic T Lymphocytes on a Human Kidney Tumor Results from Reverse Strand Transcription. The Journal of Experimental Medicine 1999; **190**:1793-800.
32. Kuckelkorn U, Stubler S, Textoris-Taube K, Kilian C, Niewienda A, Henklein P, Janek K, Stumpf MPH, Mishto M, Liepe J. Proteolytic dynamics of human 20S thymoproteasome. J Biol Chem 2019; **294**:7740-54.
33. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. Proceedings of the National Academy of Sciences 2017; **114**:8247-52.

34. Tran NH, Qiao R, Xin L, Chen X, Liu C, Zhang X, Shan B, Ghodsi A, Li M. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nature Methods 2019; **16**:63-6.
35. Rolfs Z, Müller 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". Science Immunology 2019; **4**:eaaw1622.
36. Erhard F, Dölken L, Schilling B, Schlosser A. Identification of the cryptic HLA-I immunopeptidome. Cancer Immunology Research 2020:canimm.0886.2019.
37. Faridi P, Li C, Ramarathinam SH, Illing PT, Mifsud NA, Ayala R, Song J, Gearing LJ, Croft NP, Purcell AW. Response to Comment on "A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands". Science Immunology 2019; **4**:eaaw8457.
38. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, Li B, Christian E, Knisbacher BA, Le PM, Hartigan CR, Keshishian H, Apffel A, Oliveira G, Zhang W, Chow YT, Ji Z, Jungreis I, Shukla SA, Bachireddy P, Kellis M, Getz G, Hacohen N, Keskin DB, Carr SA, Wu CJ, Regev A. Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. bioRxiv 2020:2020.02.12.945840.
39. Laumont CM, Daouda T, Laverdure JP, Bonneil E, Caron-Lizotte O, Hardy MP, Granados DP, Durette C, Lemieux S, Thibault P, Perreault C. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun 2016; **7**:10238.
40. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, Ayala R, Song J, Gearing LJ, Hertzog PJ, Ternette N, Rossjohn J, Croft NP, Purcell AW. A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. Sci Immunol 2018; **3**:eaar3947

41. Sachs A, Moore E, Kosaloglu-Yalcin Z, Peters B, Sidney J, Rosenberg SA, Robbins PF, Sette A. Impact of Cysteine Residues on MHC Binding Predictions and Recognition by Tumor-Reactive T Cells. J Immunol 2020.
42. Michaux A, Larrieu P, Stroobant V, Fonteneau JF, Jotereau F, Van den Eynde BJ, Moreau-Aubry A, Vigneron N. 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.
43. Babon JAB, DeNicola ME, Blodgett DM, Crevecoeur I, Buttrick TS, Maehr R, Bottino R, Naji A, Kaddis J, Elyaman W, James EA, Haliyur R, Brissova M, Overbergh L, Mathieu C, Delong T, Haskins K, Pugliese A, Campbell-Thompson M, Mathews C, Atkinson MA, Powers AC, Harlan DM, Kent SC. Analysis of self-antigen specificity of islet-infiltrating T cells from human donors with type 1 diabetes. Nature Medicine 2016; **22**:1482-+.
44. Gonzalez-Duque S, Azoury ME, Colli ML, Afonso G, Turatsinze JV, Nigi L, Lalanne AI, Sebastiani G, Carre A, Pinto S, Culina S, Corcos N, Bugliani M, Marchetti P, Armanet M, Diedisheim M, Kyewski B, Steinmetz LM, Buus S, You S, Dubois-Laforgue D, Larger E, Beressi JP, Bruno G, Dotta F, Scharfmann R, Eizirik DL, Verdier Y, Vinh J, Mallone R. Conventional and Neo-antigenic Peptides Presented by beta Cells Are Targeted by Circulating Naive CD8+T Cells in Type 1 Diabetic and Healthy Donors. Cell Metab 2018; **28**:946-+.
45. Wiles TA, Delong T, Baker RL, Bradley B, Barbour G, Powell RL, Reisdorph N, Haskins K. An insulin-IAPP hybrid peptide is an endogenous antigen for CD4 T cells in the non-obese diabetic mouse. Journal of Autoimmunity 2017; **78**:11-8.
46. Baker RL, Jamison BL, Wiles TA, Lindsay RS, Barbour G, Bradley B, Delong T, Friedman RS, Nakayama M, Haskins K. CD4 T cells reactive to hybrid insulin peptides are indicators of disease activity in the NOD mouse. Diabetes 2018; **67**:1836-46.

Fig1. The generation of linear, cis-spliced, reverse- cis spliced and trans-spliced peptides by proteasome and their subsequent presentation by HLA-I molecules.

Fig2. Peptide splicing rules. A) In proteolysis, a nucleophile (Thr1 of the beta subunits in proteasome) attacks the peptide bond and forms an acyl enzyme (Acyl-Ez) intermediate. The Acyl-Ez is either hydrolysed by a water molecule, or is attacked by another peptide fragment to make a spliced peptide. B) Visualization of how P1 and P1' side chain bulkiness hinders the splicing to happen. C) Different factors endorsed by several publications as transpeptidation-favouring conditions.

Fig3. The contribution of linear and spliced peptides in the presentation of cancer associated antigens in the HLA-I immunopeptidome. A total of 1608 linear and 1214 spliced HLA-I bound peptides derived from cancer associated antigens have been reported from melanoma, breast and colon cancer tumours. This figure represents cancer associated antigens that had at least 10 reported peptides.

Fig 1

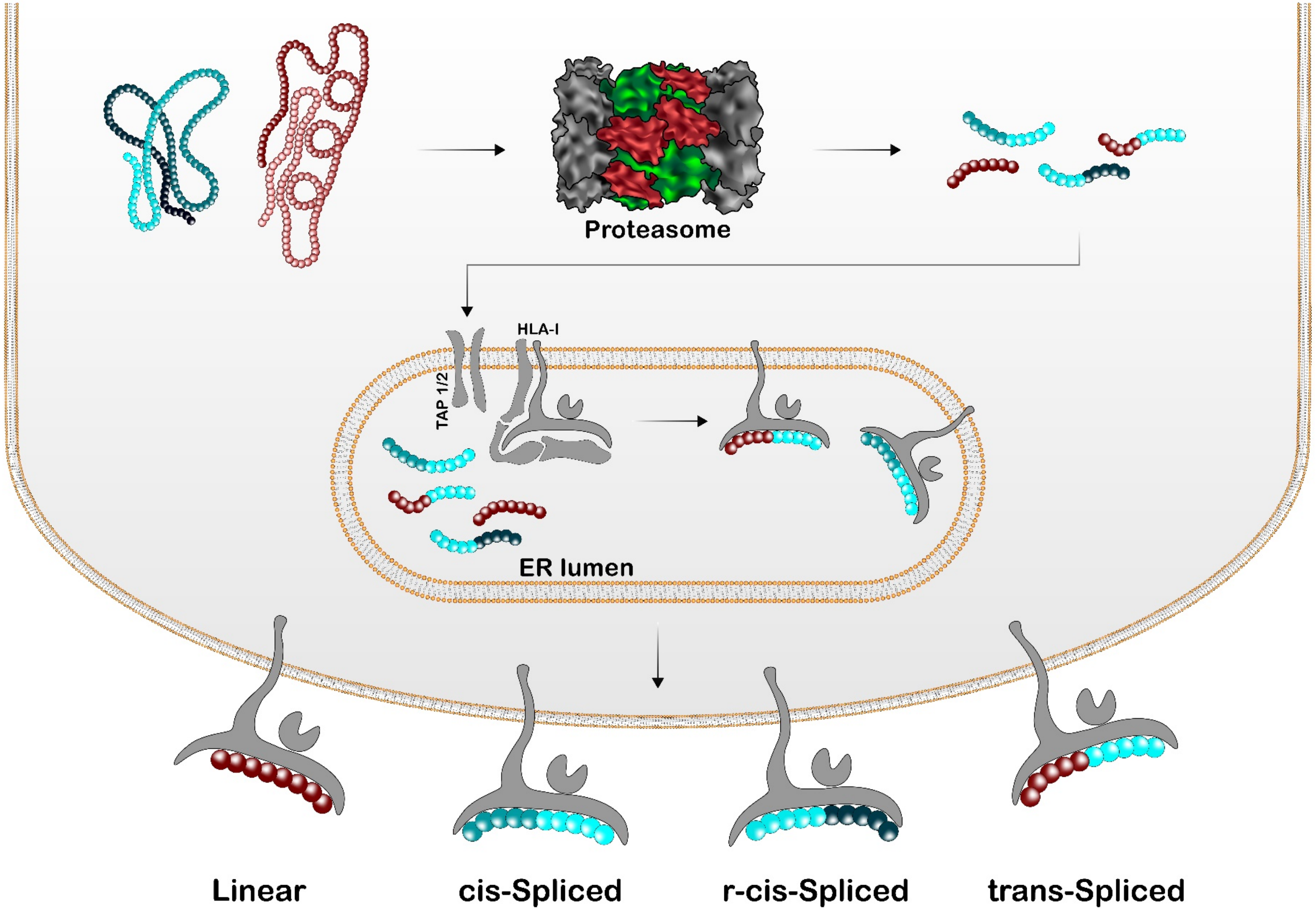

Proteasome
HLA-I
TAP 1/2
ER lumen
Linear
cis-Spliced
r-cis-Spliced
trans-Spliced

Fig 2

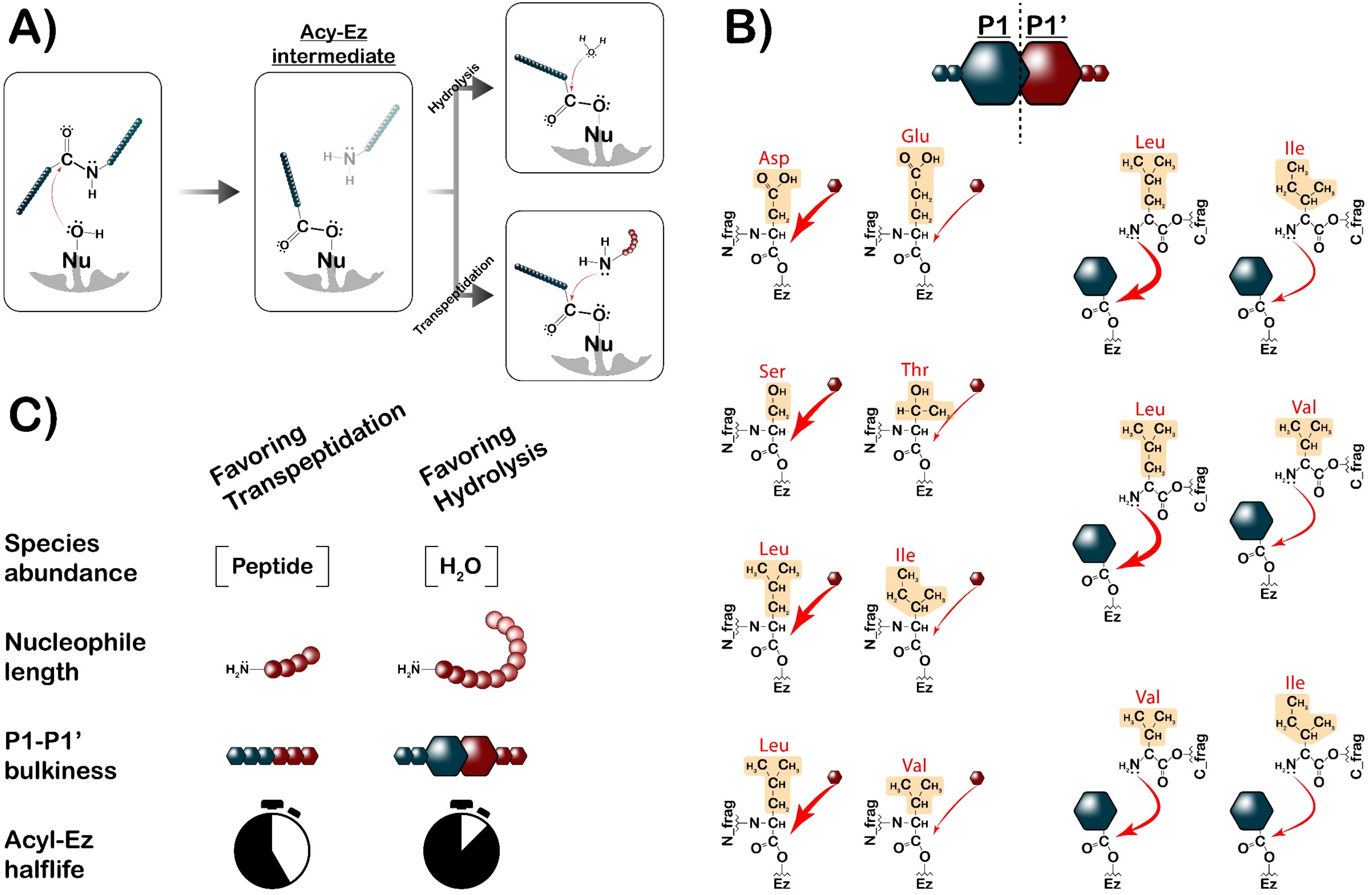
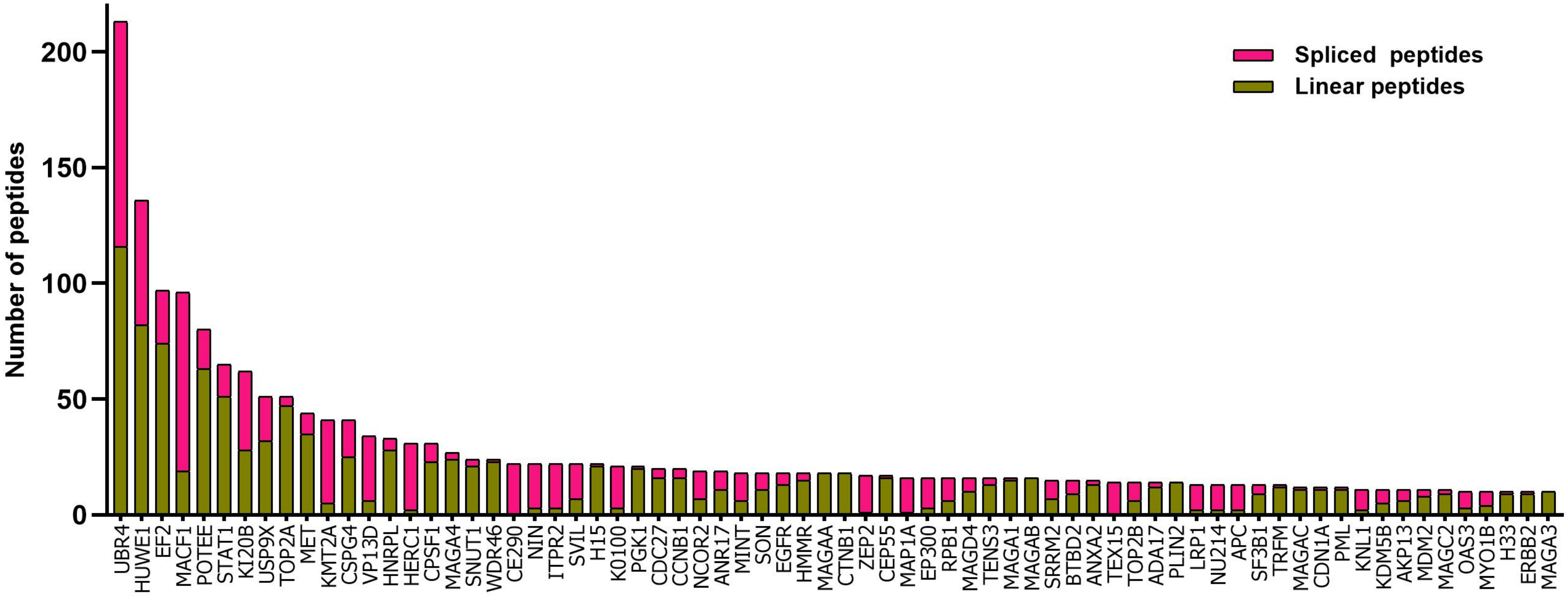

A)
Acy-Ez intermediate
Hydrolysis
Transpeptidation
Nu
B)
P1
P1'
Asp
Glu
Leu
Ile
Ser
Thr
Leu
Val
Leu
Ile
Leu
Val
Val
Ile
N_frag
C_frag
Ez
C)
Favoring Transpeptidation
Favoring Hydrolysis
Species abundance
Peptide
H2O
Nucleophile length
P1-P1' bulkiness
Acyl-Ez halflife

Fig 3


Spliced peptides
Linear peptides
Number of peptides
0
50
100
150
200
UBR4
HUWE1
EF2
MACF1
POTEE
STAT1
KI20B
USP9X
TOP2A
MET
KMT2A
CSPG4
VP13D
HNRPL
HERC1
CPSF1
MAGA4
SNUT1
WDR46
CE290
NIN
ITPR2
SVIL
H15
K0100
PGK1
CDC27
CCNB1
NCOR2
ANR17
MINT
SON
EGFR
HMMR
MAGAA
CTNB1
ZEP2
CEP55
MAP1A
EP300
RPB1
MAGD4
TENS3
MAGA1
MAGAB
SRRM2
BTBD2
ANXA2
TEX15
TOP2B
ADA17
PLIN2
LRP1
NU214
APC
SF3B1
TRFM
MAGAC
CDN1A
PML
KNL1
KDM5B
AKP13
MDM2
MAGC2
OAS3
MYO1B
H33
ERBB2
MAGA3