A subset of HLA-I peptides are not genomically templated : Evidence for cis- and trans-spliced peptide ligands

Faridi, Pouya, Li, Chen, Ramarathinam, Sri H., Vivian, Julian P., Illing, Patricia T., Mifsud, Nicole A., Ayala, Rochelle, Song, Jiangning, Gearing, Linden J., Hertzog, Paul J., Ternette, Nicola, Rossjohn, Jamie, Croft, Nathan P. and Purcell, Anthony W.

Faridi, P., Li, C., Ramarathinam, S. H., Vivian, J. P., Illing, P. T., Mifsud, N. A., Ayala, R., Song, J., Gearing, L. J., Hertzog, P. J., Ternette, N., Rossjohn, J., Croft, N. P. and Purcell, A. W. (2018). A subset of HLA-I peptides are not genomically templated : Evidence for cis- and trans-spliced peptide ligands. Science *Immunology*, *3*(28), Article eaar3947. https://doi.org/10.1126/sciimmunol.aar3947

A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands

Pouya Faridi[1], Chen Li[1,2], Sri H. Ramarathinam[1], Julian P. Vivian[1,3], Patricia T. Illing[1], Nicole A. Mifsud[1], Rochelle Ayala[1], Jiangning Song[1,4], Linden J. Gearing[5], Paul J. Hertzog[5], Nicola Ternette[6], Jamie Rossjohn[1,3,7], Nathan P. Croft[1], and Anthony W. Purcell[1],

[1]Infection and Immunity Program and Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia.
[2]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland.
[3]Australian Research Council Centre of Excellence in Advanced Molecular Imaging, Monash University, Clayton, Victoria 3800, Australia.
[4]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, Victoria 3800, Australia.
[5]Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research and Department of Molecular and Translational Science, School of Clinical Science, Monash University, Clayton, Victoria 3168, Australia.
[6]Jenner Institute, University of Oxford, Oxford OX3 7FZ, UK.
[7]Institute of Infection and Immunity, Cardiff University School of Medicine, Heath Park, Cardiff CF14 4XN, UK.

Stitching peptides for presentation

Intracellular protein–derived peptides generated by proteasomal degradation are loaded on to class I MHC molecules in the endoplasmic reticulum and presented to $CD8^+$ T cells. Although it has been assumed that these peptides are contiguous segments derived from intracellular proteins, recent studies have shown that noncontiguous peptides generated by cis-splicing of two distinct regions of an antigen can be presented by class I MHC molecules. Here, Faridi *et al.* demonstrate that class I MHC molecules can present peptides that are generated by splicing together of segments from two distinct proteins and term them to be “trans-spliced” peptides. Precisely how cis- and trans-spliced peptides are generated and how they contribute to T cell selection and expansion remain to be explored.

Abstract

The diversity of peptides displayed by class I human leukocyte antigen (HLA) plays an essential role in T cell immunity. The peptide repertoire is extended by various posttranslational modifications, including proteasomal splicing of peptide fragments from distinct regions of an antigen to form nongenomically templated cis-spliced sequences. Previously, it has been suggested that a fraction of the immunopeptidome constitutes such cis-spliced peptides; however, because of computational limitations, it has not been possible to assess whether trans-spliced peptides (i.e., the fusion of peptide segments from distinct antigens) are also bound and presented by HLA molecules, and if so, in what proportion. Here, we have developed and applied a bioinformatic workflow and demonstrated that trans-spliced peptides are presented by HLA-I, and their abundance challenges current models of proteasomal splicing that predict cis-splicing as the most probable outcome. These trans-spliced

peptides display canonical HLA-binding sequence features and are as frequently identified as cis-spliced peptides found bound to a number of different HLA-A and HLA-B allotypes. Structural analysis reveals that the junction between spliced peptides is highly solvent exposed and likely to participate in T cell receptor interactions. These results highlight the unanticipated diversity of the immunopeptidome and have important implications for autoimmunity, vaccine design, and immunotherapy.

INTRODUCTION

The proteins encoded by the major histocompatibility complex (MHC) [human leukocyte antigen class I (HLA-I) molecules in humans] play a critical role in adaptive immunity by binding to intracellular peptide antigens and presenting them on the surface of cells for recognition by $CD8^+$T cells [(1)]. The canonical mechanism for HLA-I–bound peptide (*p*-HLA) production is degradation of intracellular proteins by the proteasome, which generates peptides with lengths between 2 and 22 amino acids [(2)]. These antigenic peptide precursors are subsequently transported into the endoplasmic reticulum via the transporter associated with antigen-processing (TAP) molecule. Further peptidase trimming generates peptides typically between 8 and 12 amino acids in length that bind to nascent HLA-I molecules before being exported to the cell surface. Posttranslational modification (PTM) of *p*-HLA can profoundly influence the composition of the immunopeptidome and T cell recognition [(3, 4)]. For instance, although considered to be a rare event [(5–7)], recent studies have shown that proteasomes can also ligate distinct peptide fragments (termed here spliced peptides), producing sequences that are noncontiguous and therefore not linearly templated in the genome [(5, 6, 8)]. The origin of spliced peptide segments can be from the same protein (cis-splicing) or different proteins (trans-splicing) (Figure 1A). Although the ability of HLA-I to present linear peptides is well known [(5)], the potential for binding of cis- and trans-spliced peptides has received limited investigation. Liepe *et al.* [(9)] recently reported that a substantial (~30%) fraction of *p*-HLA are short-distance cis-spliced peptides. However, although trans-spliced peptides have been reported for HLA-II, and the generation (with similar efficiency as cis-spliced peptides) of several HLA-I trans-spliced peptides has been shown to occur in vitro and in vivo, the relevance of trans-spliced peptides in HLA-I immunopeptidomes has not yet been determined [(4, 10)].

The most comprehensive method for investigating HLA-I immunopeptidomes is through immunoaffinity purification of *p*-HLA-I and subsequent sequencing of the bound peptides by tandem liquid chromatography–mass spectrometry (LC-MS/MS). The data acquired by this approach are then interrogated using algorithms that rely on reference proteome or proteogenomic databases for spectral

matching [11, 12]. Although this method is successful for identifying linear *p*-HLA-I, the absence of sequence information for nonlinear peptides in the predicted proteome precludes the use of this workflow for the identification of spliced peptides. As a means to overcome this, Liepe *et al.* [9] generated a theoretical database containing proximal (donor segments within 20 amino acids) cis-spliced peptides for searches of nonlinear peptide antigens. However, considering the additional complexity introduced by peptides that are generated via trans-splicing and the possibility of more distally cis-spliced peptides, similar algorithms to account for these peptides would generate extremely large databases that make computational analyses impractical. One substantial barrier to comprehensively identifying trans-peptides is the availability of computational resources required to process all permutations of trans- and cis-spliced peptides [11]. Here, we have developed a bioinformatics workflow to identify *p*-HLA and discriminate between linear and spliced peptides. We have used this workflow to analyze MS data acquired from *p*-HLA purified from a multitude of monoallelic cell lines and show that both cis- and trans-spliced peptides contribute to the *p*-HLA landscape.

RESULTS

To identify spliced peptides from eluted *p*-HLA repertoires, we leveraged the capabilities of de novo peptide sequencing of high-quality MS/MS spectra combined with database searching [13] and an in-house developed algorithm ("hybrid finder") for hierarchical source protein identification. Briefly (see Figure 1B and Figure S1 and S2 for a schema), we first identified linear *p*-HLA by the PEAKS Studio 8.5 software (a de novo–based peptide library search algorithm), matching against the human reference proteome at a 1% false discovery rate (FDR) threshold. We reasoned that the remaining unidentified de novo sequences (themselves derived from high-quality MS/MS spectra) would constitute any of the following: (i) true linear sequences that fell below the stringent 1% FDR cutoff applied in the above database search; (ii) potential cis- or trans-spliced peptides; or (iii) untemplated peptides with no biological explanation at this stage or whose de novo sequencing was not of high enough accuracy. We therefore extracted these high-confidence de novo candidates (set at a maximum of five sequence assignments per spectrum) and processed each with our hybrid finder algorithm, which assigned a possible explanation as above. To make the most conservative estimate possible, these results were then ranked by likelihood, our rationale being that a linear (i.e., proteome-matched) explanation takes precedence over cis-splicing, which in turn takes precedence over trans-splicing. If no explanation could be found, then the sequence was discarded. This resulted in a single assigned sequence per spectrum, which we then built into a proteome-like FASTA database, merging it with the human reference

proteome and thus generating a combined sample-specific database. After this, all the mass spectra were researched against the combined library, and peptides were identified at a 1% FDR cutoff.

Spliced and linear HLA-I peptides share similar overall sequence features

We applied this data-driven workflow to MS data acquired from *p*-HLA purified from 17 different monoallelic cell lines, comprising expression of eight and nine different HLA-A and HLA-B alleles, respectively. We used monoallelic [14] cell lines to overcome the ambiguity in *p*-HLA motifs that may arise from the coexpression of multiple HLA alleles. In total, we have identified more than 50,000 *p*-HLA peptides (range of 978 to 11,110 peptides and median of 2781 peptides per HLA allotype; Figure 2, A and B). Although 38,345 (~72%) of these peptides could be mapped to sequences within the human proteome, a substantial (28%) fraction of the data was found to be best explained by peptide splicing (note that less than 3% of all de novo candidates could not be mapped to any splicing explanation). When assessing individual alleles, we observed a range of 12.6% (HLA-A*24:02) to 44.7% (HLA-B*15:02) (HLA-B*51:01 median, 24.6%) of spliced peptides. A high proportion of these spliced peptides could only be explained by a reaction in trans, with this pattern observed for both HLA-A and HLA-B peptidomes (Figure 2C).

Given the two main factors that affect peptide binding to HLA are sequence length (typically 8 to 12 amino acids for class I) and HLA-binding amino acid motifs (allele-specific), we next assessed the degree to which these peptide properties were maintained across linear and spliced peptides for any given allele (Figure 2, D and E). We found that spliced peptides also conformed to an 8- to 12–amino acid length profile; however, their lengths were skewed toward more 10-mers and fewer 9-mers [$P < 0.0001$, two-way analysis of variance (ANOVA) multiple-comparison test] in comparison to linear peptides (Figure 2D).

Next, we compared HLA-binding motifs in spliced and linear peptides in each dataset through statistics-based visualization using iceLogo [15, 16] (Figure 2E and Figure S3), examining amino acid enrichment at each position of the 9-mer and 10-mer peptides. For all HLA allotypes in this dataset, the major anchor positions were at position 2 (P2) and/or P3, as well as PΩ (the C-terminal anchor residue). Across all alleles, we found that the particular amino acid frequency at the PΩ position correlated strongly ($r > 0.9$, Pearson test) between spliced and linear peptides (Figure 2F). At the P2/P3 position, we noted more variance in concordance between spliced and linear peptides—for example, HLA-B*27:05 and HLA-

B*15:02 showed the weakest correlation at P2 (r = 0.4863) and P3 (r = 0.5388), respectively (Figure 2F). Nonetheless, the correlation was statistically significant ($P < 0.05$, Pearson test) across all alleles.

To examine this potential impact on HLA-binding affinity, we used NetMHC 4 or NetMHCcons to predict in silico binding affinity of both linear and spliced *p*-HLA (9-mers and 10-mers) for their corresponding HLA allomorph. We found that, on average, 77.3% of linear peptides were predicted to bind to their corresponded HLA, whereas for spliced peptides this value was significantly ($P < 0.0001$, Wilcoxon test) lower (46.86%; Figure. S4).

Experimental validation confirms spliced peptide sequence authenticity and p-HLA binding

To validate the authenticity of identified spliced peptides, we used data from the C1R-B*57:01 immunopeptidome. We selected this dataset because it contained the greatest number of overall peptides (>10,000) and because this allomorph has strict and distinctive peptide binding characteristics. One possible explanation for the existence of *p*-HLAs that are unable to be mapped to a reference proteome is that either the reference proteome does not account for all known translated transcripts or the genomic background of the cell line (in this case, the B-lymphoblastoid cell line C1R from which many of our monoallelic datasets are derived) bears nonsynonymous mutations or expresses unanticipated transcripts. To address the first possibility, we searched all B*57:01 peptides against the Ensembl protein database (converting all Ile to Leu and including all ab initio predicted peptides within this database) and matched 99.9% of linear peptides but just 0.6% of spliced peptides. Then, for the second possibility, we carried out detailed RNA-sequencing (RNA-seq) analysis of the C1R-B*57:01 (as well as parental C1R) cell lines and assessed whether any mutations or unanticipated transcripts could give rise to peptide sequences that we attributed to spliced peptides. From all RNA-seq reads, we carried out six-frame translations, and (after removing redundancy associated with isobaric amino acids and converting all Ile to Leu) we searched for any occurrences of our HLA-B*57:01 linear and spliced peptides. This analysis showed that, although 98.7% of all linear peptides could be matched to the RNA-seq data, only 12.7% of spliced peptides could be found (Table S1). Thus, even with the most far-fetched transcript explanations from this in-depth immunopeptidogenomics approach, the vast majority of spliced peptides cannot be templated to the transcriptome.

As an additional verification of the integrity of the de novo–based sequence assignment, we selected 28 identified HLA-B*57:01 spliced peptides for synthesis and compared the LC–MS/MS spectra of the synthetic to the corresponding original *p*-HLA–eluted spectra, computing their correlation score

and *P* value for similarity assessment [17]. All eluted peptides matched significantly ($P < 0.05$) to their synthetic peptide counterparts (see Figure 3, A and B; Figure S5; and Table S2), consistent with other studies that have shown the utility of de novo–based sequence assignments [13, 18–21] and highlight the accuracy of de novo sequencing within the present study. After this, we sought to validate our de novo sequencing approach for *p*-HLA genomically untemplated peptide identification on several levels (Figure S2), as detailed in Materials and methods and in the extensive Supplementary Materials. This included finding minimal (<1%) spliced peptide detection in complex proteolytic digests generated by trypsin or elastase digests of mammalian cellular proteomes, eliminating a role for common amino acid PTM in false assignment of spectra, and testing for any bias in database and search algorithms by using several search engines for the library-based searches. We have also precluded potential contamination from bovine peptides contained within the culture media and potential viral and retroviral sources of the parental antigen.

To further confirm that the spliced peptides identified from the repertoire of HLA-B*57:01 could bind and be presented by this allele, we selected a panel of peptides for in vitro stabilization binding assays (using the TAP-deficient cell line T2-B*57:01; Figure 3C) and for refolding of peptide B*57:01 complexes for determination of crystal structures (Figure 3D). Of the five (four trans- and one cis-) spliced peptides that were tested for in vitro stabilization, four were found to bind HLA-B*57:01 with similar capability to a known B*57:01-restricted linear peptide (Figure 3C). One peptide, trans-spliced LSDSTARDVTW, was not observed to stabilize B*57:01 in this assay.

For crystallization of *p*-HLA complexes, our choice of four spliced peptides included those with defined ligation site possibilities: cis-LALLTG + VRW, trans-LSDSTA + RDVTW, and trans-TSMSF + VPRPW each had only one possible ligation site, whereas the trans-peptide GSFDYSGVHLW could be spliced as either GSFDY + SGVHLW or GSFDYS + GVHLW. Crystal structures were determined to a resolution of between 2.04 and 1.83 Å (structure statistics summarized in Table S3), with the data permitting the visualization of the four *p*-HLA structures and, with the exception of the LSDSTARDVTW peptide, the junction of peptide splicing. That is, all peptide residues including the junction points of GSFDYSGVHLW, LALLTGVRW, and TSMSFVPRPW were resolved in the density; however, for LSDSTARDVTW (the same peptide that was not found to detectably stabilize T2-B*57:01 in Figure 3C), the Ala6-Arg7 splice junction was disordered, and we interpret this to reflect a high degree of flexibility in this region (Figure 3D). Analysis of these four novel structures and comparison with previously determined HLA-B*57:01 complexes [Protein Data Bank (PDB) accession codes: 2RFX [22], 3UPR [23], 3VRI [24], and 5T6Y [25]] showed

a high degree of structural conservation, with root mean square deviations between 0.21 and 0.36 Å (over Cα positions of their peptide-binding clefts amino acids 1 to 177). Further, each spliced peptide comprised canonical anchor residues at P2 (Ala/Ser) and PΩ (Trp) that were accommodated in the B and F pockets of HLA-B*57:01 in an orthodox fashion (Figure 3D). Accordingly, all four spliced peptides were bound to and presented by HLA-B*57:01 in the canonical manner.

To understand whether the overall level of expression of HLA-B*57:01 spliced peptides differed from that of linear peptides, we used a label-free quantification approach in measuring the distribution of peak areas from the mass spectra data. Consistent with the report of Liepe *et al.* [9] on the relative abundance of cis-spliced peptides, spliced peptides were present at lower abundance than linear peptides (Figure S6). Specifically, spliced peptides accounted for 16.7% of the abundance of linear peptides in the immunopeptidome of C1R-B*57:01.

Donor segment length and amino acid pairing influence transpeptidation

To understand the potential sequence preference for peptide splicing, we examined the nature of the donor peptide segments forming each spliced peptide to determine whether there was a bias underlying the ligation position and/or flanking residues of donor peptide segments (Figure 4). Because there are different segment length combinations and different donor proteins that may comprise any given spliced peptide, to aid this analysis, we first extracted all spliced peptides from all 17 datasets that had only one possibility for ligation (*n* = 3029 "unique" spliced peptides). From these peptides, we analyzed the amino acid pairing at the P1 (C terminus of the N-terminal segment) and P1' (N terminus of the C-terminal segment) positions (Figure 4A; see also Figure S7 comparing the $\log_2$ ratio between observed and expected amino acid pairs from the human proteome). These data were compared with adjacent amino acids located across the equivalent central position of linear peptides (*n* = 31,297) and also with a set of 31,000 amino acid pairs derived from randomly generated peptides whose amino acid frequency was computed to resemble that of the natural human proteome. Data show that linear peptides exhibit a similar distribution to these randomly derived pairs, which largely reflects the natural amino acid frequency (note that all Ile has been substituted for Leu in this analysis to account for redundancy in the de novo candidate sequences; Figure 4A). However, an analysis of the spliced peptide junction shows a markedly different distribution, with notable enrichment in small nonpolar residues (Gly/Ala/Ser) pairing either with the same residues or with hydrophobic Ile/Leu/Val in either orientation. Pro-Ile/Leu pairing was also of note but was only observed to be enriched in a P1-P1' direction.

Although this more limited (3029) subset of peptides allowed for an analysis of the splicing junction by virtue of only a single explanation of segment pairing per peptide, we next sought to determine how many donor sites might contribute to peptide splicing. That is, for any given spliced peptide, there may be multiple proteins that can supply a donor segment (e.g., a 9-mer peptide can be broken into 1 + 8, 2 + 7, 3 + 6, and so on segments, and thus there may be a multitude of ways of making the same peptide). To address this question, we took all unique spliced nonamers from our datasets (5806 peptides), segmented each peptide into all possibilities, and counted the number of occurrences of each segment from the reference human proteome (Figure 4B). As expected, as segment length increased, the number of possible sources decreased, with a median of just two sources once a segment length of six amino acids was reached. Using these data, it was therefore possible to compute the number of permutations of generating a spliced peptide by multiplying the number of occurrences of one segment with its corresponding pair (e.g., for a nonamer, a segment length of 2 has to be paired with its corresponding segment length of 7). Most of the peptides contribute to the 4 + 5 or 5 + 4 segment pairing (Figure 4C). However, it is notable that, although the occurrences of 1 + 8 or 8 + 1 pairing were rarer across this set of peptides, their combined permutations are ultimately higher because of the overwhelmingly large number of single amino acid sources that can pair with the segment of eight amino acids (with a similar situation being true for 2 + 7, 7 + 2, and so on, until the 4 + 5/5 + 4 trough is reached).

Thus, collectively, we observe a significant proportion of nonlinearly encoded peptides that contribute to the immunopeptidome of a number of HLA-A and HLA-B allotypes. We propose, as others have, that these peptides are generated through a reverse proteolytic mechanism, which may include the recently reported proteasomal catalyzed peptide splicing events [4, 9, 26]. Their existence in the immunopeptidome has profound implications for immunity and will be the subject of future research.

DISCUSSION

We have developed a workflow for the comprehensive identification of spliced *p*-HLA ligands, and, as a result, we report the considerable contribution of trans-spliced peptides, as well as both proximal and distal cis-spliced peptides, to the immunopeptidome. The unanticipated proportion of trans-spliced peptides reveals additional complexity of the immunopeptidome to that which has recently been documented [9]. Comparison of peptide-binding motifs across the 17 different alleles tested shows that, for any given allele, linear and spliced peptides share highly similar PΩ preferences and similar enrichment in P2/P3 anchor residues. Given that selection of peptide binding occurs downstream of

proteasomal processing [notwithstanding trimming by enzymes such as endoplasmic reticulum aminopeptidase (ERAP)], it is not unexpected that spliced peptides share similar binding motifs to linear peptides, but the discrepancies observed at P2/P3 (notable examples being that of HLA-B*27:05 at P2 and HLA-B*07:02 at P3) may perhaps be accounted for by splicing selection/ligation constraints put onto residues proximal to the splicing junction. Such subtle differences highlight the requirement to train binding prediction algorithms for this class of peptides.

One of the assumptions made about the genesis of trans-spliced peptides is that the proteasome must accommodate and process multiple polypeptide chains simultaneously [27]—the probability of two distinct protein substrates being degraded at the same time inside one proteasome is low [10]. However, we propose that, for trans-spliced peptides [as shown before in cis-spliced peptides [28]], it is not necessary that each of two different segments always originates from a particular position of a protein. For short segments of spliced peptides, the proteasome could use identical polypeptides generated from a multitude of donor proteins. For instance, for a 6 + 3 or 3 + 6 model of trans-spliced peptide generation, on average, around 7360 locations in the proteome could donate any given three–amino acid segment. Thus, although individual trans-spliced reactions may be rare, the high abundance of "trans-donors" may make this reaction more likely [29].

It should be noted that, for the present datasets, at maximum less than 3% of high-quality de novo sequences remained unassigned as spliced and therefore remain uncharacterized. The evolution of posttranslational proteasomal splicing and the impact on host immunity have yet to be fully determined. Thus, despite a number of initial reports of immunogenicity [5–7], the true physiological relevance of such prevalent peptides has yet to be comprehensively demonstrated. A high frequency of spliced peptides can increase the diversity of target antigens for T cell recognition (Figure 5). For instance, it may be that, for antigens that are highly susceptible to proteasomal degradation (where most cleavage products are too short to generate HLA-I ligands), the ligation of short oligopeptides may allow immunosurveillance of that antigen through the production of cis- or trans-spliced peptides. In the context of infectious immunity, this process may generate novel pathogen-derived peptides and enhance the breadth of the immune response. This may be particularly important for pathogens with small genomes that may not encode significant numbers of suitable HLA-I ligands. Presumably, under this circumstance, ligation of pathogen-derived peptide fragments with other antigens may create better targets for immunity, thus providing an advantage to the host. In contrast, this may also reveal an Achilles' heel of the immune system, facilitating the generation of cross-reactive T cells due to molecular mimicry of pathogen-

derived and self-peptides [30]. We found that only 235 (less than 1%) spliced peptides from all datasets have an exact sequence match in all proteome sequences stored in "NCBI RefSeq Non-redundant Proteins." Therefore, our findings suggest that peptide splicing does not necessarily predispose individuals to autoreactive responses upon encountering microorganisms. However, the possibility of pathogen-derived spliced peptides has also been reported and may contribute to equal proportions of the pathogen-derived immunopeptidome as those determined for self-derived peptides. Thus, the full repertoire of potential mimics, between both self-spliced/pathogen-linear and self-linear/pathogen-spliced, may be greater than was considered in this analysis.

With the implementation and application of this workflow, we have demonstrated the unanticipated abundance of trans-spliced peptides in the HLA class I peptidome. As more examples become apparent, we anticipate that the precise mechanism and underlying rules of peptide ligation will be systematically delineated in cellulo, leading to the generation of models to predict spliced peptides. Although we found subtle preferences at the junctional amino acids, no obvious splicing rules were apparent across our datasets, possibly reflecting the broad specificity of the various forms of proteasome potentially found in these cell lines (constitutive and immunoproteasomes, as well as mixed complexes). Thus, incorporating spliced peptides into the models of antigen presentation will broaden our understanding of T cell immunity while having implications in the context of immunotherapeutics, such as peptide vaccines, and having the potential to reinvigorate the search for autoimmune triggers.

MATERIALS AND METHODS

Cell culture and isolation of p-HLA complexes

Monoallelic cell lines were generated from C1R cells transfected with HLA alleles of interest and include C1R-A*01:01, C1R-B*07:02, C1R-B*08:01, C1R-B*15:02, C1R-B*18:01, C1R-B*27:05, C1R-B*57:01, C1R-B*57:03, and C1R-B*58:01 [24, 31–33]. These cells were grown to high density in RPMI 1640 media supplemented with 10% fetal calf serum, 7.5 mM Hepes, streptomycin (150 µg/ml), benzylpenicillin (150 U/ml), 2 mM L-glutamine (MP Biomedicals), 76 µM β-mercaptoethylamine, and 150 µM nonessential amino acids. Cells were tested for mycoplasma contamination in-house at regular intervals. Cells were harvested by centrifugation (1200*g*, 20 min, 4°C) and snap-frozen in liquid nitrogen. Clarified lysates were generated from cells with a combination of cryogenic milling and detergent-based lysis. HLA-peptide complexes were immunoaffinity-purified from cell lysates using the W6/32 monoclonal antibody in solid phase as described previously [34]. Bound complexes were eluted by acidification with

10% acetic acid and fractionated in a 4.6-mm (internal diameter) by 100-mm (length) monolithic reversed-phase C18 high-performance liquid chromatography (HPLC) column (Chromolith SpeedROD; Merck Millipore, Darmstadt, Germany) using an ÄKTAmicro HPLC system (GE Healthcare, Little Chalfont, United Kingdom). The mobile phase consisted of buffer A (0.1% trifluoroacetic acid; Thermo Fisher Scientific) and buffer B (80% acetonitrile and 0.1% trifluoroacetic acid; Thermo Fisher Scientific). HLA-peptide mixtures were loaded onto the column and separated using the following chromatographic conditions: 2 to 15% buffer B for 0.25 min (2 ml/min), 15 to 30% buffer B for 4 min (2 ml/min), 30 to 40% buffer B for 8 min (2 ml/min), 40 to 45% buffer B for 10 min (2 ml/min), 45 to 99% buffer B for 2 min (1 ml/min), and 99 to 100% for 2 min (1 ml/min), re-equilibrate 6 min in 2% buffer B at 2 ml/min. Fractions (500 µl) were collected, concatanated into 10 to 15 pools before vacuum-concentrated to 10 µl, and diluted in 0.1% formic acid to reduce the acetonitrile concentration. For the other alleles, we have used the publicly available data for monoallelic HLA-I cell lines [(35)].

LC–MS/MS sequencing of p-HLA–bound peptides

For LC-MS/MS acquisition, peptide-containing fractions were loaded onto a microfluidic trap column packed with ChromXP C18-CL 3-µm particles (300-Å nominal pore size; equilibrated in 0.1% formic acid, 2% acetonitrile) at 5 µl/min with a NanoUltra cHiPLC system (Eksigent). An analytical (75 µm × 15 cm ChromXP C18-CL, 3 µm, 120 Å; Eksigent) microfluidic column was switched in line, and peptides were separated by linear gradient elution with 0 to 30% buffer B (80% acetonitrile, 0.1% formic acid) over 50 min and 30 to 80% over 5 min flowing at 300 nl/min. Separated peptides were analyzed with a SCIEX TripleTOF 5600^{+} mass spectrometer equipped with a Nanospray III ion source and accumulating up to 20 MS/MS spectra per second. The following instrument parameters were used: ion spray voltage, 2400 V; curtain gas, 25 l/min; ion source gas, 10 l/min; and interface heater temperature, 150°C. MS/MS switch criteria included the following: ions of mass/charge ratio >200 amu; charge state, +2 to +5; and intensity, >40 counts per second. The top 20 ions meeting these criteria were selected for MS/MS per cycle. We calibrated the instrument every four LC runs using [Glu1]-Fibrinopeptide B standard.

De novo sequencing algorithm evaluation

The accuracy of PEAKS Studio 8.5 de novo sequencing algorithm has been previously described [(36)]. Nevertheless, we sought to evaluate the accuracy of this de novo algorithm in the context of *p*-HLA peptides. Therefore, we mixed 289 synthetic nontryptic peptides with lengths typical of *p*-HLA and analyzed them under identical conditions by LC-MS/MS (as mentioned above). We then used PEAKS de

novo [the parent mass error tolerance was set to 15 parts per million (ppm) and the fragment mass error tolerance to 0.1 Da] and allowed the algorithm to generate the top 10 candidates for each spectrum. Of the total 289 peptides, 220 peptides were identified by PEAKS de novo sequencing alone (Figure S3A). By using PEAKS DB ("database"; 1% FDR cutoff), we identified 239 peptides. Six peptides were identified by PEAKS de novo that were not identified at 1% FDR by the library search, whereas 25 peptides were identified by PEAKS DB that were not identified by PEAKS de novo. In total, 89.5% of peptides that were identified by the library search could also be identified by de novo sequencing. We also found that more than 99.5% of peptides that were identified by PEAKS de novo derived from the top five sequence candidates of the spectra (Figure S3B). The median of average local confidence (ALC) score for peptides identified by PEAKS de novo was 85 ± 13.5.

Peptide identification

Step 1: LC-MS/MS data were searched against the human proteome [UniProt v_05102017 with additional possible contaminations such as all UniProt entries for Epstein-Barr virus (EBV), the virus used to immortalize C1R cells and the bovine serum proteome] by PEAKS Studio 8.5 (Bioinformatics Solutions; Figure S2A). MS data files were imported into PEAKS Studio 8.5 (PEAKS de novo, PEAKS DB) and subjected to default data refinement. The parent mass error tolerance was set to 10 and 15 ppm and the fragment mass error tolerance to 0.02 or 0.1 Da for data generated by Thermo or SCIEX instruments, respectively (based on the software's default settings). Oxidation of methionine and deamidation of asparagine or glutamine were set in the de novo and database peptide searches as variable PTMs. A 1% FDR cutoff was applied, and all peptides identified by PEAKS DB were defined as linear peptides. For spectra that were just identified by PEAKS de novo (de novo–only peptides), the top five candidates (see "De novo sequencing algorithm evaluation" section) were extracted. Although we found that 85% of correct sequences appear as the first candidate (Figure S2B), we extracted multiple high-confidence candidates instead of just the highest hit per spectrum. Hence, by this logic, we reduced possible false-positive spliced peptide matching due to isobaric PTM and substitution errors. For finding the ALC cutoff for de novo candidates in each dataset, the ALC of spectra for identified linear peptides (at 1% FDR) was exported, and their corresponding median and SD were calculated. For retaining high-confidence de novo candidates for each spectrum, a mathematical model was generated by using Eq.1

ALC cutoff for de novo candidates = median ALC (linear peptides at 1% FDR) - SD (linear peptides at 1% FD (1)

For instance, for the HLA-B*57:01 dataset, the median and SD of ALC score for linear peptides were 91 and 12, respectively. Therefore, the ALC cutoff for this dataset was 79, and all candidates with ALC less than 79 were not included in the next steps. All de novo candidates that did not pass the ALC cutoff in their corresponding dataset were removed.

Step 2: An in-house algorithm ("Hybrid finder") was designed and set to execute the workflow indicated in Figure S3B. This software was designed and implemented using the JAVA programming language. De novo candidates that passed the ALC cutoff were analyzed by this hybrid finder algorithm. Given that it is not possible to distinguish between leucine (Leu) and isoleucine (Ile) residues by their mass [37], in the context of de novo sequencing, each Ile and Leu was converted to "L" in the proteome library [38]. This conversion prevents false assignment of linear peptides as spliced peptides by considering all permutations and combinations of Ile and Leu in the reference proteome. The steps of the algorithm were as follows: The algorithm first searches to find any match for the sequence in the proteome database. If this step fails to find a match, then the algorithm splits the peptide into all possible two segment pairs. The algorithm then searches each segment pair against the same library as used above. By analyzing the protein headers gathered from the search of the two segments, the program then lists proteins for potential cis-spliced and then trans-spliced peptides. There can be many possible source proteins for each segment (the algorithm identifies all possibilities), but because of limitations in space for the output, just one of the possibilities is listed in the attached tables (Table S4). If a sequence is not included in any of the above peptide categories (linear, cis-, or trans-spliced), then the algorithm assigns it as having no current biological explanation.

Step 3: De novo candidates from the same spectrum were grouped as candidates. In this step, we retrieved the ALC score for each candidate from the first de novo sequencing (in step 1) and the assigned linear or splicing type from the hybrid finder output (step 2). Then, we reranked the candidates in each spectrum group based on two different criteria. The first criterion was biological possibility, with our rationale being that a linear peptide explanation was deemed more likely than a cis-spliced peptide, which in turn was deemed more likely than a trans-spliced peptide. After this step, if multiple peptide sequences (of the same biological explanation) within a spectrum group were tied for first place, then such peptides were reranked on the basis of their de novo sequencing ALC score. For instance, if there were two cis-spliced candidates (and thus no linear candidates) tied for the first place in a group, then the one with the higher ALC would claim the first place.

Then, the first ranked candidate was kept from each candidate group, and all other candidates were removed. In the next step, only the candidates with a splicing (cis or trans) explanation were kept, and any linear and no biological explanation (NBE) candidates were removed. Subsequently, by using an in-house algorithm, all such spliced candidates were merged into a FASTA format, with sequences concatenated into representative pseudoprotein lengths (i.e., to mimic typical protein entries and thus to not bias protein scoring), and these collective sequences were appended to the original UniProt database. The resultant merged proteome database was used for the second and final PEAKS DB search, with the identical search parameters as the first PEAKS DB search. Peptides that were identified at 1% FDR that matched to our spliced candidate list were counted as spliced peptides (Table S4).

REFERENCES

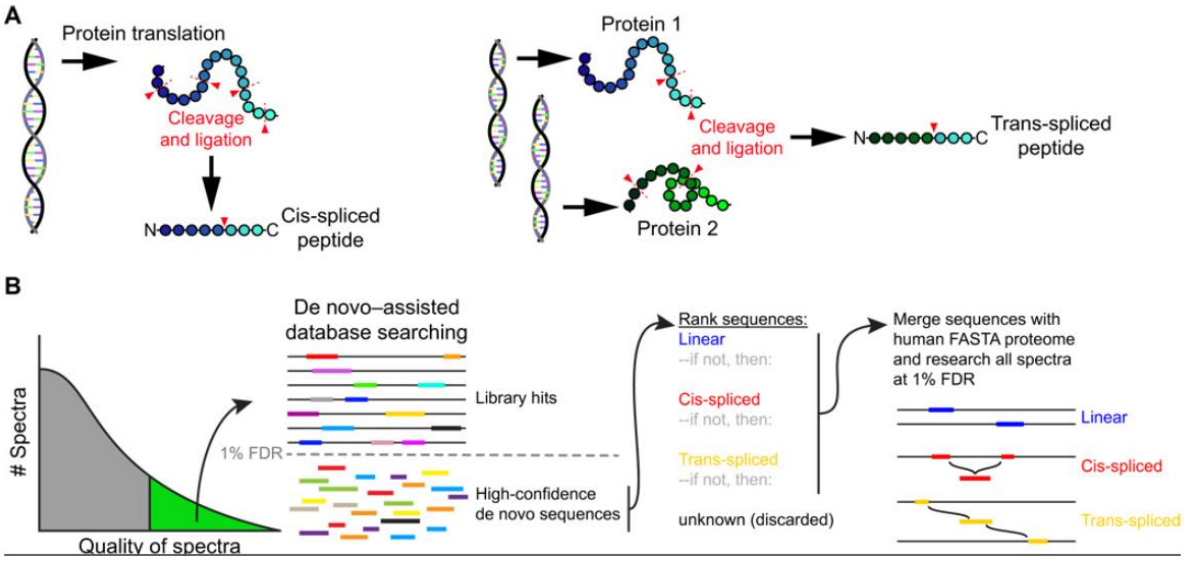
1. M. Groettrup, C. J. Kirk, M. Basler Proteasomes in immune cells: More than peptide producers? Nat. Rev. Immunol. 10, 73–78 (2010).
2. A. F. Kisselev, T. N. Akopian, K. M. Woo, A. L. Goldberg The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. J. Biol. Chem. **274**, 3363–3371 (1999).
3. J. Harbige, M. Eichmann, M. Peakman New insights into non-conventional epitopes as T cell targets: The missing link for breaking immune tolerance in autoimmune disease? J. Autoimmun. **84**, 12–20(2017).
4. M. Mishto, J. Liepe Post-translational peptide splicing and T cell responses. Trends Immunol. **38**, 904–915 (2017).
5. N. Vigneron, V. Stroobant, J. Chapiro, A. Ooms, G. Degiovanni, S. Morel, P. van der Bruggen, T. Boon, B.J. Van den Eynde An antigenic peptide produced by peptide splicing in the proteasome. Science **304**, 587–590 (2004).
6. E. H. Warren, N. J. Vigneron, M. A. Gavin, P. G. Coulie, V. Stroobant, A. Dalet, S. S. Tykodi, S. M. Xuereb, J. K. Mito, S. R. Riddell, B. J. Van den Eynde An antigen produced by splicing of noncontiguous peptides in the reverse order. Science **313**, 1444–1447 (2006).
7. K. Hanada, J. W. Yewdell, J. C. Yang Immune recognition of a human renal cancer antigen through post-translational protein splicing. Nature **427**, 252–256 (2004).
8. C. R. Berkers, A. de Jong, H. Ovaa, B. Rodenko Transpeptidation and reverse proteolysis and their consequences for immunity. Int. J. Biochem. Cell Biol. **41**, 66–71 (2009).
9. J. Liepe, F. Marino, J. Sidney, A. Jeko, D. E. Bunting, A. Sette, P. M. Kloetzel, M. P. H. Stumpf, A. J. R. Heck, M. Mishto A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science **354**, 354–358 (2016).
10. N. Vigneron, V. Ferrari, V. Stroobant, J. A. Habib, B. J. Van den Eynde Peptide splicing by the proteasome. J. Biol. Chem. **292**, 21170–21179 (2017).
11. P. Faridi, A. Wayne Purcell, N. P. Croft In Immunopeptidomics We Need a Sniper Instead of a Shotgun. Proteomics **18**, e1700464 (2018).
12. E. Caron, D. J. Kowalewski, C. Chiek Koh, T. Sturm, H. Schuster, R. Aebersold Analysis of Major Histocompatibility Complex (MHC) immunopeptidomes using mass spectrometry. Mol. Cell. Proteomics **14**, 3105–3117 (2015).
13. J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. A. Lajoie, B. Ma PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol. Cell. Proteomics **11**, M111.010587 (2012).
14. R. B. Schittenhelm, N. L. Dudek, N. P. Croft, S. H. Ramarathinam, A. W. Purcell A comprehensive analysis of constitutive naturally processed and presented HLA-C*04:01 (Cw4)-specific peptides. Tissue Antigens **83**, 174–179 (2014).
15. D. Maddelein, N. Colaert, I. Buchanan, N. Hulstaert, K. Gevaert, L. Martens The iceLogo web server and SOAP service for determining protein consensus sequences. Nucleic Acids Res. **43**, W543–W546 (2015).
16. N. Colaert, K. Helsens, L. Martens, J. Vandekerckhove, K. Gevaert Improved visualization of protein consensus sequences by iceLogo. Nat. Methods **6**, 786–787 (2009).
17. M. Fälth, M. Svensson, A. Nilsson, K. Sköld, D. Fenyö, P. E. Andren Validation of endogenous peptide identifications using a database of tandem mass spectra. J. Proteome Res. **7**, 3049–3053 (2008).
18. A. Devabhaktuni, J. E. Elias Application of de novo sequencing to large-scale complex proteomics data sets. J. Proteome Res. **15**, 732–742 (2016).
19. S. L. Lundstrom, B. Zhang, D. Rutishauser, D. Aarsland, R. A. Zubarev SpotLight Proteomics: Uncovering the hidden blood proteome improves diagnostic power of proteomics. Sci. Rep. **7**, 41929 (2017).
20. T. Muth, B. Y. Renard Evaluating de novo sequencing in proteomics: Already an accurate alternative to database-driven peptide identification? Brief. Bioinform. 10.1093/bib/bbx033 (2017).
21. Savidor, R. Barzilay, D. Elinger, Y. Yarden, M. Lindzen, A. Gabashvili, O. Adiv Tal, Y. Levin Database-independent Protein Sequencing (DiPS) enables full-length de novoprotein and antibody sequence determination. Mol. Cell. Proteomics 16, 1151–1161 (2017).

22. D. Chessman, L. Kostenko, T. Lethborg, A. W. Purcell, N. A. Williamson, Z. Chen, L. Kjer-Nielsen, N. A.Mifsud, B. D. Tait, R. Holdsworth, C. A. Almeida, D. Nolan, W. A. Macdonald, J. K. Archbold, A. D. Kellerher, D. Marriott, S. Mallal, M. Bharadwaj, J. Rossjohn, J. McCluskey Human leukocyte antigen class I-restricted activation of CD8+ T cells provides the immunogenetic basis of a systemic drug hypersensitivity. Immunity **28**, 822–832 (2008).
23. D. A. Ostrov, B. J. Grant, Y. A. Pompeu, J. Sidney, M. Harndahl, S. Southwood, C. Oseroff, S. Lu, J. Jakoncic, C. A. F. de Oliveira, L. Yang, H. Mei, L. Shi, J. Shabanowitz, A. M. English, A. Wriston, A. Lucas, E. Phillips, S. Mallal, H. M. Grey, A. Sette, D. F. Hunt, S. Buus, B. Peters , Drug hypersensitivity caused by alteration of the MHC-presented self-peptide repertoire. Proc. Natl. Acad. Sci. U.S.A. **109**, 9959–9964(2012).
24. P. T. Illing, J. P. Vivian, N. L. Dudek, L. Kostenko, Z. Chen, M. Bharadwaj, J. J. Miles, L. Kjer-Nielsen, S.Gras, N. A. Williamson, S. R. Burrows, A. W. Purcell, J. Rossjohn, J. McCluskey Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. Nature **486**, 554–558 (2012).
25. P. Pymm, P. T. Illing, S. H. Ramarathinam, G. M. O'Connor, V. A. Hughes, C. Hitchen, D. A. Price, B. K. Ho,D. W. McVicar, A. G. Brooks, A. W. Purcell, J. Rossjohn, J. P. Vivian MHC-I peptides get out of the groove and enable a novel mechanism of HIV-1 escape. Nat. Srtuct. Mol. Biol. **24**, 387–394 (2017).
26. C. M. Platteel, J. Liepe, W. van Eden, M. Mishto, A. Sijts An unexpected major role for proteasome-catalyzed peptide splicing in generation of T cell epitopes: Is there relevance for vaccine development? Front. Immunol. 8, 1441 (2017).
27. A. Dalet, N. Vigneron, V. Stroobant, K. Hanada, B. J. Van den Eynde 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. **184**, 3016–3024 (2010).
28. A. Michaux, P. Larrieu, V. Stroobant, J. F. Fonteneau, F. Jotereau, B. J. van den Eynde, A. Moreau-Aubry, N. Vigneron 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. **192**,1962–1971 (2014).
29. F. Ebstein, K. Textoris-Taube, C. Keller, R. Golnik, N. Vigneron, B. J. Van den Eynde, B. Schuler-Thurner, D.Schadendorf, F. K. M. Lorenz, W. Uckert, S. Urban, A. Lehmann, N. Albrecht-Koepke, K. Janek, P. Henklein, A. Niewienda, P. M. Kloetzel, M. Mishto Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. Sci. Rep. **6**, 24032 (2016).
30. M. F. Cusick, J. E. Libbey, R. S. Fujinami Molecular mimicry as a mechanism of autoimmune disease. Clin Rev Allergy Immunol **42**, 102–111 (2012).
31. K. Giam, R. Ayala-Perez, P. T. Illing, R. B. Schittenhelm, N. P. Croft, A. W. Purcell, N. L. Dudek A comprehensive analysis of peptides presented by HLA-A1. Tissue Antigens **85**, 492–496 (2015).
32. M. J. Rist, K. M. Hibbert, N. P. Croft, C. Smith, M. A. Neller, J. M. Burrows, J. J. Miles, A. W. Purcell, J. Rossjohn, S. Gras, S. R. Burrows T cell cross-reactivity between a highly immunogenic EBV epitope and a self-peptide naturally presented by HLA-B*18:01^{+} cells. J. Immunol. **194**, 4668–4675 (2015).
33. S. H. Ramarathinam, S. Gras, S. Alcantara, A. W. S. Yeung, N. A. Mifsud, S. Sonza, P. T. Illing, E. N. Glaros,R. J. Center, S. R. Thomas, S. J. Kent, N. Ternette, D. F. J. Purcell, J. Rossjohn, A. W. Purcell Identification of native and post-translationally modified HLA-B*57:01-restricted HIV envelope derived epitopes using immunoproteomics. Proteomics **18**, e1700253 (2018).
34. N. L. Dudek, N. P. Croft, R. B. Schittenhelm, S. H. Ramarathinam, A. W. Purcell A systems approach to understand antigen presentation and the immune response. Methods Mol. Biol. **1394**, 189–209 (2016).
35. J. G. Abelin, D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, J. Stevens, W. Lane, G. L. Zhang, T. M. Eisenhaure, K. R. Clauser, N. Hacohen, M. S. Rooney, S. A. Carr, C. J. Wu Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity 46, 315–326 (2017).
36. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. 17, 2337–2342 (2003).
37. B. Ma, R. Johnson, De novo sequencing and homology searching. Mol. Cell. Proteomics 11,O111.014902 (2012).

38. R. Flower On the utility of alternative amino acid scripts. Bioinformation 8, 539–542 (2012).
39. X. Han, L. He, L. Xin, B. Shan, B. Ma PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. J. Proteome Res. 10, 2930–2936 (2011).
40. A. M. Bolger, M. Lohse, B. Usadel rimmomatic: A flexible trimmer for Illumina sequence data.Bioinformatics 30, 2114–2120 (2014).
41. D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G.Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K.To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H.Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, P. Flicek Ensembl 2018. Nucleic Acids Res. 46, D754–D761 (2018).
42. M. Andreatta, M. Nielsen Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. Bioinformatics 32, 511–517 (2016).
43. E. Karosiene, C. Lundegaard, O. Lund, M. Nielsen NetMHCcons: A consensus method for the major histocompatibility complex class I predictions. Immunogenetics 64, 177–186 (2012).
44. L. Kostenko, L. Kjer-Nielsen, I. Nicholson, F. Hudson, A. Lucas, B. Foley, K. Chen, K. Lynch, J. Nguyen, A. H. B. Wu, B. D. Tait, R. Holdsworth, S. Mallal, J. Rossjohn, M. Bharadwaj, J. McCluskey Rapid screening for the detection of HLA-B57 and HLA-B58 in prevention of drug hypersensitivity. Tissue Antigens 78, 11–20(2011).
45. L. Kjer-Nielsen, C. S. Clements, A. G. Brooks, A. W. Purcell, M. R. Fontes, J. McCluskey, J. Rossjohn The structure of HLA-B8 complexed to an immunodominant viral determinant: Peptide-induced conformational changes and a mode of MHC class I dimerization. J. Immunol. 169, 5153–5160 (2002).
46. W. Kabsch XDS. Acta Cryst. 66, 125–132 (2010).
47. Collaborative Computational Project, Number 4 The CCP4 suite: Programs for protein crystallography.Acta Cryst. 50, 760–763 (1994).
48. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read Phaser crystallographic software. J. Appl. Cryst. 40, 658–674 (2007).
49. P. D. Adams, R. W. Grosse-Kunstleve, L. W. Hung, T. R. Ioerger,A. J. McCoy, N. W. Moriarty, R. J. Read, J. C.Sacchettini, N. K. Sauter,T. C. Terwilliger PHENIX: Building new software for automated crystallographic structure determination. Acta Cryst. 58, 1948–1954 (2002).
50. P. Emsley, K. Cowtan Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132 (2004).
51. I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, D. C. Richardson MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 35, W375–W383 (2007).

Figure 1

The nature of cis- and trans-spliced peptides and their identification from HLA immunopeptidomes sequenced by mass spectrometry.

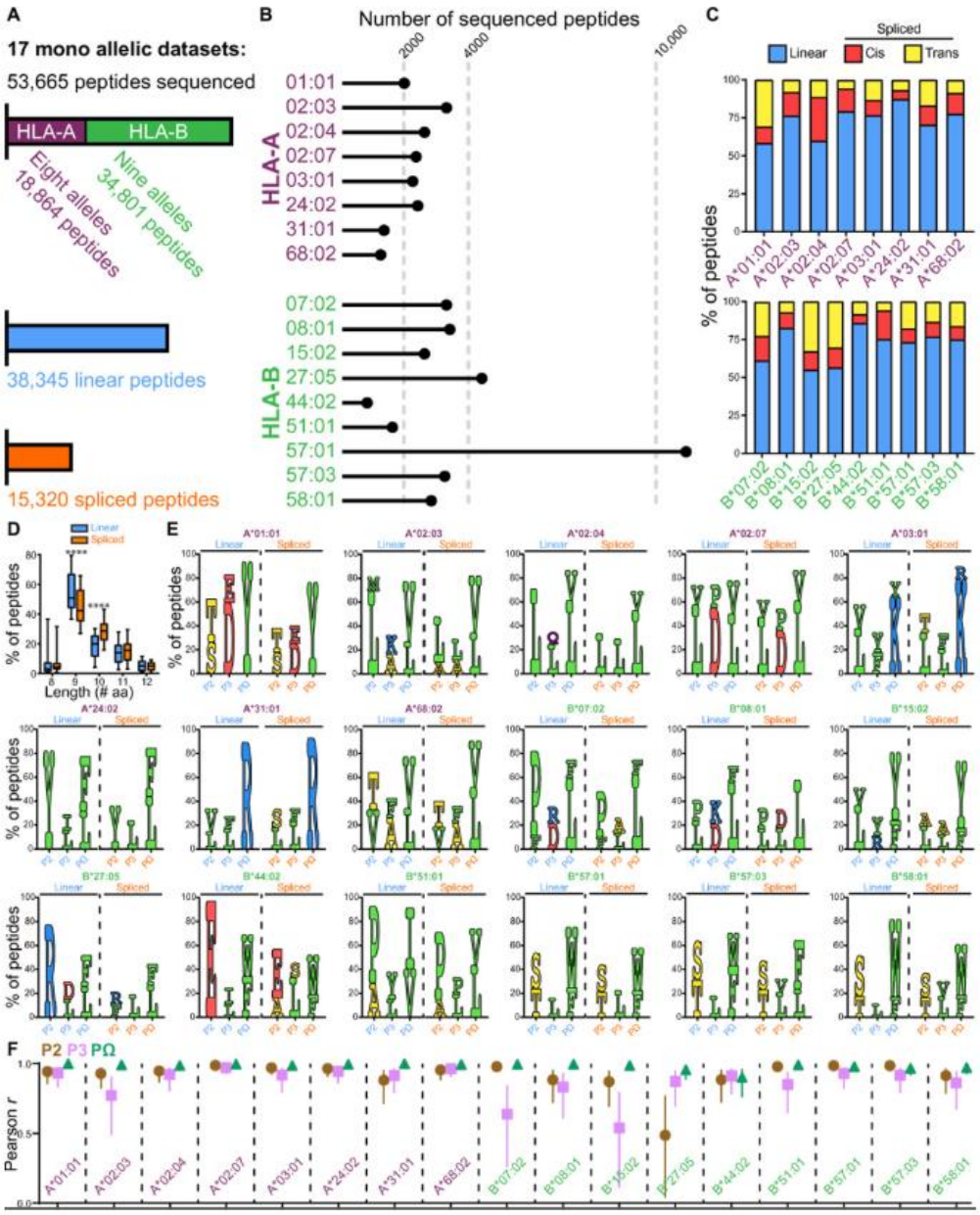


(A) Cartoon representation of (left) cis-spliced and (right) trans-spliced peptide generation. Cis-spliced peptides are formed after cleavage and ligation of segments from the same source protein; trans-spliced peptides are formed after ligation of cleaved segments from two different source proteins. Such peptides may then be subjected to HLA antigen presentation pathways for display at the cell surface.

(B) Workflow for the identification of linear and spliced peptides. From high-quality MS/MS spectra, an initial de novo–assisted database search (using the reference human proteome) was carried out, filtering the data at a 1% FDR. Subsequently, all high-quality de novo–only sequenced peptides (the top five sequences per spectrum) that fell below this threshold were searched using our in-house algorithm to hierarchically rank peptides as to whether they had a linear > cis > trans explanation (or, failing this, were discarded). The top-ranked peptides for each spectrum sequence were then built into a custom FASTA-formatted database and merged with the human proteome, and the original MS/MS data were researched, taking the 1% FDR cutoff as a final output of results.

Figure 2

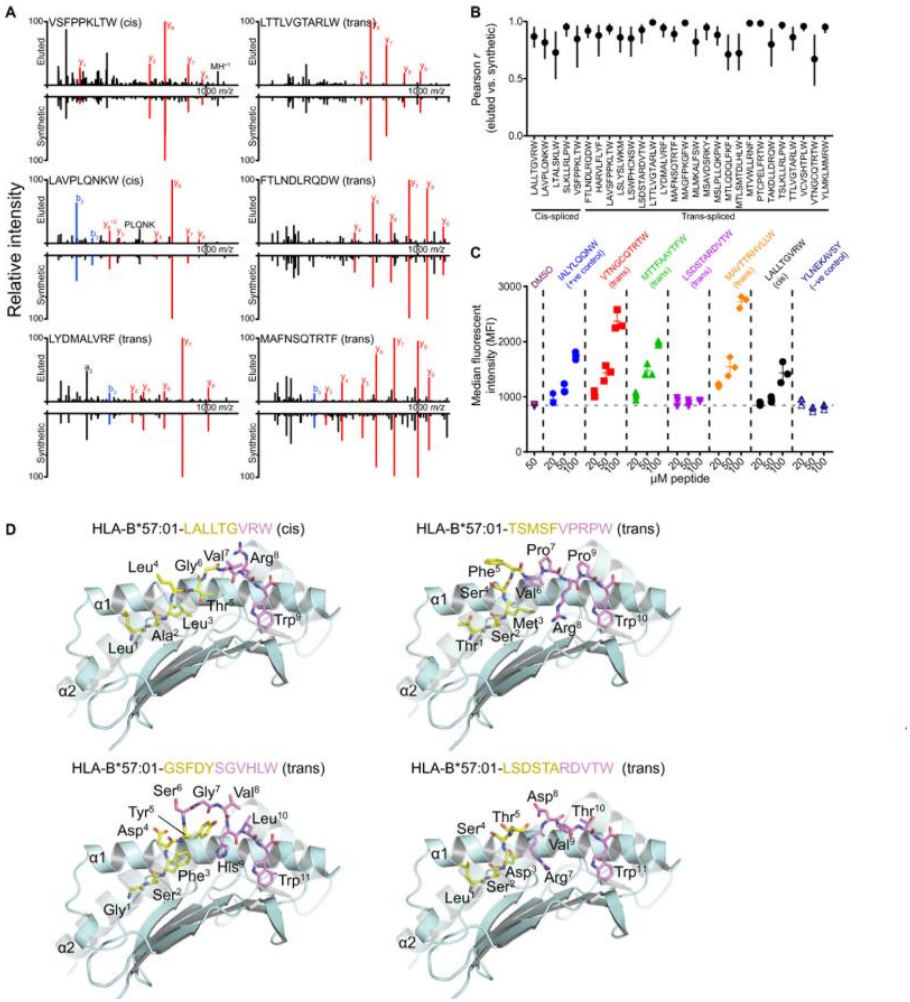
Identification, length distribution, and motif analysis of linear and spliced peptides by a combined de novo library searching hybrid workflow approach.



(A and B) More than 50,000 peptides eluted from eight HLA-A–expressing and nine HLA-B–expressing monoallelic cell lines were sequenced and defined as either linear or spliced in origin.
(C) Proportion of linear, cis-, or trans-spliced peptides contributing to each HLA allelic dataset.
(D) Length [number of amino acid (aa)] distribution of all identified linear and spliced peptides (****$P < 0.0001$, two-way multiple-comparison ANOVA test).
(E) Motif analysis for 9-mer and 10-mer linear and spliced peptides, showing the percentage of enriched amino acids (if greater than 10%) at each of positions P2, P3, and PΩ. (Note that in spliced peptides, L stands for both leucine and isoleucine.)
(F) Pearson *r* value correlation between the amino acids enriched in linear and spliced peptides for each allele at each of positions P2, P3, and PΩ (all data were $P < 0.05$).

Figure 3

Validation of identified cis- and trans-spliced peptides from C1R-HLA-B*57:01 cells.



(A) LC-MS/MS spectra from two cis- and four trans-spliced B*57:01 peptides, comparing eluted (upper panel of each peptide) and synthetic (lower panel for each peptide). Characteristic y and b ions are highlighted for each peptide. Data are displayed as relative intensity. m/z, mass/charge ratio.

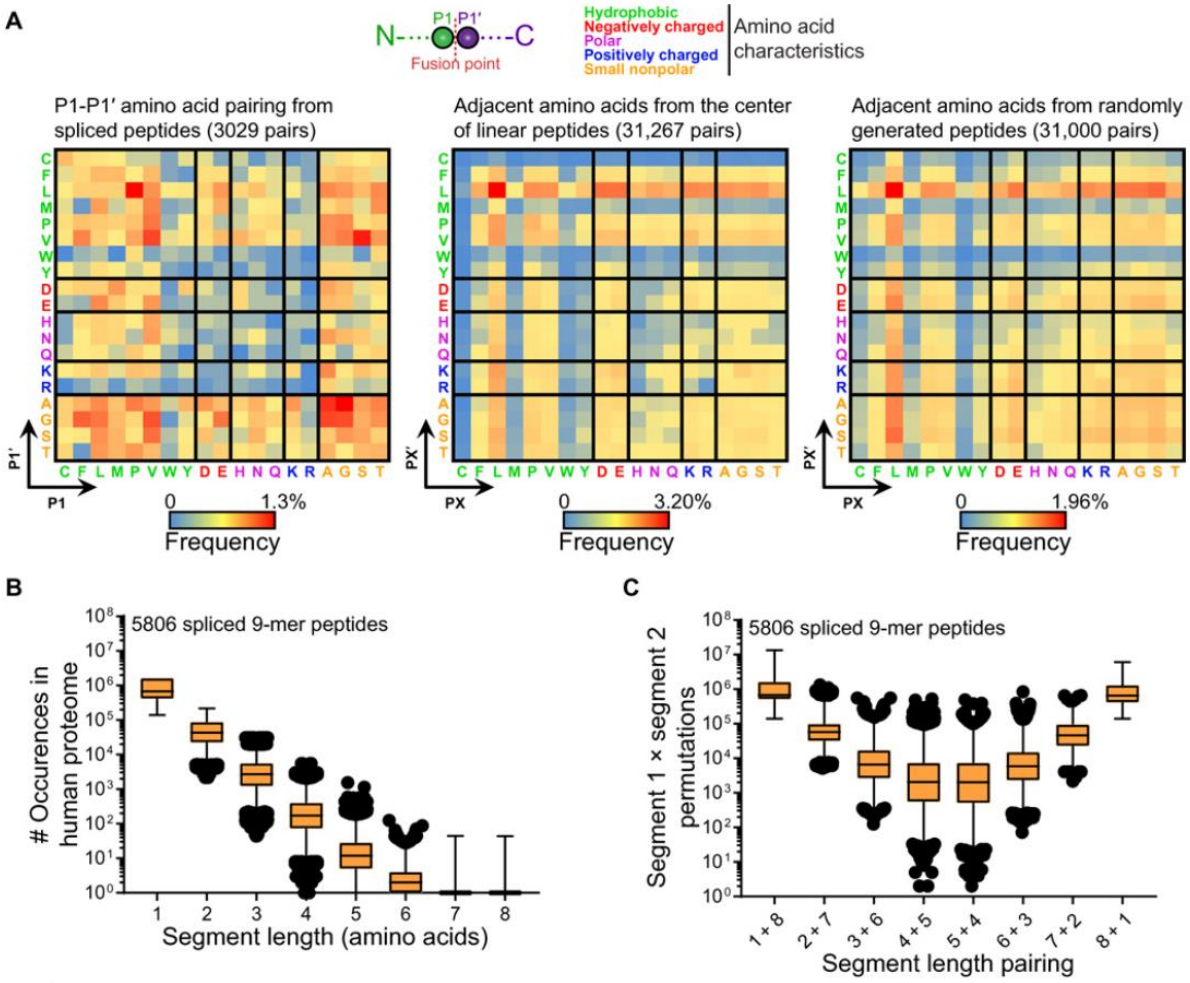
(B) Pearson r value correlation score of spectra from 28 eluted B*57:01 spliced peptides compared with their synthetic versions.

(C) Stabilization of cellular HLA-B*57:01 molecules (T2-B*57:01 cell line) by a panel of synthetic spliced peptides and the EBV-derived B*57:01-restricted positive control peptide IALYLQQNW. No peptide [dimethyl sulfoxide (DMSO)] and the non-B*57:01–restricted peptide YLNEKAVSY were used as negative controls. Peptides were tested at the indicated concentrations, and surface HLA was detected by flow cytometry using the B*57:01-specific antibody 3E12. Data show median fluorescent intensity (MFI) from three independent replicates, and error bars show mean and SD.

(D) Conformations of four spliced peptides (cis-LALLTG + VRM, trans-TSMSF + VPRPW, trans-GSFDY + SGVHLW, and trans-LSDSTA + RDVTW) presented by HLA-B*57:01. Cartoon representation of the peptide-binding groove of HLA-B*57:01 (cyan). The β2 helix has been rendered transparent to improve visibility of the peptide. Spliced peptides are represented as colored sticks, with the N-terminal segment in yellow and the C-terminal segment in purple, with individual residues as indicated. Note that, for peptide GSFDYSGVHLW, there were two possible explanations for the segments, GSFDY + SGVHLW or GSFDYS + GVHLW, and the former is indicated.

Figure 4

Spliced junction amino acid bias and analysis of donor segment frequency.



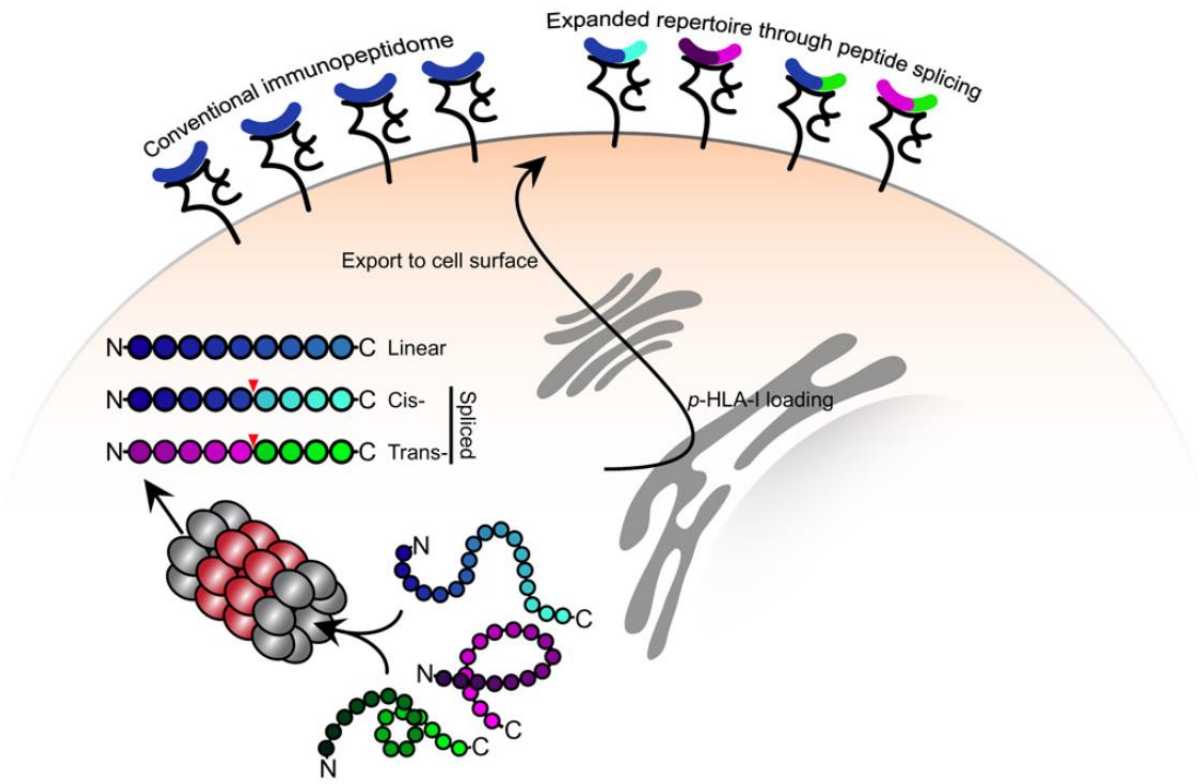
(A) A subset (3029) of spliced peptides (from all 17 analyzed HLA-A and HLA-B alleles) with only one possible splicing explanation were assessed for amino acid bias at the P1 and P1' positions (left). The central amino acid pairs from 31,267 identified linear peptides (middle) and adjacent amino acids from the center of 31,000 randomly generated (conforming to the amino acid frequency distribution of the human proteome) peptide sequences (right) were used for comparison. Heat map frequency colors are as indicated per dataset, and amino acids are colored according to broad physiochemical characteristics. All Ile residues were substituted for Leu.

(B) Number of occurrences for each possible segment of a dataset of 5806 spliced nonamers, calculated from the UniProt reference human proteome.

(C) Permutations, calculated from multiplying together the numbers of occurrences for each given segment, for generating each of the same set of 5806 spliced nonamers. For (B) and (C), data show box plots with whiskers set to the 1 to 99 percentile.

Figure 5

Cartoon model for the increased p-HLA display engendered by peptide splicing.



Although conventional, linear peptides allow sampling of (for any given HLA allele) limited regions of the proteome, we propose that the combined actions of cis- and trans-splicing enable a greater proportion of the cellular proteome to be displayed for scrutiny by T cells.