Mapping the MHC class I spliced immunopeptidome of cancer cells

Juliane Liepe, John Sidney, Felix K.M. Lorenz, Alessandro Sette, Michele Mishto

Supplementary Table S1	Characteristics of the cell lines or primary cell used in the study
Supplementary Table S2	List of synthetic peptides used in the study
Supplementary Table S3	Neoepitopes identified in the HCT116 MHC-I immunopeptidome
Supplementary Table S4	Key Resources
Supplementary Table S5	List of spliced and non-spliced peptides identified in the MHC-I immunopeptidomes of the two cancer cell lines and GR-LCL
Supplementary Table S6	List of spliced and non-spliced peptides identified in the MHC-I immunopeptidomes of the two cancer cell lines, whose assigned antigen was not detected in the cell transcriptome
Supplementary Figure S1	Proteasome-catalyzed peptide splicing (PCPS)
Supplementary Figure S2	SPI-delta: MS data analysis pipeline for the identification of spliced and non-spliced peptides in immunopeptidome datasets
Supplementary Figure S3	Characteristics of the spliced and non-spliced peptides identified in the HCT116 and HCC1143 MHC-I immunopeptidomes
Supplementary Figure S4	Sequence motif analysis of the 4 clusters of the HCC1143 immunopeptidome comparing spliced with non-spliced peptide pools
Supplementary Figure S5	MS/MS spectra of the non-spliced neoepitopes
Supplementary Figure S6	Strategy adopted to compute peptide co-localization

Supplemental Experimental Procedures

Relative quantification of spliced and non-spliced peptides

Label-free quantification comparing the amounts of spliced and non-spliced peptides has been performed using the MS ion current peak area for each identified peptide (using Mascot Distiller's label free quantification tools). By applying label-free quantification to each replicate of the HCT116 and HCC1143 cell lines, we have determined the relative differences between the ion peak area distributions of spliced and non-spliced peptides. The overall quantity of a peptide has been computed as the mean ion peak area across the replicates where the peptide has been detected.

We have found that on average spliced peptides represent 19.6 % and 19.3 % of all peptides in term of peptide amount in the HCC1143 and HCT116 cell lines, respectively. The ion peak area distribution of spliced peptides is significant smaller than that of non-spliced peptide (Kolmogorov-Smirnov test, HCC1143 p-value: 0.00156; HCT116 p-value: 0.03) (**Fig. 1B**).

Furthermore, we have computed which peptide has been observed in how many samples (**Fig. S3B-C**). We have found that the ion peak area of detected non-spliced as well as spliced peptides correlates with the number of samples in which the peptides have been observed (fitting analysis of variance model using R function 'aov'; p-value non-spliced peptides < $2.2*10^{-16}$; p-value spliced peptides = $3.14*10^{-16}$ for HCC1143 and HCT116; p-value non-spliced peptides: 0.0098 and 0.06 for HCC1143 and HCT116, respectively). Latter shows that the higher the peptide abundance is, the more likely we can identify the peptide across several biological and technical replicates (because the peptide elution protocol and the MS workflow favor higher abundant peptides and always under-sample the MHC-I immunopeptidome).

We therefore consider all peptides identified only in one sample as the low abundant portion of the immunopeptidome, while we consider peptides identified in 3 or more samples as the high abundant portion of the immunopeptidome. Computation of the spliced peptide frequency depending on the number of samples in which peptides are identified shows that the frequency of spliced peptides is highest in the low abundant portion of the immunopeptidome (around 30%), while this frequency decreases 10-fold in the high abundant immunopeptidome (around 3%) (**Fig. S3D-E**).

Mapping of spliced peptides to their antigens

In order to assign the identified spliced peptide sequences to their original antigens, we have searched the human proteome for all combinations of two non-continuous fragments of the spliced peptide within a given protein (*cis* PCPS only, we have excluded *trans* PCPS in our pipeline). If a spliced peptide sequence could be generated in multiple ways, all possible combinations have been listed (see **Supplementary Table S5**). Latter case is particularly frequent, if one of the two splice-reactants is only one or two amino acids long and if several antigen isoforms exist.

For all analyses, where the information of the peptide mapping has been needed, we have not included ambiguous mappings (*e.g.* length distributions of splice-reactants and determination of PCPS sites). This in turn resulted in the under-presentation of spliced peptides, where one of the two splice-reactants is only one or two amino acids long. Furthermore, our mapping assumes the same restrictions applied to the computation of the spliced peptide database, such as: i. maximum intervening sequence length of 25 amino acids; ii. the exclusion of *trans* PCPS events; iii. the Human reference proteome database. If we were to loosen those restrictions some spliced peptides could potentially be mapped to other spliced peptide origins.

Comparison of MS/MS spectra of synthetic peptides with the neoepitopes detected in the HCT116 immunopeptidome

The MS/MS spectra of the neoepitopes detected in the HCT116 immunopeptidome have been compared to the MS/MS spectra of their synthetic peptide counterparts. We have computed the similarity of two spectra, belonging to the MHC-I eluted peptides and the corresponding synthetic peptides, by computing the Pearson correlation coefficient between the log intensities of identified b-, y-ions (1). The closer this correlation coefficient is to 1, the more correlated are the two MS/MS spectra, *i.e.* the more similar are the intensities and patterns of the

characteristic ions. To note, the eluted data set is generated in a different laboratory with partially different settings and equipment of the MS, compared to the MS/MS spectra generated from synthetic peptides and digestions of synthetic polypeptides. One of the differences concerns the use of acetic acid compared to the use of formic acid in the MS separation system, which could influence both peptide retention time and fragmentation intensities.

Unsupervised clustering analysis of MHC-I-related peptide motifs

To test whether spliced and non-spliced peptides identified in the MHC-I immunopeptidomes of the two cancer cell lines differed in term of sequence motifs, we have focused only on 9mer peptides (**Fig. 2B, S4A**) because they are the predominant subpopulations in the MHC-I immunopeptidomes (**Fig. S3F**). First, we have applied an *in silico* unsupervised clustering approach to assign non-spliced peptides to the MHC-I variants of the cancer cell lines (**Supplementary Table S1**). To do so, we have encoded the non-spliced peptides amino acid sequence by a vector of 195 amino acid indices, resulting in a 1755 entries long vector for each 9mer peptide. In order to identify the driving features that determine the motifs of the non-spliced peptides, we have applied principle component analysis and selected the first 152 principle components (that explained 99 % of the variance in our data set). K-means clustering has been applied to the transformed data - where k = 4 (describing the four HLA-A and HLA-B clusters) for the HCT116 and the HCC1143 cell lines - to determine the peptide clusters that bind to a specific MHC-I molecule. To note, for the HCC1143 cell lines we have set k = 4 even though this cell line has 5 distinct MHC-I allotypes (**Supplementary Table S1**). However, the expected motifs of the HLA-B*35:08 complex are overlapping with the motif of the HLA-C*04:01 complex; therefore, we have considered them as a single cluster. For the HCT116 cell line, we have started with k = 6 to describe all six distinct HLA-I allotypes. Among the resulting six clusters, three have very similar characteristics compared to cluster 3 (**Fig. S4B**). This indicates that k-means clustering in combination with the selected sequence features has not been able to identify six distinct clusters corresponding to one of the six HLA-I allotypes. Therefore, we have repeated the analysis with k = 4 and identified four clusters with motifs corresponding to the known motifs of the HLA-A and HLA-B allotypes of the HCT116 cell line. The reason for failing to identify the two HLA-C allotype clusters could be various. One explanation is that the peptide sequence motifs preferred by the HLA-C*05:01 and HLA-C*07:01 molecules are not very strong, thereby not allowing the recognition of these two clusters in the dataset (2).

For the resulting four clusters of each cancer cell line, we have computed the amino acid motifs of the non-spliced peptides (**Fig. 2B** and **Fig. S4B**). If the identified spliced peptides carried the same sequence characteristics needed for binding to the MHC-I molecules, then it should be possible to describe the spliced peptides by the same clusters. We therefore have applied the same sequence transformation as for non-spliced peptides and assigned each transformed spliced peptide to the cluster possessing the most similar peptide features. The distance has been computed between the spliced peptide features and all cluster centers as sum of squares. The shortest distance determines the cluster for each spliced peptide. As a control, we have generated 1000 random 9mer peptide sequences, which we have assigned to one of the four clusters as described above for spliced peptides.

To analyze the resulting clusters, we have first computed how close all non-spliced peptides within a given cluster are to each other, by computing all pairwise distances as Euclidean distances. The resulting distributions characterize the within-cluster spread and are shown in **Fig. 2A** and **Fig. S4A** (orange lines). To test if the non-spliced peptide clusters and the spliced peptide clusters were in fact similar, we have next computed the pairwise distances between all peptides of the non-spliced clusters with all peptides of the spliced peptide clusters and the random peptide clusters, respectively (**Fig. 2A** and **Fig. S4A**, light blue and grey lines, respectively). The centers of the spliced peptide clusters and of the random peptide clusters are by definition equivalent to the non-spliced peptide clusters. We have found that the observed distances between the spliced and non-spliced peptide clusters are similar compared to the within cluster spread of non-spliced peptides only, showing that both groups of peptides share the same overall driving characteristics for binding to the MHC-I complexes, although some differences emerged. However, the observed distances between the random peptide clusters and the non-spliced peptide clusters is larger compared to the spliced peptide clusters. Comparing the resulting distributions of

distances using the Kolmogorov-Smirnov statistic (to compare empirical distributions), we have found that the random peptide clusters have a significantly larger distance to the non-spliced peptides than the spliced peptides. Next, the sequence motifs of all clusters have been computed and compared between non-spliced and spliced peptides. The sequence motifs have been computed as the Jensen Shannon divergence (JS divergence) between a given cluster and the expected sequence motifs of randomly sampled 9mer peptides under consideration of the amino acid frequency in the human proteome. Furthermore, we have used the JS divergence to compute the difference motif between spliced and non-spliced clusters. Overall, we have found that the sequence motifs of spliced and non-spliced peptides are very similar. However, specific sequence differences emerge (**Fig. 2B, S4B**), which could be best observed in the difference sequence motifs. Additionally, we have computed how frequently the 9mer spliced peptides are spliced after a given position in the peptide (*i.e.* the P1 position; see **Fig. S1**). Latter is shown as inlets in **Fig. 2B** and **Fig. S4B**. The HLA-I alleles corresponding to each cluster are reported, and they have been identified by similarities with known HLA-I-specific peptide sequence motifs.

Transcriptome analysis

Transcriptomic data for the HCT116 and HCC1143 cell lines have been taken from Klijn *et al*. (3). The expression values of the antigens of the mapped spliced and non-spliced peptides have been extracted to verify their expression. We have found that 3979 of 4016 (99.1%) non-spliced peptide sequences that are detected in the HCT116 and HCC1143 cell lines can be assigned to antigens that are also detected at expression level (**Fig. 3B**). Furthermore, we have found that 1139 of 1236 (92.2%) spliced peptide sequences that are detected in the cancer cell line immunopeptidomes can be assigned to antigens that have been detected also at expression level. To note, the transcriptome data are not generated from the same biological material as the immunopeptidome data set. Therefore, some expression values might not be exact.

Among the 97 spliced peptides, whose antigens are not detected in the transcriptome, 18 can derive from other antigens – detected in the transcriptome – if in our analysis we allowed intervening sequences longer than 25 residues (**Supplementary Table S6**). All spliced peptides could be derived from other antigens – detected in the transcriptome – if in our analysis we allowed *trans* PCPS, *i.e.* PCPS between peptide fragments derived from either 2 antigens or two molecules of the same antigen (4).

Computation of MHC-I spliced and non-spliced peptide sampling probability and fold over representation

In the following, we have investigated the relationship between antigen length, abundance and half-life with the probability of an antigen being represented by non-spliced or spliced peptides. We have found in the MHC-I immunopeptidomes of the two cancer cell lines that the number of spliced peptides per antigen correlates with both antigen length (**Fig. 4A**) and intracellular abundance (**Fig. 4B**), as shown for non-spliced peptides by others (2,5,6). To consider both factors, *i.e.* the antigen length and abundance, we have computed for each antigen the MHC-I peptide sampling probability (*D*), which takes into account the antigen length by considering the number of theoretical 9mer peptides that could be generated from an antigen of a given length (**Fig. 4C**), and which approximately indicates the likelihood of an antigen to be represented by a peptide at the cell surface. The MHC-I peptide sampling probability, *D*, is defined as:

D = number of MHC-I peptides / number of theoretical 9mer peptides.

For non-spliced peptides, the number of theoretical 9mer peptides (*N*) has been calculated as (ignoring potential duplicate sequences): $N = L-8$, where L is the antigen length in amino acids (**Fig. 4C,** red line). For spliced peptides, we have counted, for each antigen in the human proteome database, the number of theoretical possible 9mer peptides in our proteome-wide spliced peptide database. All spliced peptide sequences that could be explained through simple peptide hydrolysis (*i.e.* that resemble the sequence of a potential non-spliced peptide) have been removed. Furthermore, all spliced peptides with identical sequences have been considered as a single count for this antigen. Plotting the number of theoretical possible spliced peptides against the antigen length, we have found that both variables show a linear relationship (**Fig. 4C,** blue line). It follows that for a given antigen length, the ratio of the number of theoretical possible spliced peptides (size of spliced peptide database) and the

number of theoretical possible non-spliced peptides is constant and approximately 398 for antigens longer than 500 amino acids (**Fig. 4C,** dashed green line).

We have computed the running average of the MHC-I sampling probability for spliced and non-spliced peptides and found that the MHC-I spliced peptide sampling probability (*D*) increases with increasing antigen abundance (**Fig. 4D**), as it occurs also for non-spliced peptides (**Fig. 4D** and (2)). To note, for those antigens that are represented by both spliced and non-spliced peptides, the MHC-I sampling probability of spliced peptides is correlated with those of non-spliced peptides (**Fig. 4E**). The relationship between MHC-I sampling probability and antigen quantity could be described as a sigmoidal function, *D′=f(antigen abundance)*, for both spliced and non-spliced peptides (see also Bassani-Sternberg et al. (2) for non-spliced peptides). This sigmoidal function could be seen as the expected MHC-I sampling density in dependence of antigen abundance. Antigens with observed MHC-I sampling probabilities above the expected one (*i.e. D/D′>1*) could be considered as over-represented antigens, while antigens with observed MHC-I sampling probabilities below the expected one (*i.e. D/D′<1*) could be considered as under-represented antigens. The fold over representation is therefore defined as *D/D′*. Bassani-Sternberg *et al.* (2) showed that the *D/D′* correlates inversely with the antigen half-life (based on a database by Boisvert *et al.* (7) for mouse proteome) for non-spliced peptides. For the cancer cell lines, the *D/D′* of spliced peptides also inversely correlates with the antigen half-life – as we have confirmed also for non-spliced peptides - independently to the half-life database used in the analysis (Boisvert *et al.* (7) and MsShane *et al*. (8); **Fig. 4F**). Overall, we have found that the presentation of both, spliced and non-spliced peptides, follows similar selection criteria regarding antigen length, abundance and half-life.

Proteome coverage by spliced and non-spliced peptides

We have computed the fraction of the human proteome that is represented through non-spliced and spliced peptides in the extended MHC-I self-immunopeptidome (derived from HCT116 and HCC1143 cancer cell lines, and GR-LCL). We have chosen an approach similar to Pearson *et al*. (5), but we have computed the coverage relative to the proteome rather than relative to the human transcriptome. We have used a sliding window approach, with window sizes of 25 and 50 amino acids. For each protein in the human proteome, we have counted in how many windows we have observed a non-spliced, spliced or any peptide and have compared it to the number of windows in which no peptides have been detected. We have confirmed the results of Pearson *et al.* (5) with our dataset regarding the non-spliced peptides. Interestingly, spliced peptides cover a similar proportion as non-spliced peptides (non-spliced peptides: 9.7%; spliced peptides: 7.2%; using a 50-residue window). Furthermore, the presence of spliced peptides strongly increases the proteome coverage compared to non-spliced peptides alone, therefore confirming that spliced peptides increase the antigenic landscape (**Fig. 5C**). To investigate the spatial localization in the human proteome of peptides represented in the extended MHC-I self-immunopeptidome, we have counted how many non-spliced peptides, spliced peptides, or both peptides are observed in each window. The more peptides are observed per window, the more the peptides are spatially located together (**Fig. 5D**). To note, the values of the proteome coverage for non-spliced and spliced peptides are likely underestimated due to under-sampling of the extended MHC-I self-immunopeptidome. This also explains why the estimated coverage of the proteome by non-spliced peptides is below that of Pearson *et al.* (5), who used a larger dataset for non-spliced peptides.

Analysis of spliced and non-spliced peptides localization within the parental antigen sequence.

The analysis of proteome coverage in the extended MHC-I self-immunopeptidome suggests that spliced and non-spliced peptides might be locally clustered together. To test this hypothesis, we have considered three scenarios: (i) co-localization of non-spliced peptides, (ii) co-localization of spliced peptides, (iii) co-localization of non-spliced with spliced peptides. For all observed peptides, we have computed the shortest distance to its neighboring peptide. We define the distance between two peptides as the number of amino acids between the N-termini of the two peptides. In case of spliced peptides, the distance has been computed always relative to the N-terminus of the splice-reactant that is closest to the N-terminus of the corresponding antigen, no matter if the spliced peptide(s) are non-overlapping, overlapping with the second peptide, or the second peptide is internal (entirely surrounded) by the two splice-reactants of the spliced peptide (**Fig. S6**). If a peptide has more than one

neighboring peptide, always the distance of that peptide to its closest neighbor is considered (**Fig. S6**). The resulting distances are plotted as histograms (**Fig. 5E**). For all three scenarios, we have found highest frequency for short distances, indicating co-localization of peptides in so-called antigenic hotspots. To test whether the co-localization is statistically significant, we have generated a control distribution equivalent to Pearson *et al*. (5): we have randomly placed comparable number of spliced and non-spliced peptides in the proteome and computed their shortest distance to their neighbors. The resulting control distributions are shown as red lines in **Fig. 5E**. The observed distributions are significantly smaller than the respective control distributions (Mann-Whitney test p-values: scenario (i) $p < 1.8*10^{-16}$, scenario (ii) $p = 0.009$, scenario (iii) $p < 5.6*10^{-16}$).

Characteristics of antigens represented by spliced peptides, non-spliced peptides or both in the MHC-I immunopeptidome

The analysis has been done on the extended MHC-I self-immunopeptidome (derived from HCT116 and HCC1143 cancer cell lines, and GR-LCL). The R package "Peptides" has been used to compute the following antigen characteristics: length, hydrophobicity index and isoelectric point.

The link between the hydrophobicity index, the antigen length and the representation by either spliced or non-spliced peptides is depicted in **Fig. 6A**: antigens represented only by non-spliced peptides show decreasing hydrophobicity with increasing antigen length. Conversely, antigens represented only by spliced peptides have generally a higher hydrophobicity than those represented by only non-spliced peptides, independently to their length, and have decreasing hydrophobicity with the increase of the length until approximately 1000 residues, after which the hydrophobicity increases, reaching nearly the same hydrophobicity of very short antigens represented only by spliced peptides (**Fig. 6A**). The hydrophobicity of not represented antigens is independent of the antigen length, and generally higher than the hydrophobicity of represented antigens (black curve in **Fig. 6A**).

The average isoelectric point of the antigens seems to influence the likelihood that antigens are represented by either spliced or non-spliced peptides in a more complex way (**Fig. B-D**). For instance, we have not detected significant differences when we have plotted the average of the isoelectric point against the antigen length (**Fig. 6B)**. However, we have found a trimodal distribution of isoelectric points for all three groups of antigens (**Fig. 6C**). To model the trimodal distribution of isoelectric points, we have used a Gaussian mixture model generated through the EM algorithm by applying the R package "normalmixEM", which has allowed us to split the data into two acidic sets of antigens and one basic set of antigens. Therefore, we have obtained the proportion of antigens that belong to the basic and acid sets, respectively. We have then defined the IP-bias as:

$IP_{bias} = 100*(b-a)/N$, where N is the number of all considered antigens, b and a are the number of antigens in the basic set and acidic set, respectively. The IP-bias is strongest for the group of antigens represented by spliced peptides only (**Fig. 6D**). Latter indicates that basic antigens are more likely represented by spliced peptides, while more acidic antigens are more likely represented by non-spliced peptides.

Overall, we have found that antigens which are short, hydrophilic and acidic are more likely to be represented by non-spliced peptides, while antigens which are long, hydrophobic and basic are more likely to be represented by spliced peptides.

References

1. Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A*, et al.* A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science **2016**;354(6310):354-8.
2. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics **2015**;14(3):658-73.
3. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H*, et al.* A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol **2015**;33(3):306-12 doi 10.1038/nbt.3080.
4. Liepe J, Ovaa H, Mishto M. Why do proteases mess up with antigen presentation by re-shuffling antigen sequences? Curr Opin Immunol **2018**;52:81-6 doi 10.1016/j.coi.2018.04.016.
5. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M*, et al.* MHC class I-associated peptides derive from selective regions of the human genome. J Clin Invest **2016**;126(12):4690-701 doi 10.1172/JCI88590.
6. Hoof I, van Baarle D, Hildebrand WH, Kesmir C. Proteome sampling by the HLA class I antigen processing pathway. PLoS Comput Biol **2012**;8(5):e1002517 doi 10.1371/journal.pcbi.1002517.

7. Boisvert FM, Ahmad Y, Gierlinski M, Charriere F, Lamont D, Scott M*, et al.* A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol Cell Proteomics **2012**;11(3):M111 011429 doi 10.1074/mcp.M111.011429.
8. McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X*, et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. Cell **2016**;167(3):803-15 e21 doi 10.1016/j.cell.2016.09.015.
9. Mommen GP, Frese CK, Meiring HD, van Gaans-van den Brink J, de Jong AP, van Els CA*, et al.* Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). Proc Natl Acad Sci U S A **2014**;111(12):4507-12.
10. Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S*, et al.* An antigenic peptide produced by peptide splicing in the proteasome. Science **2004**;304(5670):587-90.
11. Mishto M, Goede A, Taube KT, Keller C, Janek K, Henklein P*, et al.* Driving Forces of Proteasome-catalyzed Peptide Splicing in Yeast and Humans. Mol Cell Proteomics **2012**;11(10):1008-23.
12. Mishto M, Liepe J. Post-Translational Peptide Splicing and T Cell Responses. Trends Immunol **2017**;38(12):904-15 doi 10.1016/j.it.2017.07.011.
13. Textoris-Taube K, Keller C, Liepe J, Henklein P, Sidney J, Sette A*, 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**;290(51):30417-28.

Supplementary Tables

Cells	Description	Expressed HLA-I allotypes
HCT116	Colon carcinoma cell line (ATCC® CCL-247™) (2)	HLA-A*01:01, -A*02:01, -B*45:01, -B*18:01, -C*05:01, -C*07:01
HCC1143	Breast carcinoma cell line (ATCC® CRL-2321™) (2)	HLA-A*31:01, -A*31:01, -B*35:08, -B*37:01, -C*04:01, -C*06:02
GR-LCL	Human EBV-immortalized lymphoblastoid cell line (9)	HLA-A*01:01, -A*03:01, -B*07:02, -B*27:05, -C*02:02, -C*07:02

Supplementary Table S1. Characteristics of the cell lines or primary cell used in the study.

Peptide	Sequence	Description	MHC-I-peptide binding affinity (IC_{50} in nM)	Figure/ Table
$CHMP7_{312-330}$	RIYASQTDQMVFN AYQAGVG	*In vitro* digestion assay		Fig. S5
$mutCHMP7_{312-330}$	RIYASQTDQMVFN TYQAGVG	*In vitro* digestion assay; A324T mutation		Fig. S5
$CHMP7[A324T]_{316-325}$	QTDQMVFNTY	Detected in the MHC-I immunopeptidome, produced *in vitro* by proteasome, binding MHC-I variants; non-spliced neoepitope; A324T mutation	HLA-A*01:01 (1), HLA-B*18:01 (29)	Fig. 3, Table S3, Fig. S5
$RBBP7_{6-25}$	FEDTVEERVINEEY KIWKK	*In vitro* digestion assay		Fig. S5
$mutRBBP7_{6-25}$	FEDTVEERVIDEEY KIWKK	*In vitro* digestion assay; N17D mutation		Fig. S5
$RBBP7[N17D]_{12-20}$	EERVIDEEY	Detected in the MHC-I immunopeptidome, produced *in vitro* by proteasome, binding MHC-I variants; non-spliced neoepitope; N17D mutation	HLA-A*01:01 (324), HLA-B*18:01 (0.4)	Fig. 3, Table S3, Fig. S5

Supplementary Table S2. List of synthetic peptides used in the study. The mutations are depicted in red.

Neoepitope	Sequence	Gene's origin	MHC-I-peptide binding affinity (IC_{50} in nM)
CHMP7[A324T]$_{316-325}$	[QTDQMVFNTY]	Q8WUX9	HLA-A*01:01 (1), HLA-B*18:01 (29)
RBBP7[N17D]$_{12-20}$	[EERVIDEEY]	Q16576	HLA-A*01:01 (324), HLA-B*18:01 (0.4)

Supplementary Table S3. Neoepitopes identified in the HCT116 MHC-I immunopeptidome. The origin, sequence and measured binding affinity to one or more MHC-I variants of the HCT116 cell line (see **Supplementary Table S1**) of two neoepitopes identified in the MHC-I immunopeptidome of the HCT116 cell line. The cancer-specific mutations are labeled in red. The binding affinity between the MHC-I variants and the synthetic peptides has been measured as described in the Materials & Methods. Only binding affinities of 1000 nM or better are shown. The two non-spliced neoepitopes identified with our strategy are among the five neoepitopes identified by Bassani-Sternberg *et al.* (2). The other three non-spliced neoepitopes identified in the latter study did not pass our 1% FDR cut-off filter applied during the assignment of MS/MS spectra. We have not identified any neoepitope in the MHC-I immunopeptidome of the HCC1143 cell line. The generation of the two neoepitopes by the proteasome have been verified by *in vitro* digestion, using purified proteasome, of the synthetic substrates mutCHMP7$_{312-330}$, and mutRBBP7$_{6-25}$, respectively. The digestion of the corresponding wild type synthetic polypeptides did not lead to a detectable generation of the corresponding wild type peptides, *i.e.* CHMP7$_{316-325}$, RBBP7$_{12-20}$ (data not shown).

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
None		
Bacterial and Virus Strains		
None		
Biological Samples		
LysC-trypsin digestion of HCC1143 intracellular proteome > 30 kDa	HCC1143 cell line in culture	
In vitro digestion of synthetic polypeptide mutCHMP7$_{312-330}$ [A324T] with purified proteasome	Peptide: Peptide Synthesis Facility of the Charité. Proteasome: in house	
In vitro digestion of synthetic polypeptide mutRBBP7$_{6-25}$ [N17D] with purified proteasome	Peptide: Peptide Synthesis Facility of the Charité. Proteasome: in house	
synthetic peptide mix containing the neoepitopes CHMP7[A324T]$_{316-325}$ and RBBP7[N17D]$_{12-20}$	Peptide: Peptide Synthesis Facility of the Charité.	
In vitro digestion of synthetic polypeptide mutCHMP7$_{312-330}$ [A324T] without purified proteasome	Peptide: Peptide Synthesis Facility of the Charité.	
In vitro digestion of synthetic polypeptide mutRBBP7$_{6-25}$ [N17D] without purified proteasome	Peptide: Peptide Synthesis Facility of the Charité.	
Chemicals, Peptides, and Recombinant Proteins		
None		
Critical Commercial Assays		
None		
Deposited Data		
New datasets generated in the study:		
filteredSearchResults.xlsx	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
synthetic_peptides.raw	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
mutCHMP7_20h.raw	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
mutCHMP7_20h_no-proteasome.raw	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
mutRBBP7_20h.raw	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
mutRBBP7_20h_no-proteasome.raw	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
LyS-tryp_HCC1143_30KDa.raw	Mendeley archive	http://dx.doi.org/10.17632/y2cvb5nvgn.1
Previously published datasets/databases used in the study are:		
20120321_EXQ1_MiBa_SA_HCC1143_1.raw	PRIDE archive	PXD000394
20120321_EXQ1_MiBa_SA_HCC1143_2.raw	PRIDE archive	PXD000394
20120322_EXQ1_MiBa_SA_HCC1143_1_A.raw	PRIDE archive	PXD000394
20120515_EXQ3_MiBa_SA_HCT116_mHLA-1.raw	PRIDE archive	PXD000394
20120515_EXQ3_MiBa_SA_HCT116_mHLA-2.raw	PRIDE archive	PXD000394
20120617_EXQ0_MiBa_SA_HCT116_1_mHLA_2hr.raw	PRIDE archive	PXD000394
20120619_EXQ6_MiBa_SA_HCT116_2_mHLA_2hr.raw	PRIDE archive	PXD000394
GRLCL immunopeptidome	Datadryad.org archive	doi:10.5061/dryad.r984n
HCC1143_mutation_cosmic_database	Cosmic	Cosmic version 17/8/2016; also available at: http://dx.doi.org/10.17632/y2cvb5nvgn.1

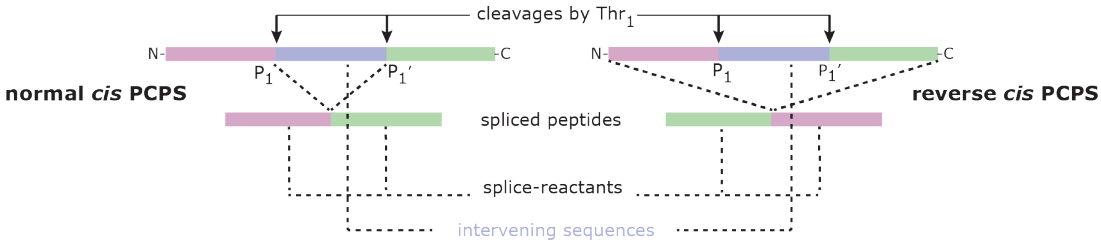
HCT116_mutation_cosmic_database	Cosmic	Cosmic version 17/8/2016; also available at: http://dx.doi.org/10.17632/y2cvb5nvgn.1
HCT116 and HCC1143 expression array	EMBL	E-MTAB-2706
Experimental Models: Cell Lines		
HCC1143	ATCC	ATCC-CRL2321
Experimental Models: Organisms/Strains		
None		
Oligonucleotides		
None		
Recombinant DNA		
None		
Software and Algorithms		
R	Team R, 2014	
Mascot software	Matrix Science Ltd	version 2.6.1.
I-Tasser	Yang et al., 2015	
Pymol	The PyMOL Molecular Graphics System	Version 1.7.4

Supplementary Table S4. Key resources.

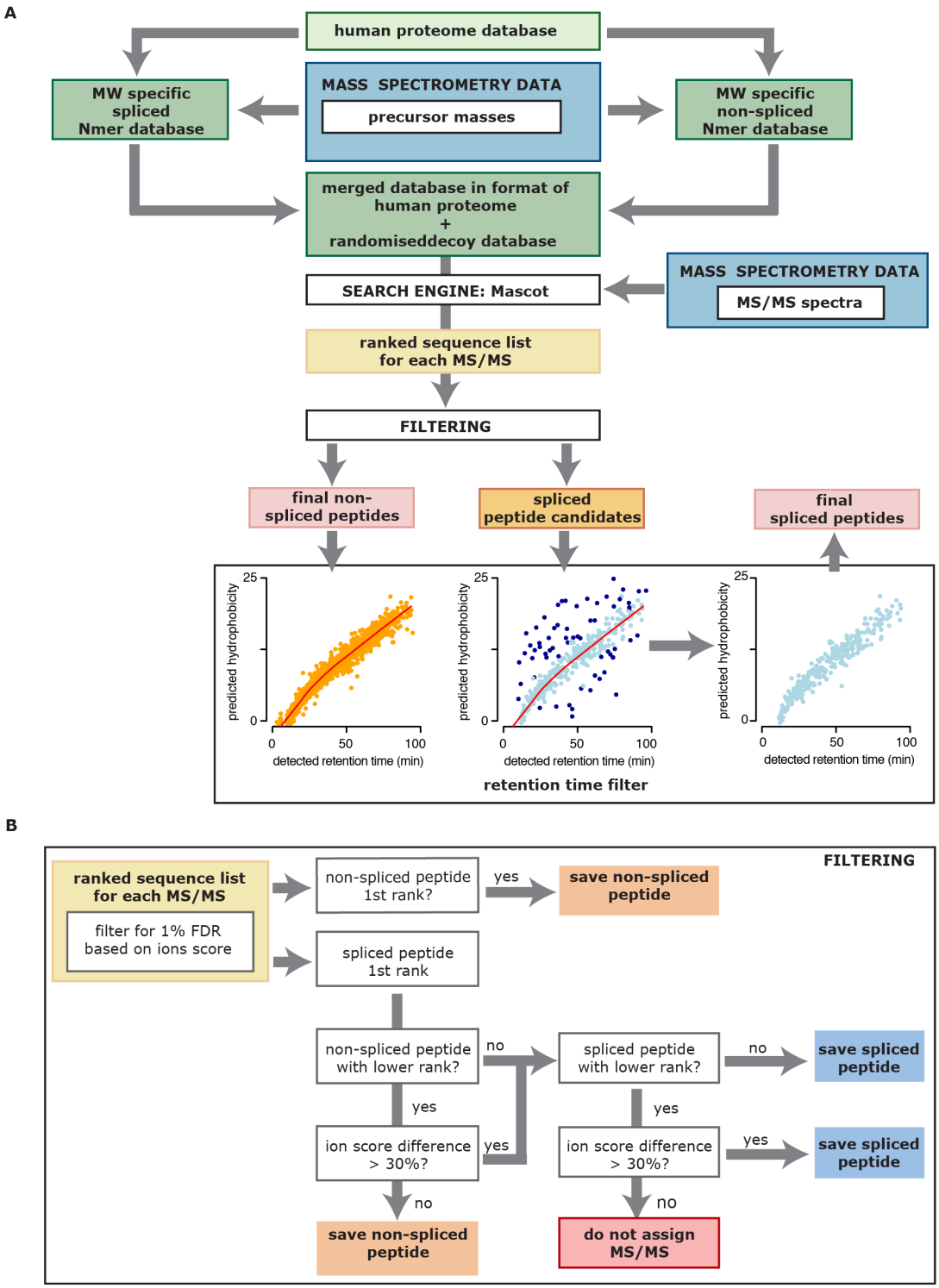
Supplementary Table S5. List of spliced and non-spliced peptides identified in the MHC-I immunopeptidomes of the two cancer cell lines and GR-LCL. The peptide sequence, the assigned antigen and the location within the antigen are reported for the HCT116 and HCC1143 and the GR-LCL 2 D immunopeptidomes. The table is reported as excel file in the Supplementary material (Supplementary Table_S5.xlsx).

Supplementary Table S6. List of spliced and non-spliced peptides identified in the MHC-I immunopeptidomes of the two cancer cell lines, whose assigned antigen was not detected in the cell transcriptome. Here the peptide sequence, the assigned antigen, the location within the antigen, and the alternative antigens and locations (if we allowed intervening sequence lengths longer than 25 residues) are reported for each cell line. The table is reported as excel file in the Supplementary material (Supplementary Table_S6.xlsx).

Supplemental Figures

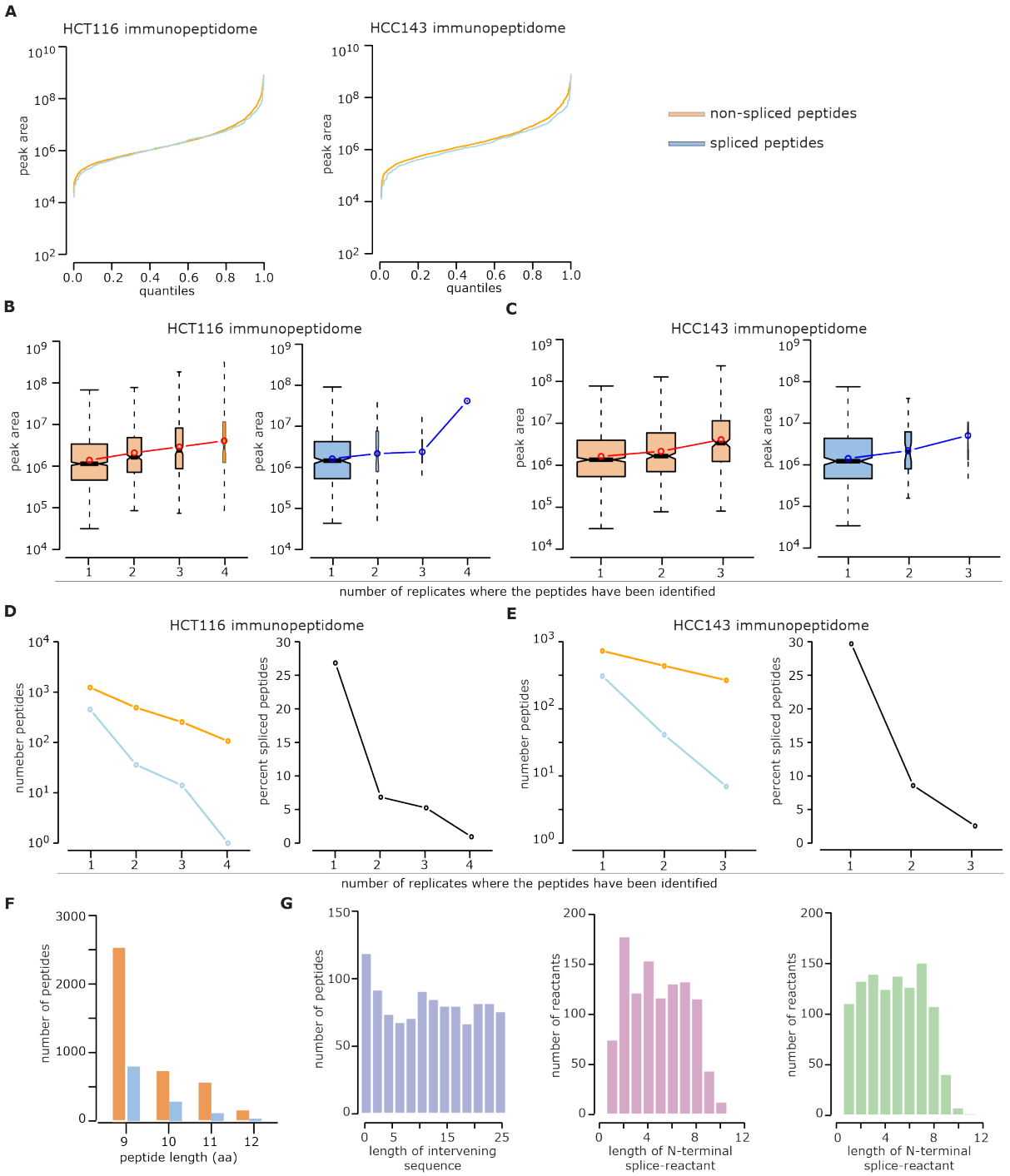


Supplementary Figure S1. Proteasome-catalyzed peptide splicing (PCPS). Spliced peptides can be formed by *cis* PCPS, *i.e.* when the two splice-reactants derive from the same polypeptide molecule. The ligation of the splice-reactants can occur in normal order, *i.e.* following the orientation from N- to C-terminus of the parental protein (normal *cis* PCPS), or in the reverse order (reverse *cis* PCPS). According to the transpeptidation model, the proteasome's catalytic Thr_1 breaks the peptide bond after the residue (P1) of the protein - thereby forming an acyl-enzyme intermediate with the N-terminal splice-reactant, coupled to the release of the intervening sequence – and, instead of catalyzing the canonical peptide hydrolysis, it catalyzes the ligation between the P1 residue of one splice-reactant with the residue P1´ of the other splice-reactant (10-12).



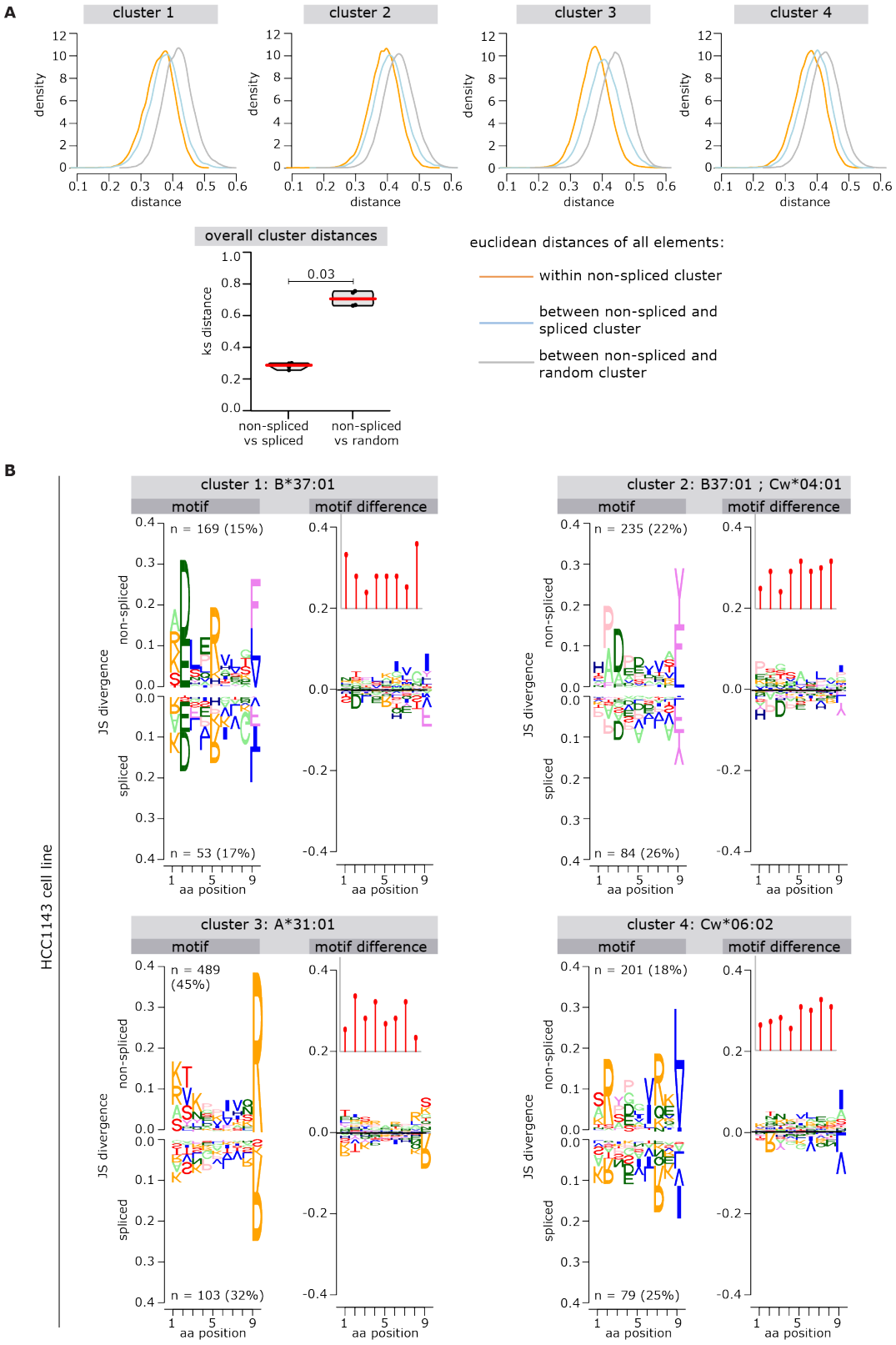
Supplementary Figure S2. SPI-delta: novel MS data analysis pipeline for the identification of spliced and non-spliced peptides in immunopeptidome datasets. (**A**) Based on the human proteome reference database all Nmer (here 9-12mer) spliced and non-spliced peptides and their molecular weight (MW) are computed with the restriction of maximum 25 amino acids intervening sequence length and allowing only *cis* PCPS. This spliced and non-spliced peptide database is filtered for all peptides with MW that matches an observed precursor mass in the MS data set, resulting in MW-specific spliced and non-spliced Nmer databases. Both spliced and non-spliced peptide databases are merged and transformed as described in Materials & Methods, resulting in the target database containing all MW-matched spliced and non-spliced peptide sequences. A decoy database is generated by randomizing the target database. The search engine Mascot is applied to assign peptides to

MS/MS spectra, which are subsequently filtered for 1% FDR based on the Mascot ion score. This results in a ranked list of possible peptide sequences assigned to each MS/MS spectrum. Each ranked list is filtered as shown in (**B**). Filtering results in the final list of identified non-spliced peptides and a list of assigned spliced peptide candidates. For both lists, the peptide hydrophobicity is predicted and correlated with the observed retention time of the assigned MS/MS spectra. The computation of the running average of predicted hydrophobicities in dependence of the observed retention times of non-spliced peptides (orange dots and red line) and the estimation of the variance from this running average allow subsequent filtering of likely wrong spliced peptide assignments (dark blue dots), thereby resulting in the final list of identified spliced peptides. (**B**) In most proteomics or peptidomics approaches the top hit (rank 1) is considered the true hit. Here, we amend this to increase accuracy in the spliced peptide identification. The ranked list is filtered based on the rules shown as flow chart. If a non-spliced peptide has rank 1 we assign this non-spliced peptide sequence to the MS/MS spectrum. If on the contrary a spliced peptide has rank 1, we first check if another non-spliced peptide with lower rank – i.e. with rank 2 or lower - exists. If the ion score difference between the rank 1 spliced peptide and the lower rank non-spliced peptide is less than 30%, we consider the non-spliced peptide as true hit. If the ion score difference is, however, larger than 30%, the non-spliced peptide is not assigned. We then check if another spliced peptide with lower rank exists, in which case we require again a minimum difference in ion scores of 30% between the rank 1 spliced peptide and the lower rank spliced peptide to assign the rank 1 spliced peptide to the MS/MS spectrum. If the ion score difference is below 30% we do not assign this MS/MS spectrum. In this way, we remove all potential spliced peptide assignments where we cannot determine the exact sequence with high confidence.



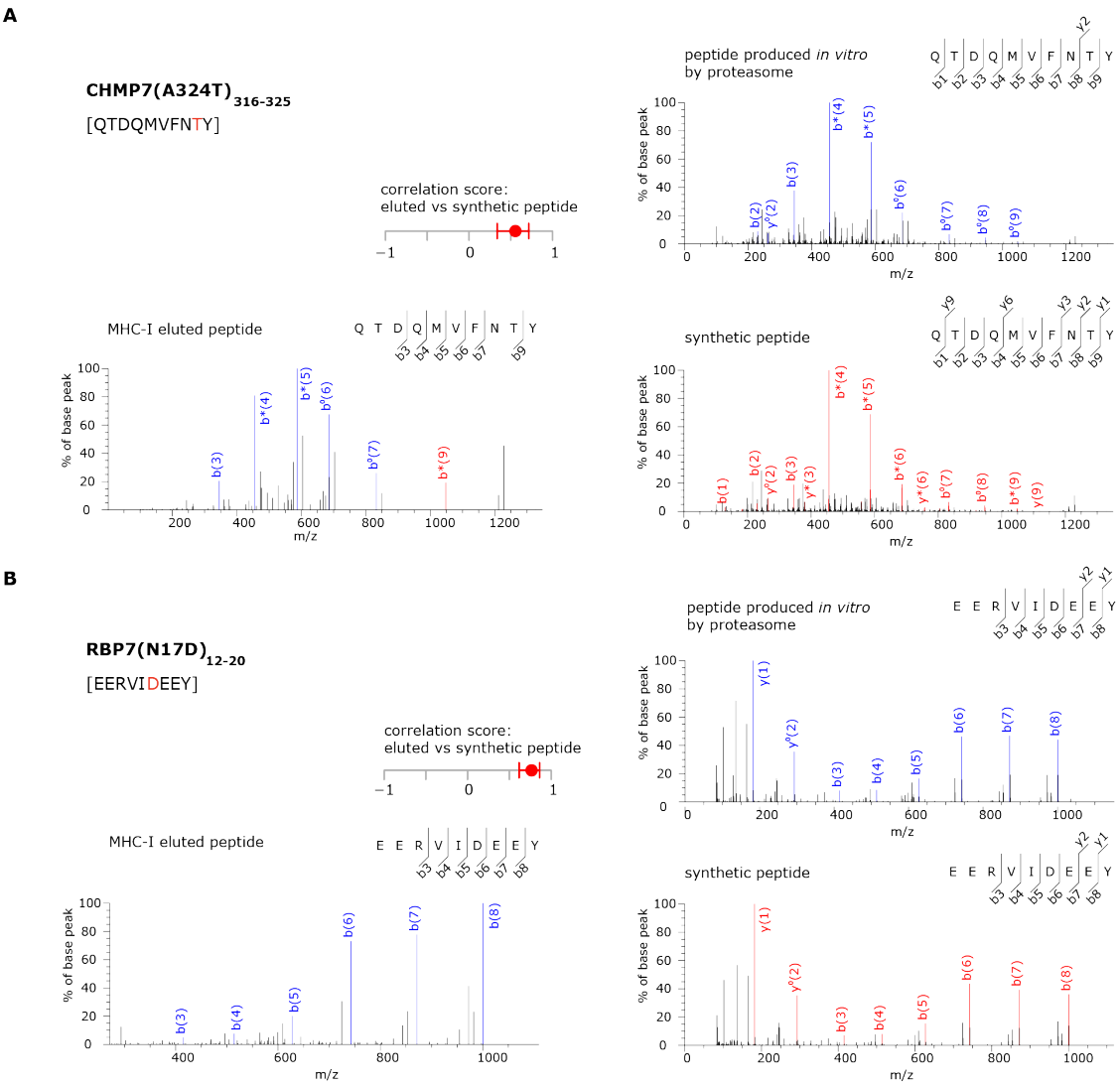
Supplementary Figure S3. Characteristics of the spliced and non-spliced peptides identified in the HCT116 and HCC1143 MHC-I immunopeptidomes. (**A-C**) Relative quantification of spliced and non-spliced peptides in the overall cancer cell line immunopeptidomes and in the technical replicates. (**A**) Distribution of the ion peak area of all spliced and non-spliced peptides, respectively. (**B,C**) Ion peak areas for peptides detected in 1 to 4 replicates for the HCT116 (**B**) and HCC1143 (**C**) immunopeptidomes shown as boxplots; boxes indicate 25% and 75%-tile, dashed bars indicate 5% and 95%-tile, red and blue lines indicate the means. The width of the boxes is proportional to the number of peptides in each group (linked to **Fig. 1B**). (**D,E**) Number of peptides detected in 1 to 4 replicates and the frequency of detected spliced peptides compared to all identified peptides in dependence of the occurrence in number of replicates for the HCT116 (**D**) and HCC1143 (**E**) (linked to **Fig. 1B**). (**F**) Length distribution of spliced and non-spliced peptides identified in the joined MHC-I immunopeptidomes of the two cancer cell lines. The antigenic spliced peptides' length distribution does not differ to that of non-spliced peptides. (**G**) Length distribution of intervening sequences, N- or C-terminal splice-reactants in the joined MHC-

I immunopeptidomes of the two cancer cell lines. The length of the intervening sequences (see **Fig. S1**) varies from 0 to 25 residues without a clear preferred length range. Spliced peptides with intervening sequence longer than 25 residues are not included in the human spliced proteome database used in this study.



Supplementary Figure S4. Sequence motif analysis of the 4 clusters of the HCC1143 immunopeptidome comparing spliced with non-spliced peptide pools. (**A**) Distribution of the distances within the cluster of non-spliced peptides (orange), between spliced and non-spliced peptides (light blue) and between non-spliced peptides and control peptides in the 4 clusters. Spliced and non-spliced peptides have been identified in the MHC-I immunopeptidome of the HCC1143 cell line. The lower panel shows the Kolmogorov-Smirnov distance between the distributions of non-spliced and spliced peptides and control peptides, respectively, which are significant different (p value = 0.03). (**B**) Comparison between the amino acids frequencies for each

position of the non-spliced and spliced 9mer peptides after clustering according to their amino acid features. For each of the 4 clusters, the amino acids frequencies are shown in the left panels. The size of the amino acid letters corresponds to their occurrence within the cluster. The number of peptides belonging to each cluster is also reported. On the right panels, the motif difference between the amino acid frequencies of the non-spliced and spliced 9mer peptides is reported as JS divergence. The inlets on the top of the right panels show the frequency of PCPS (as P1 position) for each residue. The HLA-I alleles corresponding to each cluster are reported, and they have been identified by similarities with known HLA-I-specific peptide sequence motifs.



Supplementary Figure S5. MS/MS spectra of the non-spliced neoepitopes. (**A,B**) MS/MS spectra of the non-spliced neoepitopes as observed in the MHC-I immunopeptidomes (left panel), in *in vitro* digestions of synthetic substrates by purified proteasome (top panel), and by measuring the synthetic peptides (lower panel). Detected *m/z* and charges in the MS/MS spectra from eluted peptides and from *in vitro* digestions are indicated in blue, if they are also detected in the MS/MS spectrum of the corresponding synthetic peptide. All other assigned *m/z* are indicated in red. In the spectra, charged b- and y-ions are reported. Double charged ions are marked as $^{++}$. Ions' neutral loss of water and of ammonia are symbolized by ^{o} and *, respectively. The correlation score between the MS/MS spectra from eluted peptides and the MS/MS spectra of the synthetic peptide is displayed with its confidence intervals on top of the MS/MS spectra from eluted peptides. Correlation scores above 0 indicate matching spectra.

The corresponding wild type non-spliced peptides are not produced by proteasome by degrading the wild type antigen sequences (data not shown), as already observed for other mutations (13). In the control experiments, leaving the synthetic substrates at 37°C for 20h in TEAD buffer and in absence of proteasome, none of the target peptides have been identified by MS (data not shown). The latter results exclude autocatalytic cleavage events generating the target peptides during the *in vitro* reactions at 37°C.

Supplementary Figure S6. Strategy adopted to compute peptide co-localization. To calculate the co-localization of spliced and/or non-spliced peptides we need to compute the distance between peptides. Shown is a scheme that illustrates the distance between two peptides (spliced or non-spliced) for several possible scenarios of localization (not all scenarios are shown).