Elucidating the signatures of proteasome-catalysed peptide splicing

Wayne Paes[1*], German Leonov[2], Thomas Partridge[1], Annalisa Nicastri[3], Nicola Ternette[3], Persephone Borrow[1*]

[1]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, OX3 7FZ, United Kingdom

[2]York Cross-disciplinary Centre for Systems Analysis, University of York, York, YO10 5DD, United Kingdom

[3]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, OX3 7BN, United Kingdom.

*Correspondence: wayne.paes@ndm.ox.ac.uk; persephone.borrow@ndm.ox.ac.uk

Running title: *Signatures of proteasomal splicing*

Abstract

Proteasomes catalyse the degradation of endogenous proteins into oligopeptides, but can concurrently create spliced oligopeptides through ligation of previously non-contiguous peptide fragments. Recent studies have uncovered a formerly unappreciated role for proteasome-catalysed peptide splicing (PCPS) in the generation of non-genomically templated major histocompatibility complex class I (MHC-I)-bound *cis*-spliced peptides that can be targeted by $CD8^+$ T cells in cancer and infection. However, the mechanisms defining PCPS reactions are poorly understood. Here, we experimentally define the biochemical constraints of proteasome-catalysed *cis*-splicing reactions by examination of *in vitro* proteasomal digests of a panel of viral- and self-derived polypeptide substrates using a tailored mass-spectrometry-based *de novo* sequencing workflow. We show that forward and reverse PCPS reactions display unique splicing signatures, defined by preferential fusion of distinct amino acid residues with stringent peptide length distributions, suggesting sequence- and size-dependent accessibility of splice reactants for proteasomal substrate binding pockets. Our data provide the basis for a more informed mechanistic understanding of PCPS that will facilitate future prediction of spliced peptides from protein sequences.

Keywords: peptide splicing/proteasome/splicing mechanism/peptide epitope

Introduction

Proteasomes are multi-subunit enzyme complexes in eukaryotic cells that selectively degrade "unwanted" endogenous proteins into oligopeptides. Their activity is important for protein quality control and for regulation of many intracellular processes including cell cycle progression, signalling pathways and transcription (Saeki & Tanaka, 2012). The oligopeptides generated by proteasomes can be degraded to provide a source of amino acids (aas) for protein synthesis, but a subset of these peptides is translocated from the cytoplasm into the endoplasmic reticulum (ER) by the transporter associated with antigen presentation (TAP), where they may associate with newly-synthesised major histocompatibility complex class I (MHC-I) molecules (Murata, Takahama et al., 2018). Peptide-loaded MHC-I molecules then traffic to the cell surface for display to $CD8^+$ T cells, enabling immunosurveillance of tumours and infected cells.

The peptides comprising the cellular MHC-bound repertoire (the immunopeptidome) were classically thought to be contiguous sequences originating from self- or foreign proteins. However, following initial anecdotal reports of recognition of non-templated peptides generated by proteasome-catalysed "cut-and-pasting" of non-contiguous fragments of a polypeptide by tumour-specific $CD8^+$ T cells (Hanada, Yewdell et al., 2004, Vigneron, Stroobant et al., 2004, Warren, Vigneron et al., 2006), increasing evidence has accumulated to support the concept that proteasome-catalysed peptide splicing (PCPS) reactions comprise an additional source of peptides that can be presented on MHC-I molecules for $CD8^+$ T cell recognition (Dalet, Vigneron et al., 2010, Ebstein, Textoris-Taube et al., 2016, Faridi, Li et al., 2018, Liepe, Marino et al., 2016, Michaux, Larrieu et al., 2014, Mylonas, Beer et al., 2018, Paes, Leonov et al., 2019, Platteel, Mishto et al., 2016, Platteel, Liepe et al., 2017). In addition to generating peptides composed of fragments from within the same protein/polypeptide via

cis-splicing, proteasomes have also been suggested to catalyse *trans*-splicing of fragments from polypeptides derived from separate proteins *in vitro* (Faridi et al., 2018). However, the contribution of *trans*-spliced peptides to the immunopeptidome remains controversial and no examples of $CD8^+$ T cell recognition of *trans*-spliced epitopes have been reported to date, leaving the *in vivo* significance of proteasome-catalysed *trans*-splicing unclear. Interestingly, $CD4^+$ T cell recognition of MHC-II-bound epitopes composed of proinsulin peptides fused to other peptides present in beta-cell secretory granules has been implicated in the pathogenesis of diabetes (Delong, Wiles et al., 2016), although the enzyme(s) responsible for the generation of these spliced epitopes were not identified.

The constitutive proteasome (CP) and the immunoproteasome (IP) both mediate PCPS reactions (Dalet et al., 2010, Ebstein et al., 2016), and the thymoproteasome has recently also been shown to do so (Kuckelkorn, Stubler et al., 2019). PCPS is initiated at proteasomal active sites by catalytic threonine residues which perform nucleophilic attack on carbonyl groups within an unfolded polypeptide chain, and can occur in either a forward or reverse sense (Dalet, Robbins et al., 2011, Dalet et al., 2010, Michaux et al., 2014, Warren et al., 2006). This forms an acyl-enzyme intermediate (contained in the N-terminal portion of the final spliced peptide) which is tethered to the proteasome by an ester linkage (Vigneron, Ferrari et al., 2017). Nucleophilic attack of the acyl-enzyme intermediate by a free amine group liberated by proteasomal cleavage of a non-adjacent peptide fragment within the precursor substrate then hydrolyses the ester linkage, appending the C-terminal portion of the final spliced peptide product (Vigneron et al., 2004). This process is likely to be governed by inherent properties of each proteasomal active site as well as those of the constituent splice partners such as sequence, length and affinity for the substrate binding sites. However, these constraints remain undefined.

Although progress has been made in understanding proteasomal cleavage specificities during events generating non-spliced peptides (Kuckelkorn et al., 2019, Mishto, Liepe et al., 2014, Winter, La Greca et al., 2017), the physical constraints and biochemical preferences governing peptide splicing reactions remain undefined. Systematic analysis of PCPS reactions generating a source of peptides for MHC-I presentation has been hindered by the limited number of fully-validated spliced epitopes identified to date. The only study attempting to experimentally assign preferred 'rules' for PCPS was limited by its restriction to the binding preferences of HLA-A*02:01, and only considered the ligation efficiency of N- and C-terminal sequence variants of fixed lengths in a single 9-mer peptide (Berkers, de Jong et al., 2015).

Here, by employing a liquid chromatography tandem mass spectrometry (LC-MS/MS)-based *de novo* sequencing workflow to analyse the *in vitro* digestion products of a panel of unique polypeptide substrates by the 20S CP, we have experimentally elucidated the biochemical characteristics of *cis*-PCPS reactions. We document preferred splice partner length distributions, *cis*-splicing distances and aa enrichments proximal to splice sites that are specifically attributable to 20S proteasomal activity and not confounded by other *in cellulo* antigen processing events, thus shedding light on the biochemical constraints inherent in both forward and reverse *cis*-PCPS reactions. Elucidation of these splicing signatures will enable refinement of a mechanistic basis for PCPS, with the eventual aim of facilitating the prediction of spliced peptides from protein sequences to enable study of their roles in antigen presentation and putative contribution to broader aspects of cell biology.

Results and Discussion

The relative proportions of unique non-spliced and *cis*-spliced peptides generated by the constitutive proteasome are dependent on precursor peptide length

To enable characterisation of non-spliced and *cis*-spliced peptides generated by the 20S CP, 19 unique polypeptides (14-44aa long, median = 22aa) were individually incubated with the CP for 2h or 20h at 37°C. Of the 19 polypeptides studied, 8 were derived from the clinically-important viral pathogen human immunodeficiency virus type 1 (HIV-1), 8 were "self" peptides derived from human protein sequences in the SwissProt reference proteome, and 3 originated from a publicly available short open reading frame database (Olexiouk, Crappe et al., 2016) (**Table EV1**). The total aa content of these polypeptides closely reflected that of the combined HIV-1 and human proteomes (**Table EV2**). Following digestion, purification and *de novo* sequencing of proteolytic products from precursor substrates by MS, identification of both contiguous and spliced peptide products was achieved using a bioinformatics workflow [(Paes et al., 2019); adapted as described in the Materials and Methods]. Control digestion reactions (lacking the CP) were conducted in parallel to enable removal of background products arising from non-proteasomal polypeptide degradation or peptide synthesis artefacts.

During proteasomal generation of non-spliced peptides, it was previously shown that saturation of cleavage site usage was achieved at 480 minutes (Winter et al., 2017). In line with this, saturation of available cleavage sites was not achieved at the 2h timepoint in our study, as for the majority of those polypeptides where the precursor was observed at both timepoints, the undigested precursor underwent a decrease in abundance (as determined from LC-MS/MS intensity values) from 2 to 20h (**Figure EV1A**). In addition, we observed a 13.3% increase in diversity of non-spliced peptides at 20h (**Figure EV1B**). However, the 2h timepoint was selected for analysis in order to minimise potential proteasomal re-entry of cleavage products and ensuing *trans*-splicing events, which could confound characterisation of *cis*-splicing products.

The full repertoire of unique spliced and non-spliced peptides generated following a 2h proteasomal digest is presented in **Appendix Tables S1-S2**. Overall, we observed a total of 679 unique non-spliced (50.5%) and 665 *cis*-spliced (49.5%) peptides (**Figure 1A**). The relative proportions of spliced and non-spliced peptides within all unique peptide products were also similar at the 20h timepoint, where 769 unique non-spliced (55.1%) and 625 *cis*-spliced (44.9%) peptides were observed (**Figure EV1C**), indicating that spliced peptide generation was not under-estimated by consideration of the earlier 2h timepoint. Interestingly, however, although spliced and non-spliced peptides constituted approximately equal proportions of the unique peptides generated following a 2h digest of all 19 precursor polypeptides, spliced peptides constituted only 14.4% of all unique peptides generated following digestion of the 5 longest precursors (27-44aa), and conversely constituted 88.4% of all unique peptides generated from the 5 shortest precursors (14-18aa) **(Figure 1A)**. Furthermore, within the digestion products from precursors of all lengths, when peptide fragments containing the N- or C-terminal aa of the original polypeptide and those generated from within the polypeptide substrate (that did not then contain the N- or C-terminal aa residues) were considered separately, spliced peptides constituted a markedly lower proportion of all unique peptides in the latter category, and a higher proportion of the peptide products containing one of the original peptide termini (**Figure EV1D**). This suggests that a greater diversity of non-spliced than spliced peptides may be generated as long polypeptides derived from ubiquitinated protein substrates are digested by the proteasome within cells.

Comparison of the relative abundance (based on LC-MS/MS intensity values) of the spliced and non-spliced peptides generated from all 19 precursor polypeptides following an *in vitro* 2h proteasomal digest revealed that the median abundance of the non-spliced peptide products was similar to that of the spliced peptide products (**Figure 1B**). Notably, the relative abundance

of both non-spliced and spliced peptides containing one of the terminal aas of the precursor polypeptide was significantly higher than that of peptide products generated from within the precursor **(Figure EV1E).** This is likely because non-spliced peptides containing the N- or C-termini of parental polypeptides can be generated by just a single excision event, and if higher concentrations of these peptide fragments are present they may then be preferentially incorporated into splicing events *in vitro*. Prior studies where the relative abundance of spliced and non-spliced peptides generated by other types of proteasomes was addressed have not yielded a consensus about this (Ebstein et al., 2016, Paes et al., 2019), but these utilised only very small numbers of polypeptide substrates, suggesting a need for more comprehensive future analyses to clarify whether there are differences in the efficiency with which spliced and non-spliced peptides are generated by different proteasomal isoforms both *in vitro* and *in cellulo*.

To limit potential bias introduced in determination of PCPS signatures by higher concentrations of N- or C-terminal-containing fragments that may preferentially participate in PCPS reactions, as well as bias introduced by the differential frequencies of aa residues comprising the start and end of each polypeptide precursor, both non-spliced and spliced peptides containing the precursor N- or C-terminal aa residues were excluded from subsequent analyses. Most of the splicing parameters described below remained consistent irrespective of whether analyses were performed with or without these, and any minor differences observed are noted.

***Cis*-spliced peptides follow a narrower length distribution than non-spliced peptides**

Comparison of the length distributions of *cis*-spliced and non-spliced peptides revealed that PCPS reactions generated peptides with a narrower length distribution than hydrolysis events generating non-spliced peptides (**Figure 1C**). For spliced peptides, 90% of PCPS reactions

generated products that were 13aa or smaller. In contrast, the 90th percentile length for non-spliced peptides was 21aa (**Figure 1C**). When both spliced and non-spliced peptides containing one of the terminal aas of the precursor were included, the respective 90th percentile length values were 16aa (for spliced) and 23aa (for non-spliced) peptides (**Figure EV2A**). To demonstrate that this distribution was not biased by the median length (22aa) of the precursor polypeptides studied, the length distributions of *cis*-spliced and non-spliced peptides derived from just the 5 longest precursors (27-44aa) were considered. Here we observed very similar trends, where 90% of spliced peptide products were 13aa or shorter, while the corresponding 90th percentile for non-spliced peptides was 24aa (**Figure EV2B**). When we compared the abundance of the non-spliced peptides greater than the 90th percentile of spliced peptides (i.e. >14aa) detected at the 2h and 20h timepoints we found no significant difference, suggesting that these peptides had reached saturation cleavage very early (**Figure EV2C**). Likewise, we did not observe a significant increase in the median abundance of the total non-spliced peptide repertoire generated by the CP from 2h to 20h (**Figure EV2D**). Overall, our results indicate that *cis*-PCPS reactions are typically biased towards the generation of a larger proportion of shorter peptides than canonical cleavage events giving rise to non-spliced peptides. This could be due to the nature of the substrate binding pockets and/or aa residues surrounding the catalytic subunits, which in turn govern the optimum lengths of splice partner fragments that can be accommodated in the correct orientation within the substrate binding pockets. Thus, we next sought to address whether constraints such as splice partner length or other physical and biochemical factors governing splicing reactions could be identified, and whether these differed between forward and reverse *cis*-PCPS reactions.

Forward and reverse *cis*-splicing events demonstrate unique preferences for N- and C-terminal splice partner lengths

Cis-PCPS reactions occur in either a forward or reverse sense as illustrated in **Figure 2A**. To characterise the physical and biochemical properties pertaining to each type of PCPS reaction, we separated the spliced peptides identified into forward and reverse spliced subsets. Reverse spliced peptides constituted a slightly higher proportion of all unique spliced peptides than forward spliced peptides (57% reverse versus 43% forward spliced) (**Figure 2B**), potentially because whereas generation of a forward spliced peptide from within a precursor peptide entails a minimum of 4 excision reactions, reverse spliced peptides can also be generated by 3 rather than 4 excisions (these constituted 18% of all the reverse spliced peptides considered here). Forward and reverse spliced peptides also displayed similar length distributions, indicating that both reactions predominantly generated peptides of 8-15aa in length (**Figure 2C**). Furthermore, we found no significant difference in the median abundance of forward and reverse spliced peptides (**Figure 2D**), indicating that both subsets of spliced peptides can be generated with similar efficiency by the CP *in vitro*.

In both types of PCPS reaction, N-terminal splice reactants of shorter lengths (2-5aa) were observed in a high proportion (67.8% in forward and 74.1% in reverse PCPS) of the final spliced peptide products, while longer splice reactants (6-12aa) were mainly found at the C-termini of spliced peptide products (**Figure 2E**). When we considered only the 5 longest precursors, we found that the fragment length distributions of forward and reverse spliced peptides were not biased by the selected polypeptide lengths, as only 3/70 (4.3%) of the splice partners in peptides generated from these substrates were >10aa long (**Figure EV3A**). This broadly reflected the overall distribution of splice partner fragments in **Figure 2E**, and further suggests that steric constraints within proteasomal substrate binding pockets may limit the length of splice partners that can be accommodated in PCPS reactions. Although it has previously been shown that the minimal length of a C-terminal splice reactant required to

perform nucleophilic attack on a stable acyl-enzyme intermediate was 3aa (Michaux et al., 2014), no upper bound for the length of splice partners has previously been described. Based on our results, this value may typically lie between 9-12aa. However, although unlikely at the 2h digestion time point analysed here, we cannot entirely preclude that post-PCPS processing (via proteasomal re-entry) may bias the fragment length distributions observed.

***Cis*-splicing events are biased towards splicing over shorter intervening distances within polypeptide substrates**

Although *cis*-PCPS reactions as a whole were found to occur preferentially over shorter distances (**Figure 2F**), differences were observed in the preferred intervening *cis*-splicing distances between forward and reverse PCPS reactions. Forward spliced peptides demonstrated a more marked preference for shorter *cis*-splicing distances, with the majority of splicing events (67.2%) involving excision of 1-2aa and a further 28.1% occurring over *cis*-splicing distances of 3-8aa. In contrast, only 30.5% of reverse PCPS events occurred over a distance of 1-2aa within the polypeptide, with the proportion of PCPS events spanning a 3-8aa gap increasing to 38.8%, and 11.8% of events occurring over still longer distances. Notably, we also observed that 18.8% of reverse spliced peptides occurred with no intervening distance between splice reactants (i.e. 0 aa) (**Figure 2F**).

For the 5 longest polypeptides (27-44aa) undergoing proteolysis, 5/10 (50%) of forward and 8/11 (72.7%) of reverse PCPS reactions occurred at a distance of >8aa (**Figure EV3B**). This was much greater than the overall proportion (~8%) observed across the entire polypeptide set. Thus, by using precursor polypeptides of moderate length, we may be underestimating the distances over which *cis*-PCPS can occur. Indeed, an experimentally validated FGF-5-derived spliced epitope in renal cell carcinoma was generated by a transpeptidation reaction spanning

40aa (Hanada et al., 2004). However, it was subsequently demonstrated that *cis*-PCPS reaction efficiencies displayed an inverse correlation with increasing intervening sequence length (Dalet et al., 2010), indicating that PCPS reactions are likely to occur with greater efficiency over shorter distances.

Proteasomal cleavage signatures in canonical hydrolysis and PCPS reactions demonstrate preferences for particular residues

Next, we analysed the overall frequency of cleavage after each of the 20 aa residues across the 19 synthetic polypeptides following both canonical hydrolysis and PCPS reactions. Global cleavage frequencies were first obtained by incorporating all cleavage sites for each set of peptides - i.e. 2 sites for non-spliced, 3 or 4 sites for reverse spliced (depending on the intervening sequence length between splice partners) and 4 sites for forward spliced peptides. These cleavage sites are denoted by red bars in in **Figure 3A**. Aa frequencies were then normalised relative to the overall frequency of that aa within all 19 precursor polypeptides in order to ascertain whether enrichment of cleavage after particular residues was evident (**Figure 3B**). For non-spliced peptides, cleavage after certain aas (A, D, E, F, H, L, M, N, V, and W) was higher than would be expected given their overall frequency within the substrate precursors. Although we did not observe pronounced tryptic-like activity (cleavage after basic residues) our data are largely in accordance with results from a recent analysis of non-spliced proteasomal cleavage products derived from a set of 228 rationally-designed synthetic 14-mer peptides, where preferential cleavage after A, F, G, H, L, N, R, S and T residues was observed (Winter et al., 2017). Additionally, while it is established that the catalytically active subunits (β_1, β_2 and β_5) of the CP predominantly initiate cleavage after acidic, hydrophobic and basic residues, respectively (Vigneron et al., 2017), cleavages following branched chain or small

neutral aas also predominate (Orlowski, Cardozo et al., 1993), and were likewise observed in our study.

However, for spliced peptides, cleavage preferences were more restricted, with C, F, G and L residues being over-represented relative to their respective frequencies in the substrate precursors. PCPS is characteristically thought to involve nucleophilic attack of a stable acyl-enzyme intermediate by the free amine group of a non-adjacent aa sequence (as although it has also been suggested that certain spliced peptides may be produced without the formation of a semi-stable *O*-acyl-enzyme intermediate via a condensation reaction (Ebstein et al., 2016), generation of spliced peptides from a contiguous precursor polypeptide/protein via this mechanism has yet to be demonstrated). Therefore, we sought to identify aa residues predominantly participating as either acyl-enzyme intermediates, denoted in **Figure 3A** by X_1 or X_2 (for forward and reverse spliced peptides respectively), or nucleophilic free amines (denoted as Z_1 or Z_2 for forward and reverse spliced peptides). Here, we observed differences between each reaction (**Figure 3C-D**). During reverse PCPS, C and S predominantly acted as anchor residues for stable acyl-enzyme intermediates, while E, F, G, H, I and Y were typically favoured in forward splicing reactions (**Figure 3C**). Similarly, residues involved in nucleophilic attack of the N-terminal fragments showed differences, with K and T specifically favoured in reverse splicing reactions and C, E, H and L in forward PCPS (**Figure 3D**). Residues D, F and G were favoured in both reverse and forward PCPS reactions. Although the global sequence context of individual precursor polypeptides may influence cleavage specificity, our data suggest that in addition to preferred length constraints for N- and C-terminal splice partners, certain residues may preferentially function as reactive intermediates or nucleophilic amines in forward and reverse PCPS reactions, which may facilitate optimal positioning or stability within the proteasomal chamber to enable splicing. In addition to

cleavage site residues, because proximal aas within splice partners would also dictate the affinity of each fragment for particular substrate binding pockets, we sought to determine preferred splice site signatures that may elaborate on the sequence-dependent accessibility of splice partners in forward and reverse PCPS reactions.

Unique splice site signatures are observed in forward and reverse splicing reactions

Aa pairing combinations ± 3 residues from the fusion junction were inspected to deduce the biochemical characteristics of N- and C-terminal splice partners preferentially accommodated within substrate binding pockets during PCPS (**Figure 4A**). Here, distinct patterns were observed for forward and reverse PCPS reactions (**Figures 4B** and **C**). In forward PCPS reactions, immediately proximal to the fusion junction at [P1]-[P1'] we observed preferred [P1]-[P1'] pairings (e.g. [G/H/I]–[L], [F]-[E], [G]-[I] ad [I/K]-[S]) (**Figure 4B**). Conversely, in reverse PCPS reactions, we observed hotspots at [C/I/L/S]-[K], [N]-[G] and [S]-[E/H/K/Q] at the [P1]-[P1'] position that were not evident in forward PCPS reactions (**Figure 4C**). Notably, nucleophilic attack was frequently mediated by K residues of the C-terminal splice partner in reverse PCPS events, which was not evident in forward PCPS, where L was more predominant. While most acyl-enzyme intermediate and nucleophilic free amine residues participating in both forward and reverse PCPS reactions at the [P1]-[P1'] position corresponded to the dominant cleavage events (depicted in **Figure 3B**), residues from 'minor' cleavage sites were also over-represented at [P1]-[P1'], indicating that PCPS is not entirely dictated by the efficiency of cleavage events producing the constitutive splice reactants, but is also sequence-dependent. Indeed, further differences between forward and reverse PCPS reactions were observed at [P2]-[P2'] and [P3]-[P3'] positions (**Figures 4B** and **C**).

Importantly, inspection of splice site signatures defined following *in vitro* proteasomal digests in our study revealed that these differed from those of MHC-I bound spliced peptides identified from MS-based immunopeptidome profiling datasets (Faridi et al., 2018, Liepe et al., 2016). Previously proposed aa splicing preferences determined by inspection of immunopeptidomic datasets could be biased by selective pressures imposed on proteasome-generated peptide pools during subsequent steps of antigen processing/presentation, e.g. TAP-mediated peptide translocation into the ER, MHC-I binding preferences and ER-mediated peptide trimming (Neefjes, Jongsma et al., 2011). The repertoire of spliced peptides detected in MHC-I-bound peptide repertoires may also be restricted by the peptide length preferences of class I alleles. For instance, in our study, ~30% of *cis*-spliced peptides generated following *in vitro* proteasomal polypeptide digestion were >12aa in length, and would thus be highly under-represented in MHC-I bound peptide datasets.

Despite over a decade of research, the properties governing PCPS are poorly understood. By employing a tailored MS-based *de novo* sequencing approach to identify peptides generated by proteasomal splicing *in vitro*, our study has provided novel insight into both the steric length constraints and biochemical preferences of PCPS mediated by the CP. Future analysis of *in vitro* peptide splicing by other proteasome types, and comparative analysis of the MHC-I bound spliced peptides generated in cells expressing different proteasomal subunits, and in immunopeptidomes generated by direct versus cross-presentation would be of interest. Importantly, the delineation of cleavage and splice site signatures reported here will facilitate a more informed investigation of the mechanisms of PCPS, and the development of algorithms to predict spliced peptide generation from protein sequences. This will in turn expedite future work to define spliced peptides derived from pathogen and tumour antigens that could be targeted for prophylaxis or therapy. In addition, a more thorough understanding of the

repertoire of spliced peptides produced within cells could uncover roles of these peptides in aspects of cell biology beyond antigen presentation.

Materials and Methods

Synthetic peptides

Eight peptides derived from the sequences of NL4.3 or IIIB laboratory-adapted HIV-1 strains or patient viruses, eight "self" peptides from different human protein sequences from the SwissProt Homo sapiens reference proteome and three polypeptides originating from a publicly available sORF database (Olexiouk et al., 2016) (**Table EV1**) were employed as substrates for proteasomal digestion. Peptides were synthesised by Genscript (USA) using Fmoc solid phase chemistry.

In vitro proteasomal digests

For constitutive proteasomal digests, 5µg of synthetic polypeptide were incubated with 500ng of the 20S constitutive proteasome (CP) (Enzo Life Sciences) in 20mM HEPES (pH 7.8), 5mM magnesium chloride, 2mM DTT for 2h or 20h at 37°C. Control incubations (lacking the CP) were performed for each peptide in parallel. Following incubation, reactions were terminated by addition of 5µL of acetic acid, and digest material was bound to a C18 ZipTip column (Merck), eluted in 30% acetonitrile (ACN) in 5% trifluoroacetic acid (TFA), dried down and resuspended in 20µl LC-MS/MS loading buffer (1% ACN, 0.1% TFA in water).

LC-MS/MS Analysis

1µl of each digest sample was injected onto a 3µm particle size 0.075mm x 150mm PepMap C18 trap column followed by loading onto a 2µm particle size, 75µm x 50cm analytical column on the Ultimate 3000 nUPLC system (Thermo Scientific). A gradient was applied for 30min

from 8-50% buffer B (0.1% TFA in 100% ACN) in buffer A (1% ACN, 0.1% TFA in water). Peptides were introduced to the Fusion Lumos mass spectrometer (Thermo Scientific) using an Easy-Spray source at 2000V and 40ºC. The ion transfer tube temperature was set to 305ºC. Full MS spectra were recorded from 300-1500 m/z in the Orbitrap at 120,000 resolution with an automatic gain control (AGC) target of 400,000. Precursor selection was performed using Top Speed mode at a cycle time of 2s. Peptide ions were isolated using a width of 1.2amu, followed by trapping at a maximal injection time of 120ms, setting an AGC target of 300,000. Higher-energy collisional dissociation (HCD) fragmentation was induced at an energy setting of 28 for peptides with a charge state of 2-4, while singly charged peptides were fragmented at an energy setting of 32 at lower priority. Fragments were analysed in the Orbitrap at 30,000 resolution.

Discovery workflow for identification of non-spliced and spliced peptides

PEAKS v8.0 (Bioinformatic Solutions) software was employed for the analysis of all LC-MS/MS datasets (.raw files). Matching against precursor polypeptide sequences was conducted without enzyme specification, using a mass tolerance precursor setting of 5ppm for peptides and 0.03Da for fragment ions. PEAKS *de novo* assisted sequencing was implemented for the assignment of non-spliced peptides derived from individual polypeptide sequences following proteasomal digest of each of the 19 precursor substrates. A false discovery rate of 5% was set using a parallel decoy database search. Following matching of spectra to respective polypeptide sequences, MS/MS spectra assigned as post-translationally modified or originating from the 20S CP were removed from the analysis, as were peptides assigned a PEAKS -10lgP score of <20, yielding a final list of non-spliced peptide sequences.

For identification of spliced peptides, a tailored bioinformatics workflow was employed similar to that previously described (Paes et al., 2019). *De novo* sequences originating from peptide spectra with average local confidence (ALC) scores of ≥50% that did not match contiguous sequences within the polypeptide precursors were considered. These were termed '*de novo* unmatched peptides' (DNUPs), and only the top 5 scoring sequence interpretations for each scan were included in the analysis. As the mass:charge ratios of leucine (L) and isoleucine (I) are identical, LC-MS/MS cannot differentiate between L and I residues within peptides. Hence, PEAKS *de novo* sequencing reports all L or I residues as Ls. Therefore, for DNUPs containing a total of 'n' leucine residues, all permutations (2^n) of L/I variants were computed prior to *in silico* splicing – e.g. for *de novo* sequence DLLGCFKPFLC originating from polypeptide precursor KTFKPFICSLDLIGCPM, 2^3 combinations of the original *de novo* sequence would be computed (DLLGCFKPFLC, DLIGCFKPFLC, DILGCFKPFLC, DLLGCFKPFIC, DLIGCFKPFIC, DILGCFKPFIC, DIIGCFKPFLC and DIIGCFKPFIC) and each sequence input to the splicing algorithm. If a match to a contiguous polypeptide sequence within the SwissProt proteome resulted following L/I permutations, the whole LC-MS/MS spectrum/scan was removed from the analysis.

In silico splicing of the resulting DNUP sequence interpretations in each scan was then implemented. Briefly, each DNUP of length 'n' amino acids was split into two fragments from (n-1)aa to 1aa. The $(n-1)^{th}$ fragment was first scanned for a contiguous match across the polypeptide, and when found, its corresponding splice partner fragment was scanned for a contiguous match within the remainder of the polypeptide sequence. *Trans*-spliced peptides were omitted from the analysis. Contiguous and spliced peptides identified in control experiments involving incubation of polypeptides without the 20S CP were removed prior to

further data analysis, to exclude degradation products that may not arise directly from proteasome-induced hydrolysis.

Data analysis

Data analysis, plotting and determination of statistical significance was implemented using Graphpad Prism version 8.0. Non-spliced peptide identifications were retrieved from PEAKS PTM searches and corresponding LC-MS/MS intensity values were extracted to plot abundance comparisons. For extraction of abundance values for *de novo* sequenced *cis*-spliced peptides, matching of scan number and m/z value for a given identified *cis*-spliced peptide (obtained from the all *de novo* candidates PEAKS file) was implemented to extract corresponding intensity values from the *de novo* output. Peptide length and fragment length distributions were plotted using Prism, and peptide splicing heat maps were generated using gnuplot version 5.2.4.

Statistical analysis

Unless otherwise stated, two-tailed unpaired Mann-Whitney t-tests were carried out to assess the statistical significance of differences between groups using GraphPad Prism v8.0. Differences were considered statistically significant at a p-value of <0.05.

Acknowledgements

The authors would like to thank Professor Benoit van den Eynde for helpful discussions about the study. This work was funded by Medical Research Council (MRC) programme grant MR/K012037 (PB), and by grants from the National Institutes of Health (R01 AI 118549 and the Duke Centre for HIV/AIDS Vaccine Immunology-Immunogen Discovery (UM1 AI 100645)) (PB). Mass spectrometry samples were acquired in the Target Discovery Institute

Mass Spectrometry laboratory led by Benedikt Kessler. PB and NT are Jenner Institute Investigators.

Author contributions

WP, GL, TP, NT and PB conceived the study. WP designed the study, conducted experiments, analysed data, prepared figures and drafted the manuscript. AN and NT acquired MS data. GL, TP, NT and PB contributed to data analysis and manuscript editing. All authors read and approved the final version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

Berkers CR, de Jong A, Schuurman KG, Linnemann C, Meiring HD, Janssen L, Neefjes JJ, Schumacher TN, Rodenko B, Ovaa H (2015) Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. *J Immunol* 195: 4085-95

Dalet A, Robbins PF, Stroobant V, Vigneron N, Li YF, El-Gamil M, Hanada K, Yang JC, Rosenberg SA, Van den Eynde BJ (2011) An antigenic peptide produced by reverse splicing and double asparagine deamidation. *Proc Natl Acad Sci U S A* 108: E323-31

Dalet A, Vigneron N, Stroobant V, Hanada K, Van den Eynde BJ (2010) 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: 3016-24

Delong T, Wiles TA, Baker RL, Bradley B, Barbour G, Reisdorph R, Armstrong M, Powell RL, Reisdorph N, Kumar N, Elso CM, DeNicola M, Bottino R, Powers AC, Harlan DM, Kent SC, Mannering SI, Haskins K (2016) Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion. *Science* 351: 711-4

Ebstein F, Textoris-Taube K, Keller C, Golnik R, Vigneron N, Van den Eynde BJ, Schuler-Thurner B, Schadendorf D, Lorenz FK, Uckert W, Urban S, Lehmann A, Albrecht-Koepke N, Janek K, Henklein P, Niewienda A, Kloetzel PM, Mishto M (2016) Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. *Sci Rep* 6: 24032

Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, Ayala R, Song J, Gearing LJ, Hertzog PJ, Ternette N, Rossjohn J, Croft NP, Purcell AW (2018) A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Sci Immunol* 3

Hanada K, Yewdell JW, Yang JC (2004) Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* 427: 252-6
Kuckelkorn U, Stubler S, Textoris-Taube K, Kilian C, Niewienda A, Henklein P, Janek K, Stumpf MPH, Mishto M, Liepe J (2019) Proteolytic dynamics of human 20S thymoproteasome. *J Biol Chem* 294: 7740-7754
Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A, Kloetzel PM, Stumpf MP, Heck AJ, Mishto M (2016) A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* 354: 354-358
Michaux A, Larrieu P, Stroobant V, Fonteneau JF, Jotereau F, Van den Eynde BJ, Moreau-Aubry A, Vigneron N (2014) 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: 1962-71
Mishto M, Liepe J, Textoris-Taube K, Keller C, Henklein P, Weberruss M, Dahlmann B, Enenkel C, Voigt A, Kuckelkorn U, Stumpf MP, Kloetzel PM (2014) Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation. *Eur J Immunol* 44: 3508-21
Murata S, Takahama Y, Kasahara M, Tanaka K (2018) The immunoproteasome and thymoproteasome: functions, evolution and human disease. *Nat Immunol* 19: 923-931
Mylonas R, Beer I, Iseli C, Chong C, Pak HS, Gfeller D, Coukos G, Xenarios I, Muller M, Bassani-Sternberg M (2018) Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome. *Mol Cell Proteomics* 17: 2347-2357
Neefjes J, Jongsma ML, Paul P, Bakke O (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 11: 823-36
Olexiouk V, Crappe J, Verbruggen S, Verhegen K, Martens L, Menschaert G (2016) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* 44: D324-9
Orlowski M, Cardozo C, Michaud C (1993) Evidence for the presence of five distinct proteolytic components in the pituitary multicatalytic proteinase complex. Properties of two components cleaving bonds on the carboxyl side of branched chain and small neutral amino acids. *Biochemistry* 32: 1563-72
Paes W, Leonov G, Partridge T, Chikata T, Murakoshi H, Frangou A, Brackenridge S, Nicastri A, Smith AG, Learn GH, Li Y, Parker R, Oka S, Pellegrino P, Williams I, Haynes BF, McMichael AJ, Shaw GM, Hahn BH, Takiguchi M et al. (2019) Contribution of proteasome-catalyzed peptide cis-splicing to viral targeting by CD8(+) T cells in HIV-1 infection. *Proc Natl Acad Sci U S A* 116: 24748-24759
Platteel AC, Mishto M, Textoris-Taube K, Keller C, Liepe J, Busch DH, Kloetzel PM, Sijts AJ (2016) CD8(+) T cells of Listeria monocytogenes-infected mice recognize both linear and spliced proteasome products. *Eur J Immunol* 46: 1109-18
Platteel ACM, Liepe J, Textoris-Taube K, Keller C, Henklein P, Schalkwijk HH, Cardoso R, Kloetzel PM, Mishto M, Sijts A (2017) Multi-level Strategy for Identifying Proteasome-Catalyzed Spliced Epitopes Targeted by CD8(+) T Cells during Bacterial Infection. *Cell Rep* 20: 1242-1253
Saeki Y, Tanaka K (2012) Assembly and function of the proteasome. *Methods Mol Biol* 832: 315-37
Vigneron N, Ferrari V, Stroobant V, Abi Habib J, Van den Eynde BJ (2017) Peptide splicing by the proteasome. *J Biol Chem* 292: 21170-21179
Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S, van der Bruggen P, Boon T, Van den Eynde BJ (2004) An antigenic peptide produced by peptide splicing in the proteasome. *Science* 304: 587-90

Warren EH, Vigneron NJ, Gavin MA, Coulie PG, Stroobant V, Dalet A, Tykodi SS, Xuereb SM, Mito JK, Riddell SR, Van den Eynde BJ (2006) An antigen produced by splicing of noncontiguous peptides in the reverse order. *Science* 313: 1444-7
Winter MB, La Greca F, Arastu-Kapur S, Caiazza F, Cimermancic P, Buchholz TJ, Anderl JL, Ravalin M, Bohn MF, Sali A, O'Donoghue AJ, Craik CS (2017) Immunoproteasome functions explained by divergence in cleavage specificity and regulation. *Elife* 6

Figure Legends

Figure 1. Diversity, abundance and peptide length distribution of proteasome-derived spliced and non-spliced peptides

A. Proportion of unique spliced and non-spliced peptides following a 2h *in vitro* digestion of 19 self- and HIV-1-derived polypeptides (**Table EV1**) by the constitutive proteasome. Proportions of spliced and non-spliced peptides within all unique peptides (n=1,344), unique peptides originating from only the 5 longest polypeptide substrates (n=501) and unique peptides originating from only the 5 shortest polypeptide precursors (n=177) are shown.

B. Violin plots showing abundance of all unique spliced and non-spliced peptides as measured by LC-MS/MS intensity values. Median and quartile abundance values are indicated.

C. Length distributions of unique spliced (n=149) and non-spliced (n=487) peptides.

In B and C, $P<0.05$ was used as the threshold for significance following a non-parametric unpaired Mann-Whitney t-test. ** $P<0.01$.

Figure 2. Physical properties of forward and reverse PCPS reactions

A. Schematic depicting forward and reverse PCPS and indicating the N- and C-terminal fusion partners.

B. Relative proportions of forward and reverse spliced peptides among all unique spliced peptides (n=149) identified following 2h *in vitro* digestion of 19 self- and HIV-1-derived polypeptide sequences by the constitutive proteasome.

C. Length distributions of forward (n=64) and reverse (n=85) spliced peptides.

D. Relative abundance of forward and reverse spliced peptides. Median and quartile abundance values are indicated. A non-parametric unpaired Mann-Whitney t-test was used to determine whether abundance values differed between groups, with $P<0.05$ as the significance threshold.

E. Fragment length distributions of N-terminal and C-terminal splice partners involved in forward and reverse PCPS reactions.

F. *Cis*-splicing distances in forward and reverse PCPS reactions.

Figure 3. Proteasomal cleavage preferences differ between canonical hydrolysis and PCPS reactions

A. Schematic depicting cleavage sites employed to generate non-spliced and forward and reverse spliced peptides from precursor polypeptides. X_1 and Z_1 denote residues at cleavage sites acting as O-acyl enzyme intermediates and nucleophilic free amine groups (respectively) in forward PCPS reactions, while X_2 and Z_2 denote the corresponding residues in reverse PCPS reactions.

B. Comparison of global amino acid cleavage site usage for spliced and non-spliced peptides. Amino acid frequencies are normalised relative to the overall frequency of that amino acid within all 19 precursor polypeptides.

C. Normalised frequencies of amino acid residues acting as acyl-enzyme intermediates in forward and reverse PCPS reactions.

D. Normalised frequencies of amino acid residues acting as nucleophilic free amines in forward and reverse PCPS reactions.

In B-D, amino acids are defined using single letter code.

Figure 4. Forward and reverse PCPS reactions display unique splice site signatures

A. Schematic depicting forward and reverse PCPS. Red arrows indicate fusion sites. Amino acid pairings proximal to the fusion site are labelled P1/P1', P2/P2' and P3/P3'.

B. Forward PCPS splice site signature.

C. Reverse PCPS splice site signature.

In B and C, heat maps show frequencies of amino acid residue pairings at ± 1, 2 or 3 aa adjacent to the fusion site. Scale represents the frequency of each amino acid pairing as a percentage of the overall pairings, and amino acids are defined using single letter code.

List of Expanded View Tables

Table EV1 – Synthetic polypeptides subjected to *in vitro* proteasomal digestion

Table EV2 – Amino acid frequencies within synthetic polypeptides and combined HIV-1 and UniProt human proteomes

Expanded View Figure Legends

Figure EV1. Diversity and abundance of proteasome-derived spliced and non-spliced peptides

A. Comparison of the relative abundance at the 2h and 20h timepoints of precursor polypeptides (**Table EV1**) that were observed at both timepoints.

B. Area-proportional Venn diagram illustrating the diversity of constitutive proteasome-derived non-spliced peptides following *in vitro* digestion of precursor polypeptide substrates for 2h or 20h.

C. Proportion of unique spliced and non-spliced peptides following 20 h *in vitro* digestion of 19 self- and HIV-1-derived polypeptide sequences (**Table EV1**) by the constitutive proteasome. Proportions of the unique peptides generated from all 19 polypeptide substrates (n=1,614), unique peptides originating from only the 5 longest polypeptide substrates (n=585)

and unique peptides originating from only the 5 shortest polypeptide precursors (n=194) are shown.

D. Number of unique spliced and non-spliced peptide products identified following 2h *in vitro* digestion of 19 precursor polypeptides (**Table EV1**) by the constitutive proteasome. Numbers of all unique spliced and non-spliced peptides, those originating from within the polypeptide substrate and not containing the terminal amino acid (-T) and those containing terminal amino acid(s) of the precursor substrate (+T) or are shown.

E. Violin plots showing abundance of the spliced and non-spliced peptide products containing terminal amino acids of precursor substrates (+T) and those originating from within the polypeptide substrate (-T), as measured by LC-MS/MS intensity values. Median and quartile abundance values are indicated. $P<0.05$ was used as the threshold for significance following a non-parametric unpaired Mann-Whitney t-test. $*P\leq0.05$.

Figure EV2. Length distribution and abundance of peptide products generated from 2h or 20h *in vitro* proteasomal digests

A. Length distributions of unique spliced (n=665) and non-spliced (n=697) peptide products generated from all 19 precursor polypeptide substrates after a 20h *in vitro* digest.

B. Length distributions of unique spliced (n=72) and non-spliced (n=429) peptides generated from the 5 longest precursor polypeptide substrates (27-44 aa in length) after a 2 h *in vitro* digest.

C. Comparison of the relative abundance of non-spliced peptides of length greater than the 90th percentile of spliced peptides (i.e. ≥14 aa) generated at 2h or 20h (as measured by LC-MS/MS intensity values). Median and quartile abundance values are shown.

D. Comparison of the relative abundance of non-spliced peptides generated at 2h or 20h (as measured by LC-MS/MS intensity values). Median and quartile abundance values are shown.

In C and D, $P<0.05$ was used as the threshold for significance following a non-parametric unpaired Mann-Whitney t-test.

Figure EV3. Fragment length and *cis*-splicing distributions for in forward and reverse PCPS reactions occurring during digestion of the 5 longest polypeptide precursors

A. Fragment length distributions of N-terminal and C-terminal splice partners involved in forward and reverse PCPS reactions generating spliced peptides from the 5 longest polypeptide precursors (27-44aa in length).

B. *Cis*-splicing distances in forward and reverse PCPS reactions generating spliced peptides from the 5 longest polypeptide precursors (27-44aa in length).

Figure 1

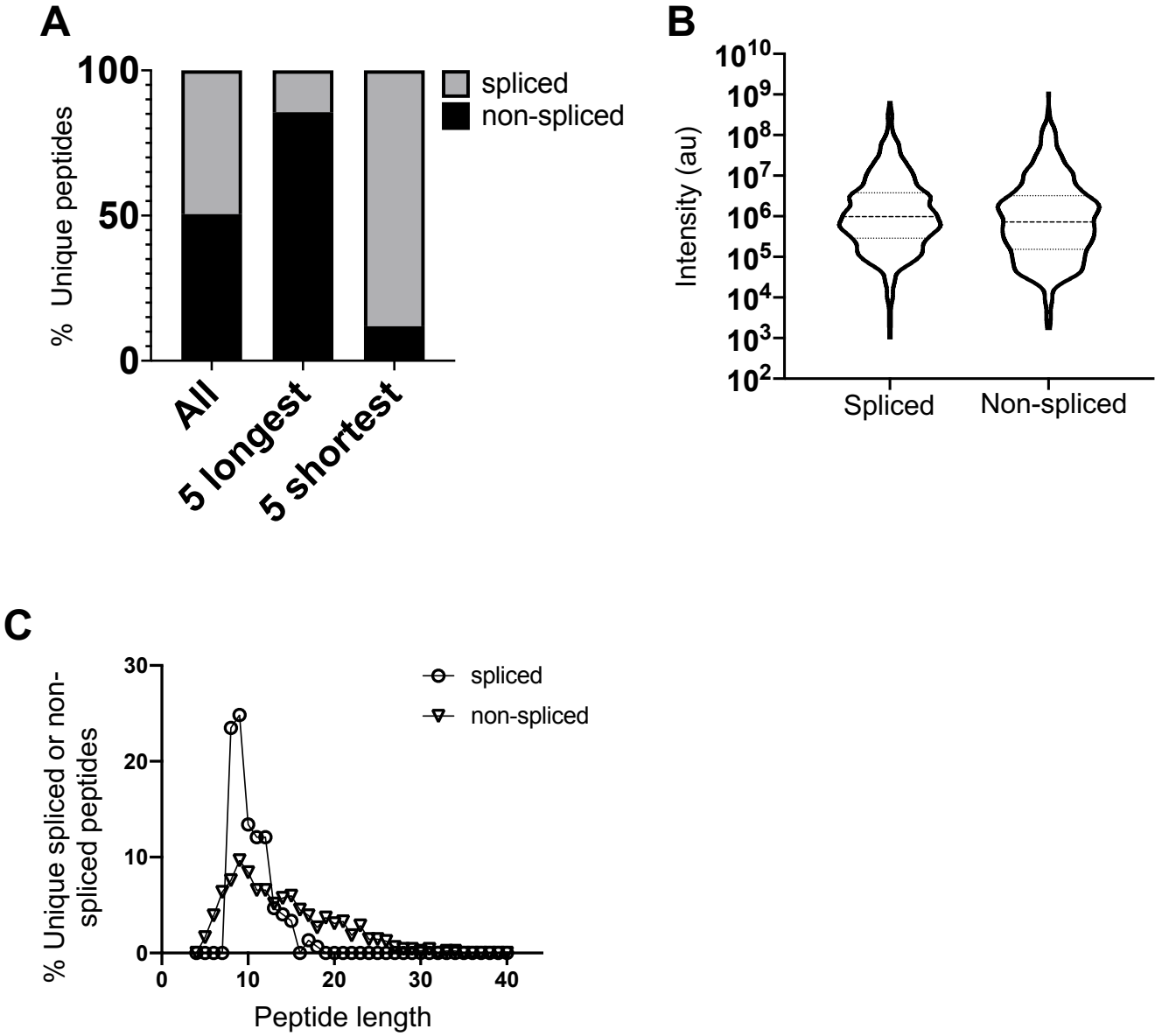

A
100
50
0
% Unique peptides
spliced
non-spliced
All
5 longest
5 shortest
B
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
Intensity (au)
Spliced
Non-spliced
C
30
20
10
0
% Unique spliced or non-spliced peptides
spliced
non-spliced
0
10
20
30
40
Peptide length

Figure 2

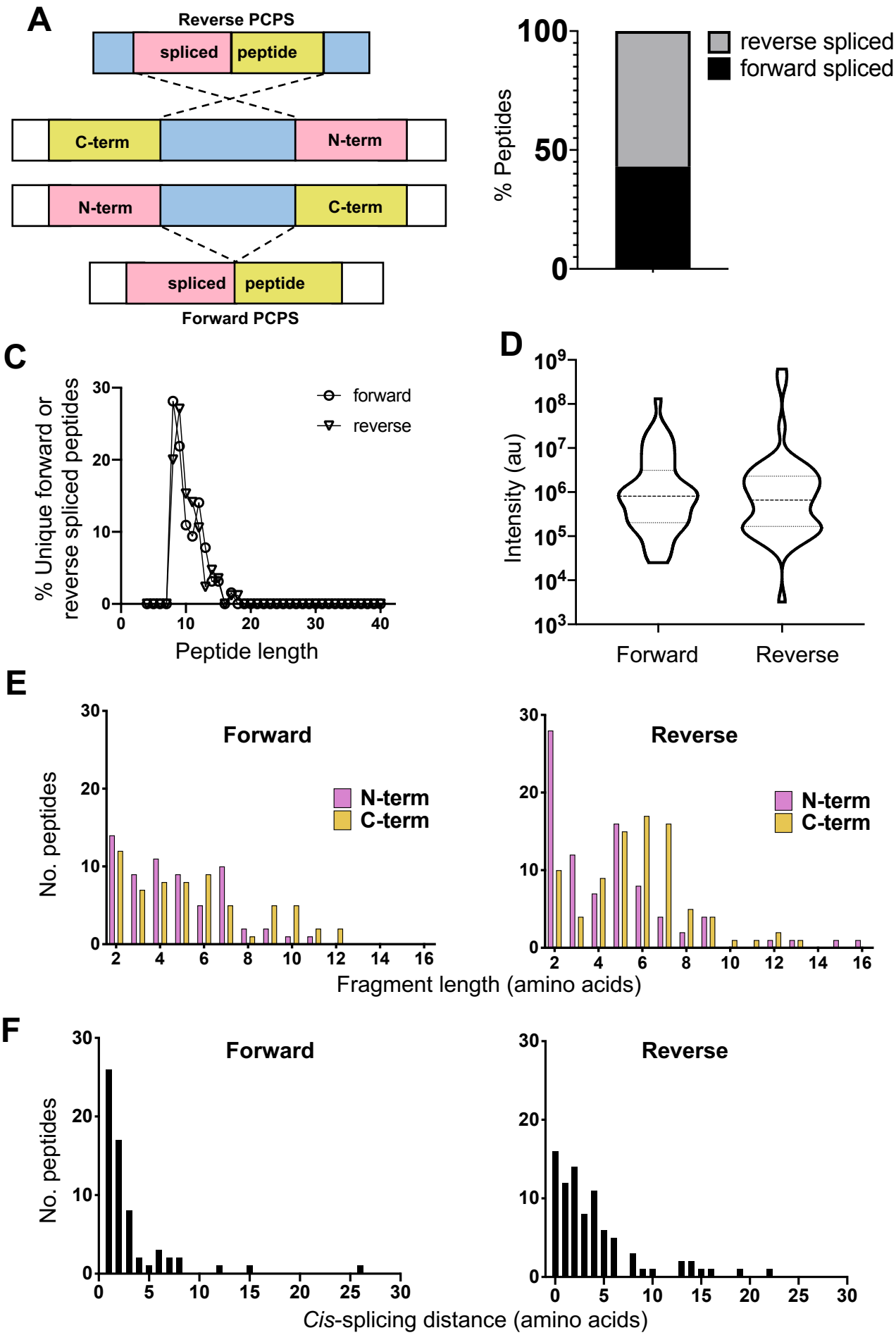

A
Reverse PCPS
spliced
peptide
C-term
N-term
N-term
C-term
spliced
peptide
Forward PCPS
B
100
50
0
% Peptides
reverse spliced
forward spliced
C
% Unique forward or reverse spliced peptides
30
20
10
0
0
10
20
30
40
Peptide length
forward
reverse
D
Intensity (au)
10^9
10^8
10^7
10^6
10^5
10^4
10^3
Forward
Reverse
E
Forward
Reverse
No. peptides
N-term
C-term
2
4
6
8
10
12
14
16
Fragment length (amino acids)
F
Forward
Reverse
No. peptides
0
5
10
15
20
25
30
Cis-splicing distance (amino acids)

Figure 3

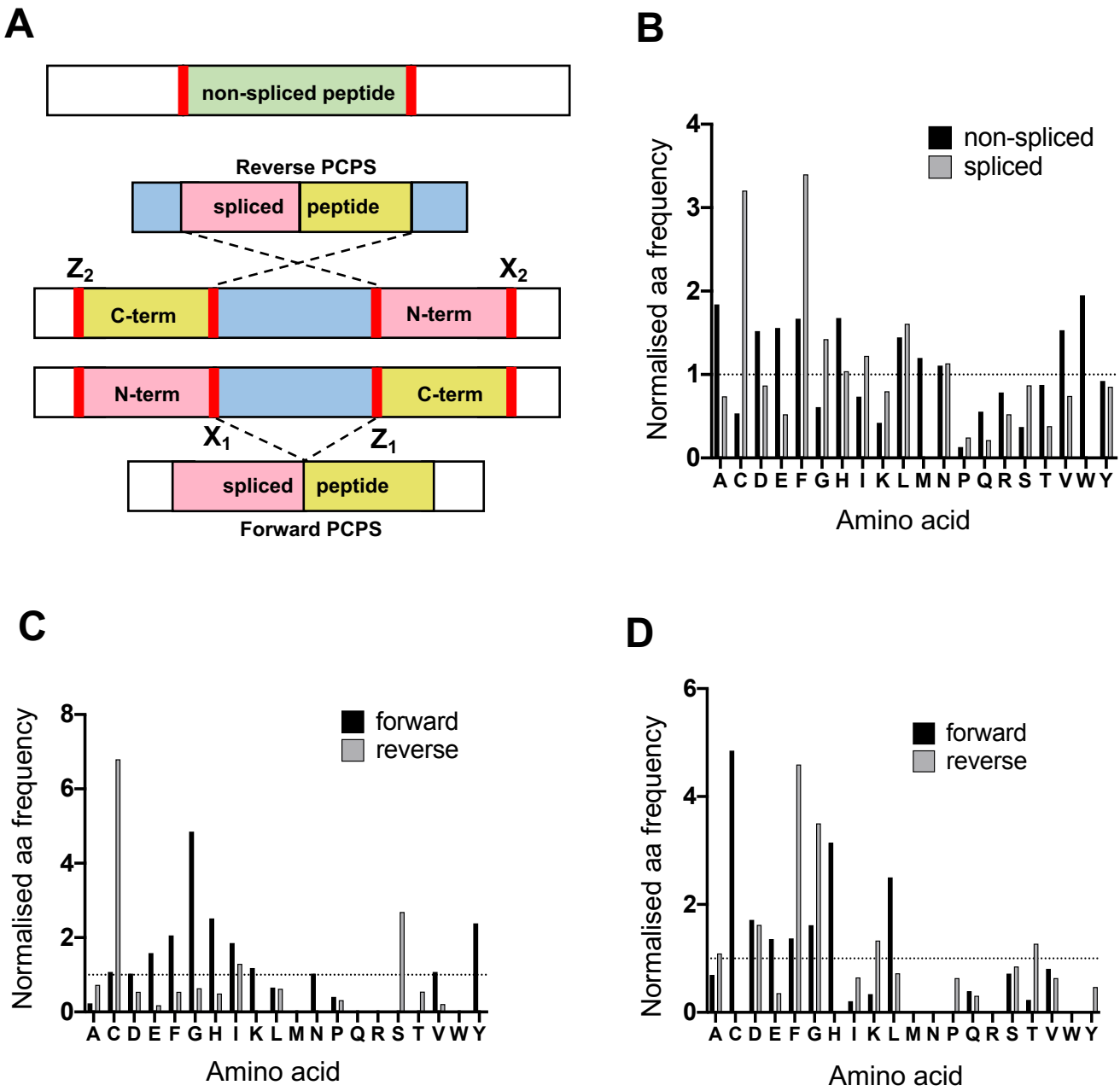
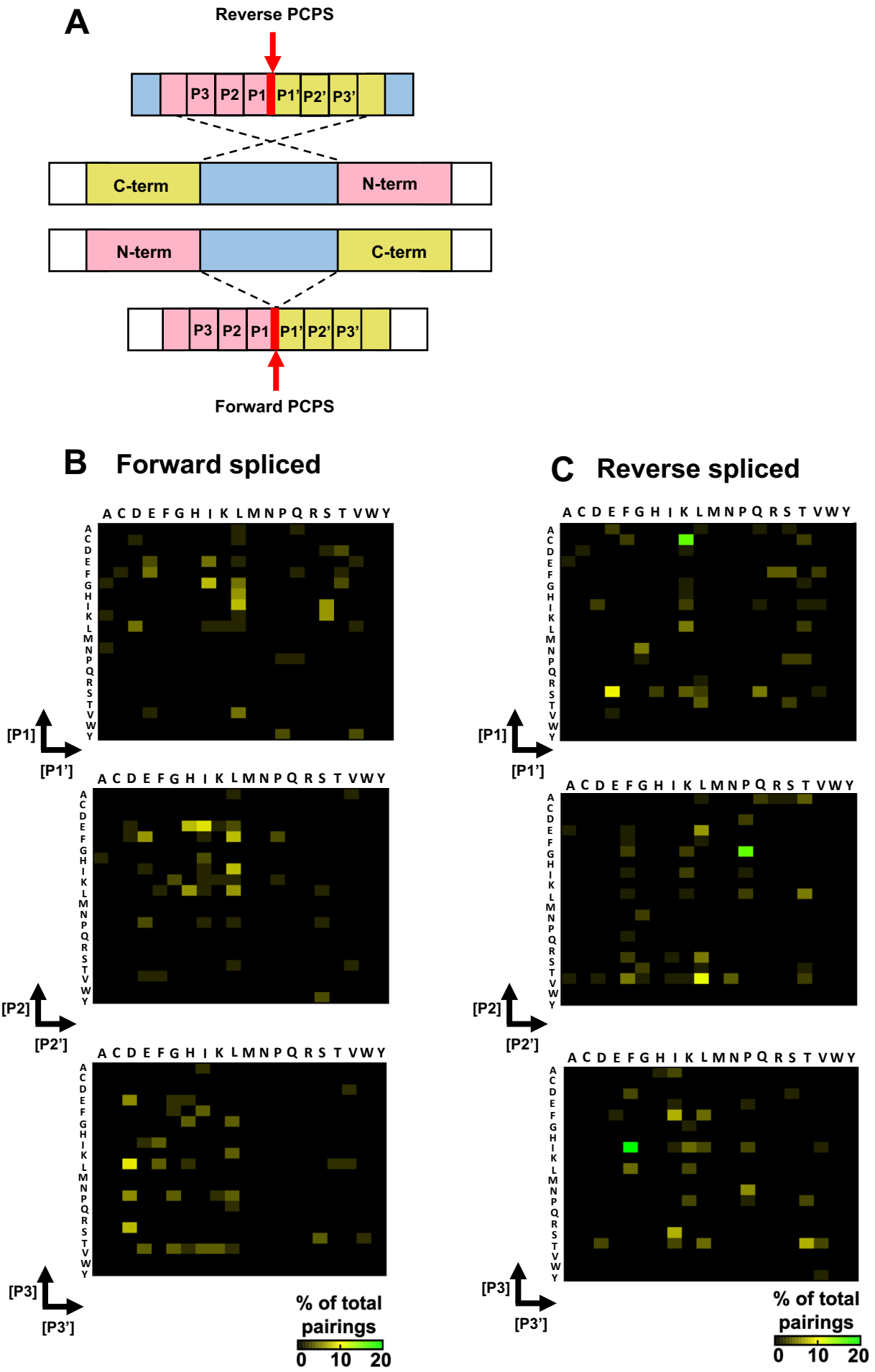

A
non-spliced peptide
Reverse PCPS
spliced
peptide
Z2
X2
C-term
N-term
N-term
C-term
X1
Z1
spliced
peptide
Forward PCPS
B
non-spliced
spliced
Normalised aa frequency
Amino acid
A C D E F G H I K L M N P Q R S T V W Y
C
forward
reverse
Normalised aa frequency
Amino acid
A C D E F G H I K L M N P Q R S T V W Y
D
forward
reverse
Normalised aa frequency
Amino acid
A C D E F G H I K L M N P Q R S T V W Y

Figure 4


A
Reverse PCPS
P3
P2
P1
P1'
P2'
P3'
C-term
N-term
N-term
C-term
P3
P2
P1
P1'
P2'
P3'
Forward PCPS
B Forward spliced
C Reverse spliced
A C D E F G H I K L M N P Q R S T V W Y
[P1]
[P1']
[P2]
[P2']
[P3]
[P3']
% of total pairings
0 10 20