A large fraction of HLA class I ligands are proteasome-generated spliced peptides

Juliane Liepe [1,#], Fabio Marino [2,3], John Sidney [4], Anita Jeko [2,3], Daniel E. Bunting [1], Alessandro Sette [4], Peter M. Kloetzel [5,6], Michael P.H. Stumpf [1], Albert J.R. Heck [2,3], Michele Mishto [5,6,#].

[1] Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, SW7 2AZ, UK

[2] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, 3584 CH Utrecht, The Netherlands

[3] Netherlands Proteomics Centre, CH Utrecht, The Netherlands

[4] Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, United States.

[5] Institut für Biochemie, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany

[6] Berlin Institute of Health, 10117 Berlin, Germany

Keywords: proteasome, antigen presentation, EThcD, HLA class I-restricted epitopes, proteasome-catalyzed peptide splicing.

Abbreviations: electron-transfer higher-energy collision dissociation (EThcD); higher-energy collision dissociation (HCD); false discovery rate (FDR); HLA class I (HLA-I); lymphoblastoid cell line (LCL); mass spectrometry (MS); molecular weight (MW); proteasome-catalyzed peptide splicing (PCPS).

The proteasome generates the epitopes presented on HLA class I molecules that elicit CD8$^+$ T cell responses. While reports of proteasome-generated spliced epitopes exist, they have been regarded as rare events. Here, however, we show that the proteasome-generated spliced peptide pool accounts for one third of the entire HLA class I immunopeptidome in terms of diversity and one fourth in terms of abundance. They also represent a unique set of antigens, possessing particular and distinguishing features. We validated this observation using a range of complementary experimental and bioinformatics approaches, and multiple cell types. The widespread appearance and abundance of proteasome-catalyzed peptide splicing events will inform immunobiology and autoimmunity theories, and may provide a previously untapped source of epitopes for uses in vaccines and cancer immunotherapy.

The presentation of epitopes on the cell surface is a key mechanism by which organisms identify the presence of pathogens, metabolic malfunctioning, or tumors. The HLA class I (HLA-I) immunopeptidome, *i.e.* the epitopes allocated onto the HLA-I molecules, impinges upon the CD8$^+$ T cell repertoire and the cell-mediated immune response (*1*). HLA-I immunopeptidomes are usually investigated by sequence identification of peptides eluted from HLA-I molecules using LC-tandem mass spectrometry (MS) (**Fig. S1**). The key step for the transformation of a protein into HLA-I-restricted epitopes is usually processing by the proteasome (*1*), which cuts proteins into peptides; alternatively, the proteasome can also cut and paste peptide sequences, thereby releasing peptide antigens that do not correspond to the original protein sequence (*2*) (**Fig. S2**). This proteasome-catalyzed peptide splicing (PCPS) has long been considered to occur only very rarely; partly this has been because the screening of the HLA-I immunopeptidome for proteasome-generated spliced peptides was impeded by methodological challenges.

To overcome these problems we developed an analytical strategy, which accounts for recent discoveries underpinning the PCPS mechanism, and which can handle the vast proteome-wide human spliced peptide database (**Fig. S3**). With this strategy we initially analyzed the HLA-I-eluted immunopeptidome of the GR lymphoblastoid cell line (GR-LCL) adopting, for a deeper coverage of the immunopeptidome, a 2D peptide pre-fractionation strategy, followed by a hybrid peptide fragmentation method, electron-transfer higher-energy collision dissociation (EThcD) for peptide identification (*3, 4*) (**Fig. S1**), supplemented by an adapted target-decoy approach (**Fig. S4**). Our analysis led to the identification of 6592 non-spliced and 3417 spliced 9-12mer peptides (**Table S1**). The latter thus represent 34 % of the total of identified antigenic peptides (**Fig. 1A**) increasing the number of identified HLA-I ligands by some 50%. The authenticity of the identified spliced epitopes is confirmed by comparing the LC-MS/MS spectra of 98 exemplary spliced peptides with their corresponding synthetic peptides and the computation of their correlation score (**Table S2** and **Fig. S5**). In addition, the proteasome-dependent generation of the spliced epitopes was verified *in vitro* for three examples by digestions of synthetic polypeptides harboring the corresponding antigenic peptides by purified 20S proteasome (**Fig. S6**).

We furthermore queried the HLA-I immunopeptidome mass spectrometry data of GR-LCL against the standard Swissprot human proteome database, which does not account for spliced peptides; this reveals that 655 peptides (*i.e.* 9 % of the total non-spliced antigenic peptides) would be erroneously matched against this incomplete database as non-spliced peptides, since they have much better hits as spliced peptides in our search against the combined spliced and non-spliced peptide database (**Fig. 1A**). The correct sequences for a set of these antigenic spliced peptides are verified by comparing the LC-MS/MS spectra of the synthetic spliced and non-spliced candidates with the corresponding LC-MS/MS of the GR-LCL HLA-I immunopeptidome (**Fig. S7A-I**). Our identifications are further supported by the ion score distributions (see **Fig. S4**) of the non-spliced peptides and of those spliced peptides that were wrongly assigned as non-spliced peptides using the standard Swissprot human proteome database, which differ only slightly in their median, but not in the overall shape (**Fig. S7J**).

In independent technical replicates of the GR-LCL HLA-I immunopeptidome analyzed without pre-fractioning (see **Fig. S1, S8A**) through EThcD or higher-energy collision dissociation (HCD) we identified thousands of peptides, where the spliced peptide pools represent between 25 % and 33 % of the HLA-I immunopeptidome diversity (**Fig. 1B**, **Table S1A**). To corroborate whether this unexpected finding would not be a peculiarity of the GR-LCL cell line we investigated the HLA-I immunopeptidome of unrelated cell lines. Here, similar results were obtained in the analysis by EThcD of a non-related C1R lymphoid cell line (*5*), where one third of the HLA-I immunopeptidome variety is represented by spliced peptides (**Fig. 1A, Table S1B**). This large prevalence of spliced peptides in the HLA-I immunopeptidome is not a peculiar characteristic of lymphoblastoid cell lines since in the HLA-I immunopeptidome of primary human fibroblasts (*6*) 29 % of the identified antigenic peptides were spliced peptides (**Fig. 1A, Table S1C**). Again, if we were to query those datasets only against the standard Swissprot human proteome database, we would wrongly assign 4 - 7 % of the antigenic peptides as non-spliced peptides (**Fig. 1A, B**), while missing all spliced peptides.

Spliced peptides were prevalent not only in the HLA-I immunopeptidome but also in the unsorted pool of GR-LCL cell lysate peptides with molecular weight (MW) smaller than 3 kDa, the maximum size of peptides produced *in vitro* by proteasome (*7*) (**Fig. 1C**). The numbers of both non-spliced and spliced 9-12mer peptides declined when we used an inhibitor of proteasome activity such as epoxomicin (**Fig. 1C, S9**). Inhibition of proteasome activity led to a longer median length of the non-spliced peptides (**Fig. S9**), which was to be expected since proteasome generates peptides with an average length of 11 residues (*7, 8*). It also almost eliminated all spliced 9-12mer peptides (**Fig. 1C, S9**), thereby confirming that the identified spliced peptides are generated by proteasome.

We further verified the proteasome-dependency of the spliced peptides by querying the HLA-I immunopeptidome of the T2 cell line, which lacks a functioning transporter associated with antigen processing (TAP). As for other TAP-deficient cell lines the few antigenic peptides identified so far in its HLA-I immunopeptidome derive from both proteasome- and SPase-mediated proteolysis in similar manner (*9, 10*). As expected, compared to other HLA-I immunopeptidomes we identified a dramatically reduced number of both spliced and non-spliced peptides eluted from the HLA-I molecules of T2 cells. Among them, spliced peptides represented only 13 % of the whole T2 HLA-I immunopeptidome (**Fig. 1B**). By contrast, the analysis of a cytosolic unsorted pool of 9-12mer peptides of the T2 cells peptides (with MW < 3 kDa) showed a normal frequency of spliced peptides (25 %, **Fig. 1C**). Together these experiments provide further evidence that spliced peptides presented by HLA-I molecules are produced by proteasome, and that about half of the TAP-independent antigenic peptide pool is generated through proteasome activity (*9*).

As a final (negative) control we performed LC-MS/MS analysis of a tryptic lysate of the C1R cell line (**Fig. S8**). This time the cell lysate was first filtered to include only proteins with MW > 30 kDa to exclude intracellular peptides generated by the proteasome in the mixture. In this tryptic digest we identified about 1300 9-12mer non-spliced peptides, but only a few 9-12mer spliced sequences (**Fig. 1C**). In fact, only 2.7 % of them are (likely incorrectly) annotated as spliced peptides, which provides us with an experimental limit on the overall false discovery rate (FDR) (see **Fig. S4B**).

We thus have independent experiments that provide evidence that about one third of peptides bound to HLA-I are generated by PCPS.

Although spliced peptides represent one third of the HLA-I immunopeptidome variety, their relevance from an immunological point of view could be undermined if they were not abundantly presented at the cell surface. We thus set out to quantify the abundance of each HLA peptide by label-free quantification, based on the intensity of the MS ion peak area. Although this method is not applicable for single peptides (*6, 8*), it has recently become well accepted when analyzing large proteomics datasets (*6, 11, 12*). We substantiate this strategy by titrating two pools of synthetic non-spliced and spliced antigenic peptides and observing no significant differences between non-spliced and spliced peptides (**Fig. 2A**). Using label-free quantification there appear to be significant differences in the abundance distribution between non-spliced and spliced peptides. Indeed, the spliced antigenic peptides are on average 16 – 26 % less abundant than non-spliced peptides in the HLA-I immunopeptidomes of three independent cell lines (**Fig. 2B**, **Fig. 2C**). Therefore, based on our results we conclude that spliced peptides not only represent one third of the HLA-I immunopeptidome variety but also approximately one fourth of the total amount of peptide molecules presented onto the HLA-I complex. In agreement with this finding we recently reported that the abundance of peptides of a small group of melanoma-associated spliced epitopes exposed onto the HLA-I complexes is comparable to that of non-spliced melanoma-associated epitopes (*13*). Moreover, we observed a specific response against those spliced epitopes in the peripheral blood of half of the melanoma patients we studied (*13*), highlighting the potential biological impact of the much larger set of spliced epitopes presented in this study.

Spliced epitopes could also represent a distinct pool of antigenic peptides with particular characteristics. The few pioneering studies on PCPS had already provided hints about generative mechanisms. For instance, they suggested that PCPS prefers specific peptide sequences, although the limited number of spliced peptides/epitopes identified has so far precluded an analysis with sufficient statistical power (*2, 8, 14-21*). With our large pool of spliced antigenic peptides this became possible. Contrasting the characteristics of the spliced and non-spliced peptides in the GR-LCL HLA-I immunopeptidome we observed no significant differences in terms of: (i) peptide length distribution (**Fig. S10A**) in agreement with our previous observation on *in vitro* proteasome-catalyzed reactions (*8*); (ii) frequency of the number of putative parental proteins that can generate each spliced or non-spliced peptide (**Fig. S10B**); and (iii) the frequencies of spliced and non-spliced peptides derived from a given antigen (**Fig. S10C**). Remarkably, one third of self-antigens are represented on the GR-LCL cell surface only by spliced peptides (**Fig. 3A** and **Fig. S10D**), showing that PCPS increases the antigen exposure and expands the HLA-I immunopeptidome-mediated surveillance of the immune system.

Among the identified spliced antigenic peptides in the GR-LCL 2D HLA-I immunopeptidome we observed a similar number of spliced peptides generated by the ligation of two splice-reactants following their orientation in the parental protein or by inverting their order, *i.e.* reverse PCPS (50.1% were normal *cis* spliced peptides; see **Fig. S2**). We did not observe a clear preference for a specific length of intervening sequences, *i.e.* the sequences excised between two splice-reactants (**Fig. S2, 3B**). Also the length of the N- and C-terminal splice-reactants (see **Fig. S2**) was almost equally distributed with the exception of a seemingly preferred length of two residues in the N-terminal splice-reactant, which is most apparent for 9mer, 10mer and 11mer peptides (**Fig. 3C** and **Fig. S10E**). Since the second residue of the antigenic peptide often corresponds to the anchor site of the specific HLA-I molecules, we can speculate that preference for specific amino acids for the ligation could have introduced an evolutionary pressure on the HLA allotype selection, as has previously been hypothesized for the proteasome-dependent peptide hydrolysis and the C-terminus of the HLA-I-restricted peptides (*22*).

In the GR-LCL dataset we also compare spliced and non-spliced antigenic peptide motifs in relation to HLA-I haplotypes by predicting *in silico* their binding to HLA-I molecules applying the (NetMHC) ANN and (IEDB) SMM algorithms (*23, 24*). The two algorithms perform similarly if we consider the non-spliced antigenic peptides (**Fig. S11A-D**). They do, however, differ significantly in the prediction of how efficiently the spliced peptides bind the specific HLA-A and –B variants (**Fig. S11A-D**). Often spliced peptides predicted by (NetMHC) ANN to be barely compatible with the specific HLA-I cleft bind it with an experimentally determined IC_{50} below 5000 nM (**Table S2, Fig. S11E**). Such a phenomenon might be due to intrinsic differences in the motifs of spliced and non-spliced antigenic peptides. Indeed, since the algorithms have been trained exclusively on non-spliced epitopes (or non-randomized peptide libraries) their predictive power for that peptide type could be limited. This hypothesis is also supported by our previous *in vitro* PCPS analysis (*8*), where we observed that several spliced peptides were produced by proteasomal cutting at rarely used substrate cleavage sites. Such differences emerge when considering the sequence motifs of the spliced and non-spliced 9mers within the GR-LCL HLA-I immunopeptidome (**Fig. 3D**) and the other immunopeptidome datasets (**Fig. S12**), especially in the positions 2 and 9, which correspond to frequently used anchor sites.

To gather information about sequence preference of PCPS we exploited the large number of spliced antigenic peptides available here, and we computed the amino acid distribution at the P_N, P1, P1′ and P_C positions (see **Fig. S2**) of the GR-LCL HLA-I immunopeptidome. This outcome matched the data obtained for the replicate GR-LCL immunopeptidome (1D EThcD), but differs from the immunopeptidomes obtained from the C1R cell lineage and human primary fibroblasts (**Fig. 3E** and **Fig. S13**). The difference between the patterns of the spliced peptide P_N, P1, P1′ and P_C positions could be due to the HLA anchor site differences between the three cell lines. Our more general analysis differs markedly from the HLA-A*02:01-restricted P1–P1′ position pattern recently published by Berkers and colleagues (*20*), confirming that an HLA-unbiased strategy to identify spliced peptide patterns may be essential for developing PCPS prediction algorithms.

In summary, our study shows that the spliced peptides bound to the HLA-I molecules are very frequent and comparable in their amount to the non-spliced peptides but represent a distinct pool of antigens and possess particular immunological characteristics. One of the key features that may have maintained PCPS through evolutionary history (*8, 25, 26*) might be its higher degree of freedom of selecting antigenic peptide sequences. Targeting antigens through non-spliced peptides can be limited by the sequence restrictions that antigens have due to their function. PCPS is a solution to this problem as hinted at by the fact that a significant portion of the antigens are represented by spliced peptides only. Of course, the unexpectedly large frequency and amount of HLA-I-restricted spliced peptides may – and we strongly expect will – have profound implications for the concept of self/non-self peptide presentation: the large variety of potential spliced antigenic peptides would dramatically increase the number of antigenic peptides with overlapping sequences derived from either human or pathogen proteomes with direct implications for autoimmunity (*25, 27, 28*). On the other hand, the frequency of antigenic spliced peptides and their features could also in turn have positive implications for therapies involving HLA-I-restricted epitopes such as anti-viral vaccinations or mutation-specific adoptive T cell therapy against tumors (*29-33*): mutant antigens lacking HLA-I-restricted non-spliced epitopes could finally be targeted through spliced epitopes.

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. N. Vigneron *et al.*, An antigenic peptide produced by peptide splicing in the proteasome. *Science* **304**, 587-590 (2004).
3. C. K. Frese *et al.*, Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *J Proteome Res* **12**, 1520-1525 (2013).
4. G. P. Mommen *et al.*, Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc Natl Acad Sci U S A* **111**, 4507-4512 (2014).
5. E. Caron *et al.*, An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife* **4**, (2015).
6. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, M. Mann, Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics* **14**, 658-673 (2015).
7. 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).
8. M. Mishto *et al.*, Driving Forces of Proteasome-catalyzed Peptide Splicing in Yeast and Humans. *Mol Cell Proteomics* **11**, 1008-1023 (2012).
9. A. O. Weinzierl *et al.*, Features of TAP-independent MHC class I ligands revealed by quantitative mass spectrometry. *Eur J Immunol* **38**, 1503-1510 (2008).
10. E. Lorente *et al.*, Diversity of natural self-derived ligands presented by different HLA class I molecules in transporter antigen processing-deficient cells. *PLoS One* **8**, e59118 (2013).
11. J. Cox *et al.*, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526 (2014).
12. E. L. de Graaf, P. Giansanti, A. F. Altelaar, A. J. Heck, Single-step enrichment by Ti4+-IMAC and label-free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. *Mol Cell Proteomics* **13**, 2426-2434 (2014).
13. F. Ebstein *et al.*, Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. *Sci Rep* **6**, 24032 (2016).
14. K. Textoris-Taube *et al.*, The T210M Substitution in the HLA-a*02:01 gp100 Epitope Strongly Affects Overall Proteasomal Cleavage Site Usage and Antigen Processing. *J Biol Chem* **290**, 30417-30428 (2015).
15. E. H. Warren *et al.*, An antigen produced by splicing of noncontiguous peptides in the reverse order. *Science* **313**, 1444-1447 (2006).
16. A. Michaux *et al.*, A Spliced Antigenic Peptide Comprising a Single Spliced Amino Acid Is Produced in the Proteasome by Reverse Splicing of a Longer Peptide Fragment followed by Trimming. *J Immunol*, (2014).
17. J. Liepe *et al.*, The 20S Proteasome Splicing Activity Discovered by SpliceMet. *PLOS Computational Biology* **6**, e1000830 (2010).
18. 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).

19. A. Dalet *et al.*, An antigenic peptide produced by reverse splicing and double asparagine deamidation. *Proc Natl Acad Sci U S A*, (2011).

20. C. R. Berkers *et al.*, Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. *J Immunol* **195**, 4085-4095 (2015).

21. C. R. Berkers *et al.*, Peptide Splicing in the Proteasome Creates a Novel Type of Antigen with an Isopeptide Linkage. *J Immunol* **195**, 4075-4084 (2015).

22. G. Niedermann *et al.*, The specificity of proteasomes: impact on MHC class I processing and presentation of antigens. *Immunol Rev* **172**, 29-48 (1999).

23. B. Peters, A. Sette, Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* **6**, 132 (2005).

24. C. Lundegaard, O. Lund, M. Nielsen, Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* **24**, 1397-1398 (2008).

25. A. Grignolio *et al.*, Towards a liquid self: how time, geography, and life experiences reshape the biological identity. *Front Immunol* **5**, 153 (2014).

26. F. A. Steven, Immunology and evolution of infectious disease. *Princeton University Press*, (2002).

27. E. Bellavista *et al.*, Current understanding on the role of standard and immunoproteasomes in inflammatory/immunological pathways of multiple sclerosis. *Autoimmune Dis* **2014**, 739705 (2014).

28. H. G. Rammensee, Immunology: protein surgery. *Nature* **427**, 203-204 (2004).

29. S. A. Rosenberg, N. P. Restifo, Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62-68 (2015).

30. S. H. van der Burg, R. Arens, C. J. Melief, Immunotherapy for persistent viral infections and associated disease. *Trends Immunol* **32**, 97-103 (2011).

31. T. Blankenstein, M. Leisegang, W. Uckert, H. Schreiber, Targeting cancer-specific mutations by T cell receptor gene therapy. *Curr Opin Immunol* **33**, 112-119 (2015).

32. M. M. Gubin *et al.*, Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515**, 577-581 (2014).

33. P. G. Coulie, B. J. Van den Eynde, P. van der Bruggen, T. Boon, Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat Rev Cancer* **14**, 135-146 (2014).

34. M. Mishto *et al.*, Modeling the in vitro 20S proteasome activity: the effect of PA28-alphabeta and of the sequence and length of polypeptides on the degradation kinetics. *J Mol Biol* **377**, 1607-1617 (2008).

35. M. Mishto *et al.*, The immunoproteasome beta5i subunit is a key contributor to ictogenesis in a rat model of chronic epilepsy. *Brain Behav Immun*, (2015).

36. M. Mishto *et al.*, Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation. *Eur J Immunol*, (2014).

37. K. Textoris-Taube *et al.*, The T210M substitution in the HLA-A*02:01 gp100 epitope strongly affects overall proteasomal cleavage site usage and antigen processing. *J Biol Chem*, (2015).

38. J. Sidney *et al.*, Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr Protoc Immunol* **Chapter 18**, Unit 18 13 (2013).

39. K. Gulukota, J. Sidney, A. Sette, C. DeLisi, Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol* **267**, 1258-1267 (1997).

40. C. Lundegaard *et al.*, NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res* **36**, W509-512 (2008).

41. M. Falth *et al.*, Validation of endogenous peptide identifications using a database of tandem mass spectra. *J Proteome Res* **7**, 3049-3053 (2008).
42. R-Core-Team, in *R Foundation for Statistical Computing, Vienna, Austria*. (2014).
43. J. E. Elias, S. P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-214 (2007).
44. S. Tenzer *et al.*, Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance. *Nat Immunol* **10**, 636-646 (2009).

Supplementary Materials are available in the online version of the paper and are the following:

Figure S1. Mass spectrometry workflow – data acquisition.

Figure S2. Proteasome-catalyzed peptide splicing (PCPS).

Figure S3. Workflow for the identification of non-spliced and spliced peptides among the HLA class I-eluted peptides.

Figure S4. Mass spectrometry data analysis.

Figure S5. Comparison of EThcD MS/MS spectra of 98 spliced peptides eluted from the GR-LCL HLA-I molecules to their corresponding synthetic peptides.

Figure S6. EThcD LC-MS/MS spectra of three representative spliced peptides produced by PCPS *in vitro*.

Figure S7. MS characteristics and MS/MS spectra of antigenic spliced peptides that would be wrongly assigned as non-spliced peptides.

Figure S8. Schematic of the datasets analyzed.

Figure S9. Impact of the epoxomicin on the GR-LCL cell lysate peptidome.

Figure S10. Characteristic of spliced peptides (blue) compared to non-spliced peptides (orange) identified in the 2D EThcD - GR-LCL HLA-I immunopeptidome.

Figure S11. HLA-I – peptide IC_{50} prediction and experimental measurements in the spliced and non-spliced antigenic peptide pool.

Figure S12. Sequence motifs of spliced and non-spliced antigenic peptides.

Figure S13. Distribution of amino acids in the P_N-P1-P1′- P_C positions of spliced epitopes.

References (*34-44*) are called only in the Supplementary Materials.

Acknowledgements. The project was supported by a grant of the Berlin Institute of Health (BIH, CRG1-TP1) and the Einstein Stiftung Berlin (A2013-174) to P.M.K., by NC3Rs through a David Sainsbury Fellowship to J.L. (NC/K001949/1), by BBSRC (BB/G007934/1), HFSP (RGP0061/2011), The Leverhulme Trust (F/07058/BP) and the Royal Society through a Wolfson Research Merit Award to MPHS. F.M. and A.J.R.H are supported by Proteins@Work (project number 184.032.201), and by the Gravity Program Institute for Chemical Immunology, both funded by The Netherlands Organisation for Scientific Research. We thank Dr. P. Henklein, P.Kunert and

B. Brecht-Jachan for the peptide synthesis. We thank C. Keller, Dr. M. Bassani-Sternberg, and Dr. P. Hitchen for technical support, Dr. S. Islam for his help to set-up the mass spectrometry analysis pipeline and John Cottrell (Matrix Science Ltd) for his technical support with Mascot. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M. (michele.mishto@charite.de) and J.L. (j.liepe@imperial.ac.uk). The RAW files used in this study are available at the Datadryad.org archive (doi:10.5061/dryad.r984n). The scripts for the identification of HLA-I spliced peptides will be available upon request.

Figure legends

Figure 1. Sizes and characteristics of spliced and non-spliced peptide pools. A-B) Summary of the 9-12mer peptides presented by HLA-I molecules on the GR-LCL and C1R cell line, and human primary fibroblasts (**A**), or, as controls, GR-LCL and T2 cell lines (**B**). **C**) Summary of the 9-12mer peptides identified among the cell lysates of T2 and GR-LCL cell lines pre-filtered for peptides smaller than 3 kDa or of C1R cell line pre-filtered for polypeptide bigger than 30 kDa and trypsin digested. Samples are analyzed by different LC-MS/MS methods, as depicted. Light blue shaded areas are part of the spliced peptide pool.

Figure 2. Semi-quantitative comparison of HLA-I eluted spliced and non-spliced peptides. A) Correlation between the MS ion peak area and the concentration of a pool of 75 non-spliced and 78 spliced peptides from two experimental replicates. The correlation coefficients with confidence intervals are reported in the charts. Linear regression was applied and the resulting regression lines with their confidence intervals are depicted for spliced (dark blue line and light blue shaded area) and non-spliced (red line and pink shaded area) peptides. Neither the correlation coefficients nor the parameters of the regression lines significantly differ between spliced and non-spliced peptides. **B**) Distribution of the MS ion peak area of the spliced and non-spliced peptides eluted from the HLA-I molecules of GR-LCL, C1R cell line or human fibroblasts. Dashed lines indicate the median of the distribution for spliced (blue) and non-spliced peptides (red). The MS ion peak area distribution of the non-spliced peptides is significantly larger than the distribution of the spliced peptides (Kolmogorov-Smirnov test; p-value $<10^{-16}$). The relative amount of spliced peptides estimated from the integral of the MS ion peak areas of spliced peptides relative to the integral of the peak area of all peptides is reported. **C**) Medians of the MS ion peak area of spliced and non-spliced peptides identified in the HLA-I immunopeptidome of GR-LCL, C1R cell line or human primary fibroblasts. All samples were not pre-fractioned and were analyzed by EThcD (GR-LCL, C1R cell line and the synthetic peptide pools) or by HCD (human fibroblasts).

Figure 3. Antigens, characteristics and sequence motifs of spliced and non-spliced peptides of the GR-LCL HLA-I immunopeptidome. (**A-E**) refer to the GR-LCL 2D – EthcD HLA-I immunopeptidome. **A)** Number of antigens represented onto the HLA-I molecules by only spliced, only non-spliced or both spliced and non-spliced peptides. **B)** Length distribution of the sequence between the two splice-reactants (*i.e.* intervening sequence). **C)** Length distribution of N- and C-terminal splice-reactants generating the antigenic spliced peptides. **D**) Frequencies of amino acids at each residue of the non-spliced and spliced 9mer peptides. **E)** Distribution of amino acids in the P_N, P1, P1´ and P_C positions (see **Fig. S2**) of the spliced peptides. **D-E**) All frequency values are normalized for the frequency of the amino acids in the human proteome, *i.e.* they indicate the probability of observing a certain amino acid at a certain position given the frequency of this amino acid in the human proteome.

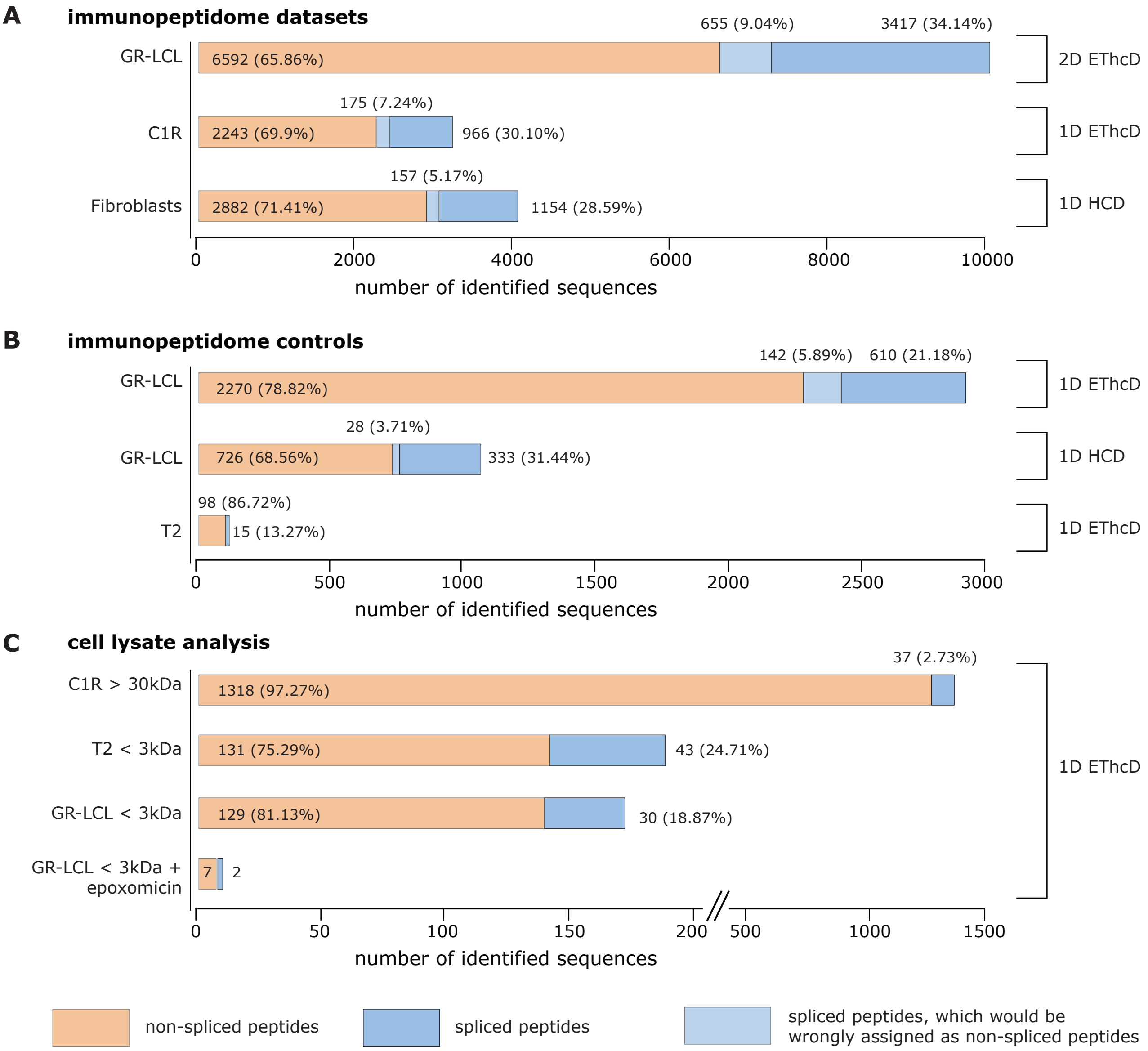
A
immunopeptidome datasets
GR-LCL
6592 (65.86%)
655 (9.04%)
3417 (34.14%)
2D EThcD
C1R
2243 (69.9%)
175 (7.24%)
966 (30.10%)
1D EThcD
Fibroblasts
2882 (71.41%)
157 (5.17%)
1154 (28.59%)
1D HCD
0
2000
4000
6000
8000
10000
number of identified sequences
B
immunopeptidome controls
GR-LCL
2270 (78.82%)
142 (5.89%)
610 (21.18%)
1D EThcD
GR-LCL
726 (68.56%)
28 (3.71%)
333 (31.44%)
1D HCD
T2
98 (86.72%)
15 (13.27%)
1D EThcD
0
500
1000
1500
2000
2500
3000
number of identified sequences
C
cell lysate analysis
C1R > 30kDa
1318 (97.27%)
37 (2.73%)
T2 < 3kDa
131 (75.29%)
43 (24.71%)
GR-LCL < 3kDa
129 (81.13%)
30 (18.87%)
GR-LCL < 3kDa + epoxomicin
7
2
1D EThcD
0
50
100
150
200
500
1000
1500
number of identified sequences
non-spliced peptides
spliced peptides
spliced peptides, which would be wrongly assigned as non-spliced peptides

Figure 1

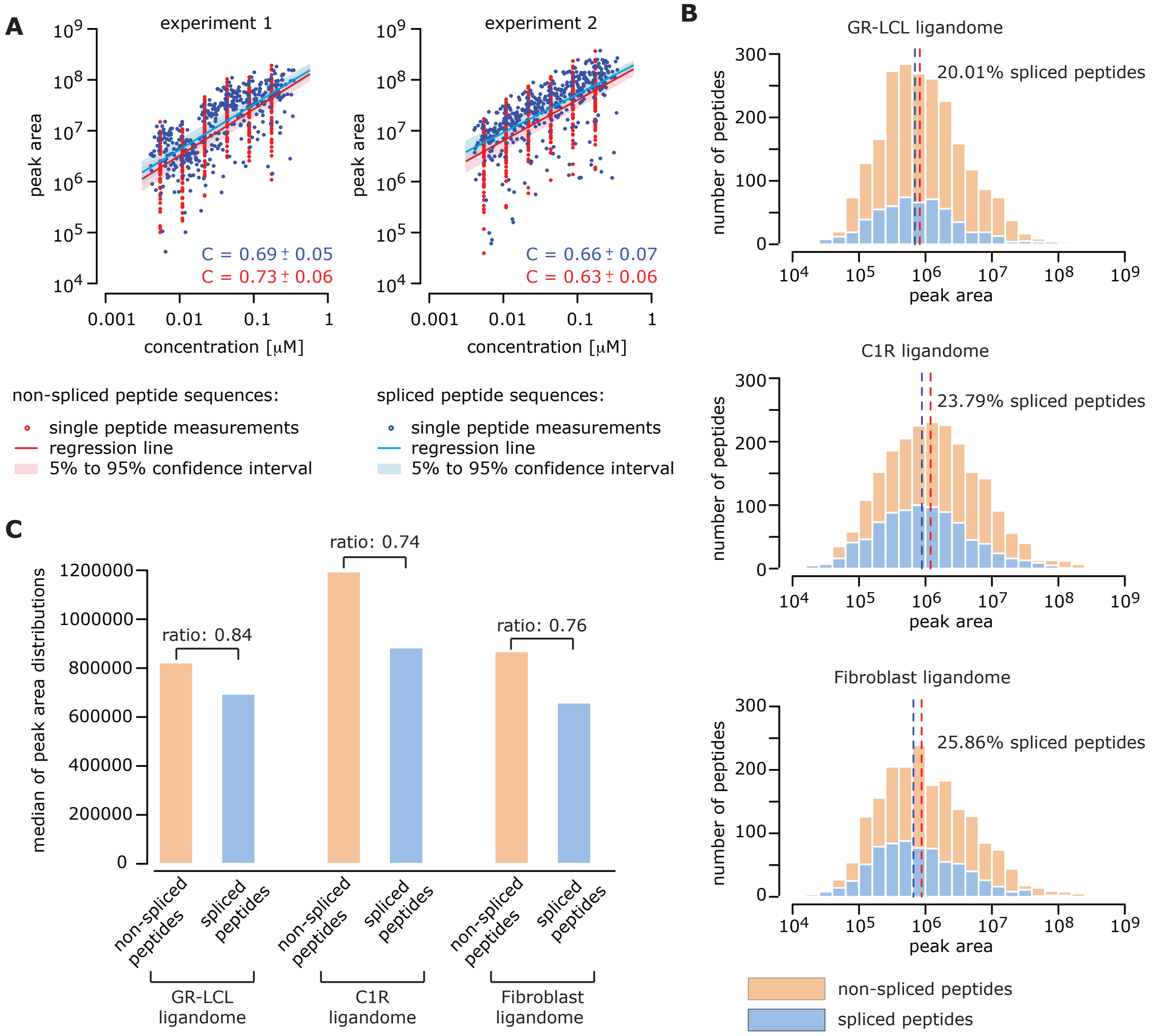
A
experiment 1
peak area
10^4
10^5
10^6
10^7
10^8
10^9
0.001
0.01
0.1
1
concentration [μM]
C = 0.69 ± 0.05
C = 0.73 ± 0.06
experiment 2
peak area
concentration [μM]
C = 0.66 ± 0.07
C = 0.63 ± 0.06
non-spliced peptide sequences:
single peptide measurements
regression line
5% to 95% confidence interval
spliced peptide sequences:
single peptide measurements
regression line
5% to 95% confidence interval
B
GR-LCL ligandome
20.01% spliced peptides
C1R ligandome
23.79% spliced peptides
Fibroblast ligandome
25.86% spliced peptides
number of peptides
0
100
200
300
peak area
non-spliced peptides
spliced peptides
C
median of peak area distributions
0
200000
400000
600000
800000
1000000
1200000
ratio: 0.84
ratio: 0.74
ratio: 0.76
non-spliced peptides
spliced peptides
GR-LCL ligandome
C1R ligandome
Fibroblast ligandome

Figure 2

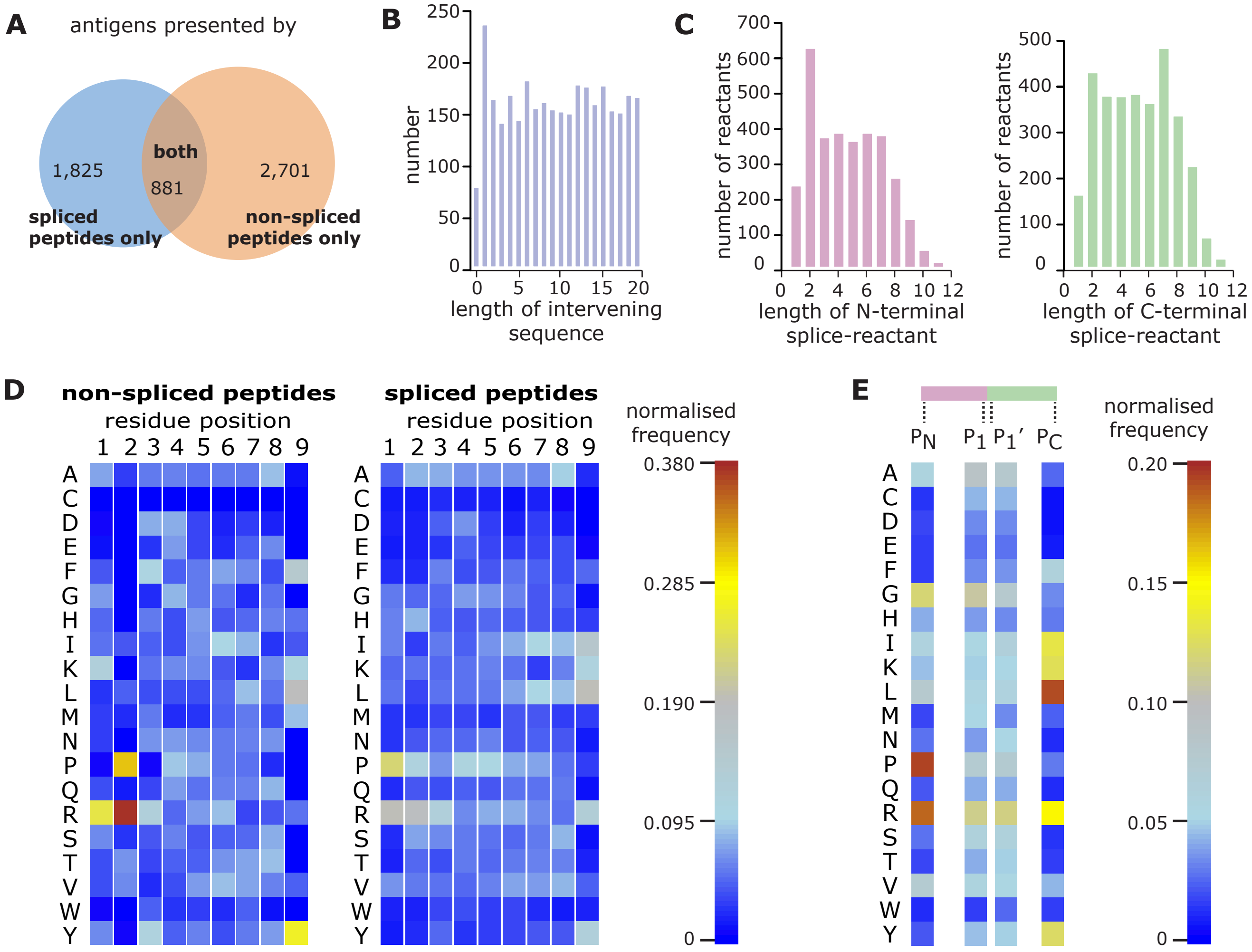
A
antigens presented by
both
1,825
881
2,701
spliced peptides only
non-spliced peptides only
B
number
250
200
150
100
50
0
0 5 10 15 20
length of intervening sequence
C
number of reactants
700
600
500
400
300
200
100
0
0 2 4 6 8 10 12
length of N-terminal splice-reactant
number of reactants
500
400
300
200
100
0
0 2 4 6 8 10 12
length of C-terminal splice-reactant
D
non-spliced peptides
residue position
1 2 3 4 5 6 7 8 9
spliced peptides
residue position
1 2 3 4 5 6 7 8 9
A C D E F G H I K L M N P Q R S T V W Y
normalised frequency
0.380
0.285
0.190
0.095
0
E
P_N P_1 P_1′ P_C
A C D E F G H I K L M N P Q R S T V W Y
normalised frequency
0.20
0.15
0.10
0.05
0

Figure 3