The role of proteasome generated spliced peptides in the adaptive immune response

Dissertation

for the award of the degree
Doctor of Philosophy
Division of Mathematics and Natural Science
of the Georg-August-Universität Göttingen

within the doctoral program International Max Planck Research School for Genome Science
of the Georg-August University School of Science (GAUSS)

Submitted by
Artem Mansurkhodzhaev

from Moscow, Russia

Göttingen, 2022

Thesis Committee

Dr. Juliane Liepe	Quantitative and Systems Biology Max Planck Institute for Multidisciplinary Sciences Göttingen
Prof. Dr. Henning Urlaub	Bioanalytical Mass Spectrometry Max Planck Institute for Multidisciplinary Sciences Göttingen Institute for Clinical Chemistry University Medical Center Göttingen
Prof. Dr. Patrick Cramer	Department of Molecular Biology Max Planck Institute for Multidisciplinary Sciences Göttingen

Members of the Examination Board

Reviewer:	
Dr. Juliane Liepe	Quantitative and Systems Biology Max Planck Institute for Multidisciplinary Sciences Göttingen
2nd Reviewer:	
Prof. Dr. Henning Urlaub	Bioanalytical Mass Spectrometry Max Planck Institute for Multidisciplinary Sciences Göttingen Institute for Clinical Chemistry University Medical Center Göttingen

Further members of the Examination Board

Prof. Dr. Patrick Cramer	Department of Molecular Biology Max Planck Institute for Multidisciplinary Sciences Göttingen
Dr. Sonja Lorenz	Ubiquitin signaling specificity Max Planck Institute for Multidisciplinary Sciences Göttingen
Dr. Johannes Soeding	Computational Biology Max Planck Institute for Multidisciplinary Sciences Göttingen
Dr. Alexander Stein	Membrane Protein Biochemistry Max Planck Institute for Multidisciplinary Sciences Göttingen

Date of the oral examination: April 20th, 2022

Declaration of Authorship

I, Artem Mansurkhodzhaev, declare that this thesis titled, "The role of proteasome generated spliced peptides in the adaptive immune response" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

Acknowledgements

I would like to express my gratitude to Dr. Juliane Liepe for her guidance and supervision as well as her continued support over the course of my research. I thank everyone in Quantitive and Systems Biology group for the great working environment. Thank you to Dr. Michele Mishto's group from King's College of London for generating the samples for the analysis and the productive project discussions. I would also like to thank Prof. Dr. Henning Urlaub, Dr. Sabine König, Dr. Ralf Pflanz, Monika Raabe, Uwe Plessmann and Dr. Christof Lenz for providing access to the Mass Spectrometry facilities at MPI-NAT and all the help with sample handling and measurements. I would like to thank Dr. Wai Tuck Soh, Dr. Nyet Cheng Chiam, Hanna Rötschke, Yehor Horokhovskyi and John Cormican for proofreading this thesis. My special thanks goes to Prof. Dr. Henning Urlaub and Prof. Dr. Patrick Cramer for being the members of my Thesis Advisory Committee and to Dr. Sonja Lorenz, Dr. Alexander Stein and Dr. Johannes Soeding for reviewing this thesis . Finally, I acknowledge the financial support from International Max Planck School for Genome Science.

Contents

List of Figures

List of Tables

Abstract

CD8+ T-cells are the crucial component of the adaptive immune system as they survey the cells of the body for the signs of pathogens or tumours and eliminate them when necessary. To do this, T-cells recognise short antigenic peptides presented on the cell surface by Major Histocompatibility Complex class I (MHCI) molecules. The majority of MHCI bound peptides are generated by proteasome through peptide bond hydrolysis. Moreover, proteasome also generates hybrid peptides via proteasome catalysed peptide splicing (PCPS). In the recent years these peptides have been shown to be frequently produced by the proteasome, presented by MHCI molecules and elicit CD8+ T-cell responses, suggesting the importance of spliced peptides for adaptive immune response. However, the mechanisms of spliced peptides generation as well their contribution into adaptive immunity remain elusive. A particular area of concern is the effect of spliced peptides on the ability of the immune system to distinguish self and non-self and hence the repertoire of circulating CD8+ T-cells. Large sequence variability of spliced peptides could lead to the emergence of a large number of pathogenic peptides that are identical or highly similar to human antigenic peptides. In such cases, such matching pathogenic-human peptides may not be recognised by the T-cells as non-self which enables the immune evasion of pathogens. The mater of molecular mimicry is further complicated by the existence of T-cell receptor cross-reactivity. We showed that proteasome generated spliced peptides substantially enlarge the number of matching viral-human peptides on the theoretical level but owing to the expected frequencies of spliced peptides presented on the cell surface, the practical impact of the spliced peptides on the viral immune evasion is modest, even when considering T-cell cross-reactivity. An alternative scenario would be the failure to eliminate potentially auto-reactive CD8+ T-cells cells and the activation of such cells upon viral or bacterial infection leading to an autoimmune disease such as Type 1 Diabetes (T1D). Accordingly, we explored a potential role of such matching viral-human peptides in T1D and identified putative spliced peptides that could be presented in the pancreas and trigger its autoimmune destruction by self reactive CD8+ T-cells. A special type of MHCI presented peptides that could be viewed in the intersection of self and non-self peptides are the peptides that are derived from the tumor associated antigens. These peptides carry tumor specific mutations which simultaneously make them self and non-self. The knowledge of the impact of such mutations on the processing of antigens by the proteasome is not only important for a better understanding of the mechanisms of PCPS but also has implications for discovery of MHCI presented peptides for anti-cancer immunotherapies. We performed a comprehensive comparison of peptide products generation dynamics between the proteasomal digestions of a variety of wild type and mutated polypeptides and discovered that the impacts of the single amino acid exchanges were highly variable and dependent on the physical and chemical properties of amino acid that was replaced. In addition, we identified and quantified a number of potential MHCI binders that were derived from the digestions of the mutated polypeptides, a large fraction of which were spliced peptides.

Chapter 1

Overview and Aims

Proteasome is a large multi-subunit enzymatic complex [1, 2, 3, 4, 5]. It is the central component of the Ubiquitin Proteasome System (UPS). The main purpose of the UPS and proteasome is the degradation of cellular proteins that have either been damaged or misfolded due to defects in transcription and translation or have outlived their usefulness and need to be recycled. Proteasome processes proteins and polypeptides by cutting them into short peptides which are further cleaved into individual amino acids by the cytosolic enzymes called peptidases. The cutting occurs at the peptide bonds and is performed by catalytically active subunits located at the core of the proteasome. They induce peptide bond hydrolysis [5]. The second important function of the proteasome is the generation of the short peptides ranging in 8 to 14 amino acids in length that are bound to Major Histocompatibility Complex class I (MHCI) molecules and are presented on the cell surface [6]. Presentation of the peptides by MHC molecules is a crucial component of the adaptive immune system. It allows to to develop a specific and a directed response against various pathogens and tumors. The mounting of these responses is performed by the cells of the immune system call lymphocytes. Specifically, a subset of T- called cytotoxic T lymphocytes (CTLs) patrol the body and scan for presence of pathogenic proteins or mutated self proteins, driving tumorogenesis. The identification of these proteins occurs via the interaction of the cell surface receptors of CTL with peptides derived from these protein and bound to MHCI molecules. If it's determined that those peptides are derived from foreign organisms, the CTLs secrete a number of molecules that destroy the infected cells and prevent the further spread of the contagion or the tumor growth. Only a minor fraction of the peptides produced by the proteasome is presented on the cell surface, however. This is due to the fact the peptides have to be produced in the sufficient quantities and possess the optimal MHCI binding proteins. In addition, after the peptides are generated, they have to pass through the number of strict selection steps to be presented. This is referred to as the antigen presentation pathway (APP). The generated peptides have to be transported to the endoplasmic reticulum (ER) and bound to the vacant MHCI molecules which are then transported to the cell surface [6, 7].

Initially, it was thought that the proteasome exclusively generates peptides by peptide bond cleavages. However, eventually a major alternative catalytic activity of the proteasome was discovered. It was shown that the proteasome is capable of inducing ligation reaction between two shorter non-contiguous peptide fragments derived either from the same protein or polypeptide or from two separate peptide/protein molecules present in the proteasome. As a result of this ligation reaction (also known as transpeptidation) a novel peptide is formed, which sequence can't be re-capitulated in the original protein/polypeptide. This activity of the proteasome was termed proteasome catalysed peptide splicing (PCPS). The first spliced peptides in were identified almost two decades ago [8]. The generation of proteasome generated spliced peptides was thereafter demonstrated by multiple groups [2, 8, 9, 10, 11, 12, 13]. It was shown that the spliced peptides are generated *in vitro* in the polypeptides' digestions by purified proteasomes, *in cellulo* in a variety of cell lines [13, 14, 15]. Multiple groups also discovered that PCPS is not a random process but follows a set of rules [16, 17, 18, 19].

Initially, the role of PCPS was puzzling. However, later on It was shown that the spliced peptides enlarge the antigenic repertoire of the human cells [13, 14, 15]. It was theorised that the spliced peptides have a significant contribution into adaptive immune response by broadening the range of available targets for T-cells. This is because, when it comes to non-spliced peptides, there is a finite number of ways each given protein could be represented which significantly limits the diversity of epitopes that could trigger immune response responses. Spliced peptides on the other hand, could bypass limitations imposed by a protein sequence and efficiently present viral epitopes or epitopes with tumor-specific mutations [20, 21]. This opened up a potential to use spliced peptides for creation of peptide based anti-viral and bacterial vaccines and as a valuable strategy in the promising field of targeted cancer immunotherapies. For example, in one study a population of T-cells capable of uniquely recognising a spliced epitope from a melanoma associated protein was identified [22]. Treatment of melanoma patients with these isolated T-cells resulted in the complete recovery, highlighting the potential usefulness of spliced peptides. Several different groups subsequently managed to isolate T-cells reactive only to spliced peptides and inducing strong immunogenic responses [2, 20, 23].

The discovery of the PCPS further challenged our understanding of the adaptive immune response. The spliced peptides could be both beneficial for the adaptive immune response but also detrimental as they could increase vulnerability to microbial infections and autoimmune disease. Viruses and bacteria utilise a variety of strategies to evade the immune response. It has been hypothesised that pathogens such as viruses and bacteria could mimic the self-proteins via the short peptides derived from the digestion of pathogenic proteins that are identical or very similar to self derived peptide [24, 25]. As a result of presentation of such matching peptides, the immune system wouldn't be able to distinguish self from non-self and mount an effective immune response. On the other hand the presentation of such matching non-self peptides and their subsequent recognition could instead trigger pathogen mediated autoimmune diseases such as type 1 diabetes and multiple sclerosis [25, 26].

The other important gap of knowledge about peptide splicing, are the mechanisms that drive PCPS. Despite the fact that PCPS is a non-random process and is tightly controlled by proteasome, not much is understood about sequence preferences of proteasome that make PCPS more likely than canonical peptide-bond cleavage [27]. One of the open questions about the mechanisms of PCPS is the effect of single amino acid substitutions on the degradation of polypeptides by the proteasome. Several groups showed that introduction of even single substitution into the polypeptide sequence drastically changed the quantiative profile of the *in vitro* proteasomal digests [28, 29]. Understanding of systematic impacts of such amino acid substitutions on the generation of peptide products has significant implications for the discovery of promising MHCI peptide candidates, that could be used in anti-cancer immunotherapies.

There are still many unanswered questions about PCPS related to both its mechanisms and a relevance for the immune response. In the next chapter I will provide the theoretical background and highlight the current knowledge about the immune response, UPS, peptide splicing and the role of the spliced peptides in the adaptive immunity. I will then give insight into the several aspects of PCPS, that were the focus of my research. Within the scope of this Ph.D. thesis, I focused on the investigation of the following topics:

- The impact of PCPS on sequence identity between human and viral proteomes which leads to potential gaps in the T-cell mediated immunity and enables the immune evasion by viruses. We explore the topics of a viral immune evasion, a potential contribution of cis-spliced peptides into a molecular mimicry of viruses and a contribution of T-cell cross reactivity in the chapter 3 of this thesis

- The impact of flexibility of T-cell cross recognition on a molecular mimicry of the invading pathogens and a subsequent immune evasion. The contribution of T-cell cross reactivity into an extent of molecular mimicry, while considering cis-spliced peptides is discussed in the chapter 4 of this thesis

- The potential association of proteasome spliced peptides with CD8+ T cell mediated autoimmune destruction of pancreatic beta islets in T1D. We discuss a theoretical contribution of cis-spliced peptides into a molecular mimicry that leads to T1D associated autoimmune responses in the chapter 5 of this thesis

- The effects of single point amino acid substitutions on peptides generation by proteasome. The analysis of qualitative and quantitative impacts of such mutations are investigated in the chapter 6 of this thesis

Chapter 2

Background

2.1 Innate and adaptive immunity

Over the course of life, mammals are subjected to a vast array of pathogens and allergens which could significantly disturb normal functioning of organisms. The ever-present danger requires the organisms to possess an adequate protection. This protection is granted by immune system. On the one hand, it has to eliminate pathogens and toxins, however simultaneously, it's responses must be strictly regulated. The immune system is comprised of multiple cells and functions which goal is to defend host against both foreign (viruses, fungi, bacteria and parasites) and disregulated self antigens (*e.g.* cancer). The immune system consists of two broad elements - innate and adaptive immune system. Innate immune system is the branch of the immune system that performs rapid (minutes to hours) but non-specific responses [30, 31] (Figure 1). Adaptive immune system, on the other hand is highly antigen specific and is capable of generating immunological memory - *i.e.* an ability to remember the signatures of invading pathogens and rapidly activate upon subsequent encounters with them [30, 31]. However, upon the initial encounter, the adaptive immune system needs time to mount the necessary defences (days).

The innate immune system is comprised of four basic elements: structural barriers such as epithelial cells with tight cell-cell contacts and mucous membranes, physiologic barriers, endocytic and phagocytic barriers and finally inflammatory barriers [30, 31] (Figure 1). A key element of the innate immune system are membrane bound and cytoplasmic pattern recognition receptors (PRRs) such as Toll-like receptors and Leucine-Rich Repeat (NLR) proteins which recognise the molecular signatures of pathogens known as microbial associated molecular patterns (MAMPs) such as lipopolysaccharides (LPS), flaggelar proteins, peptidoglycans, lipoteichoic acid and viral nucleic acids [30]. An activation of the innate immune system is necessary for the subsequent activation of the adaptive immune system. Once activated, the innate immune system promotes production of specific proteins and bioactive small molecules, such as inflammatory molecules including cytokines that regulate the response of a variety of immune cells and chemokines that attract leukocytes during inflammation, such as tumor necrosis factor alpha (TNFalpha), interleukin 1 and 6 (IL-1, IL-6). The secretion of these molecules promotes the infiltration of the immune cells where they're needed. [30, 31]. An important component of the innate immune system is complement

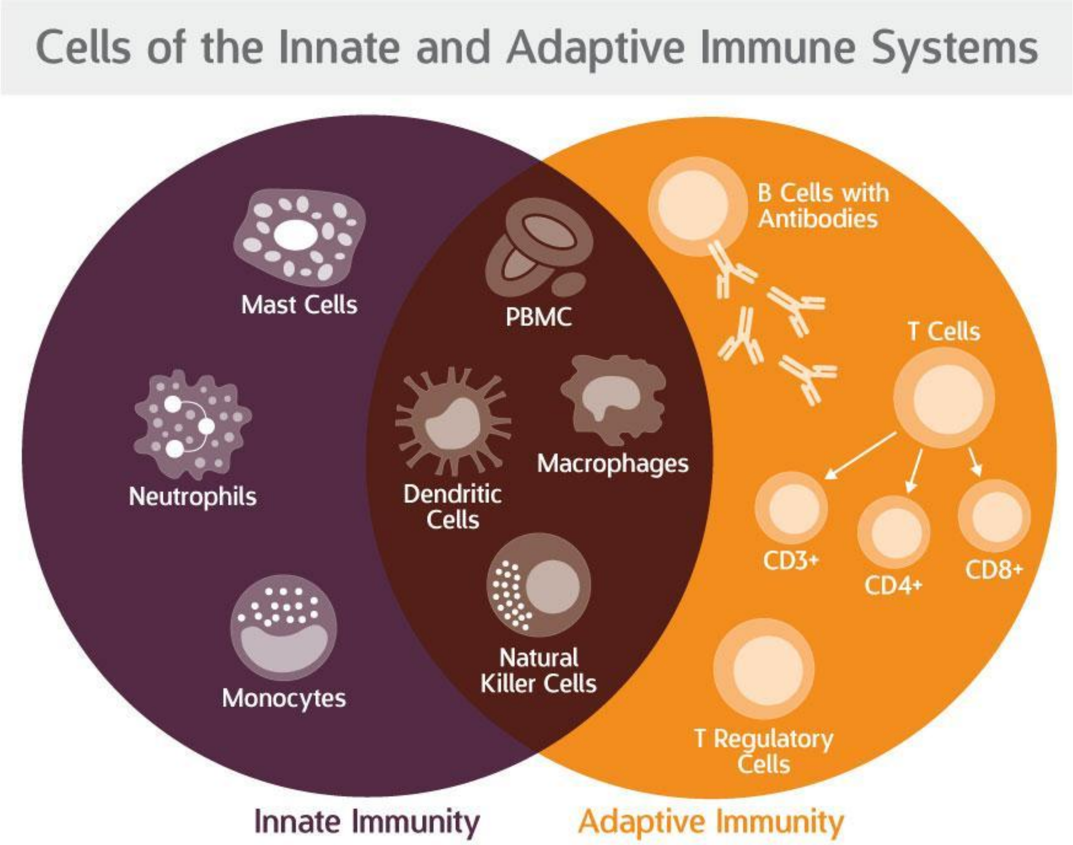


FIGURE 1: The key components of the adaptive immune system
The figure was retrieved from https://cellero.com/blog/immunology-for-non-immunologists-innate-vs-adaptive-immunity/

which makes the invading pathogens susceptible to phagocytosis - engulfing of the pathogens and the infected cells by immune cells [30]. The complement consists of a network of 25 plasma and cell surface proteins, many of which are proteinases. Complement tags targets (both the infected cells as well as bacteria and viruses) for destruction.The elements of the innate immune system are omnipresent and can be encountered in many different cells [31].

Several different cell types comprise the innate immune system. These includes monocytes phagocytes (macrophages and neutrophils), dendritic cells, mast cells, basophils, eosinophils, natural killer cells (NKs) and innate lymphoid cells [30]. Importantly, a number of these cells are critical for the adaptive immune response, highlighting the close cross-talk between the two branches. Phagocytes surround the invading pathogens and eliminate them via a variety of enzymatic pathways. Macrophages produce IFN-gamma, IL-6, IL-12 and TNF alpha and possess strong anti-bacterial capabilities [31]. Monocytes and macrophages also act as antigen presenting cells (APCs). Neutrophils are additionally involved in the antigen presentation. They produce reactive oxygen species that is cytotoxic to the bacteria [31]. They were also shown to produce TNF alpha and IL-12. Monocytes are likewise are phagocytic towards pathogens. Dendritic cells are also capable of both directly eliminating pathogens via phagocytosis but also act as major antigen presenting cells (APCs) in the adaptive immunity. Mast cells

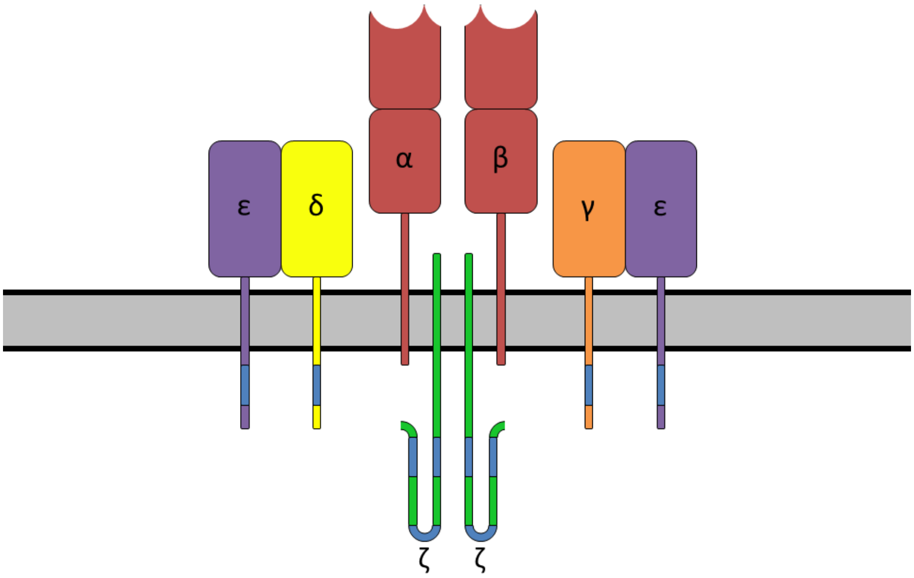


FIGURE 2: The schematic representation of the structure of the T-cell receptor
The figure was retrieved from https://en.wikipedia.org/wiki/T-cell_receptor

and basophils are important for the induction of inflammatory responses. Mast cells are mostly found in the connective tissue around the blood vessels, while basophils are circulating. Mast cells additionally produce cytokines to attract the infiltrating immune cells. Eosinophils primary purpose is destruction of large parasites (*e.g.* parasitic worms) that can' be engulfed. This is achieved via cytoplasmic granules than contains potent enzymes [31] NK cells are instrumental in the destruction of tumor and virus infected cells via the secretion of granzymes and perforins from their granules. NK cells produce a major cytokine - IFN gamma which is crucial for the anti-viral immune responses, and destroy cells infected with intracellular pathogens [30]. They are derived from pluripotent hematopoietic stem cells which differentiate into the the common myeloid progenitor cell [31]. The important function of NK is destruction of virus infected cells which were manipulated to downregulate their MHCI molecules [31]. Finally, Innate lymphoid cells secrete cytokines such as IL-4 and IL-17 important for the modulation of the strength of the immune response. Neutrophils, monocytes, macrophages, eosinophils, basophils, and mast cells are derived from the myeloid stem cells [31]

While the innate immune system serves as the first line of defence against the pathogens and alerts the immune system at large to their presence it's not effective at eliciting a directed and specific responses (Figure 1). Such responses are accomplished by a multitude of cells of the adaptive immune system. The primary functions of the adaptive immune cells are the recognition of the signatures of pathogens or disregulated self antigens, the mounting of the specific responses and the elimination of both the pathogens in the intercellular space as well as the cells infected with said pathogens. Finally, the adaptive immune system generates an immunological memory that allows it to rapidly recognize and deal with the pathogens that were previously encountered. Upon recognition the cells of the adaptive immune system expand and multiply in order to respond to the pathogens and tumors. Due to this requirement for recognition and proliferation, the adaptive immune system isn't activated rapidly but after some time following an elicitation of the innate immune response [31]. Broadly, the adaptive immune system can be split into three equally important parts: antigen presenting cells (APCs) that present antigens to T-cells, T-cells that are activated by APCs and B-cells which generate antibodies [30, 31]. Cell surface receptors of B-cell and T-cells are encoded by genes, variable junction regions of which mix and match in the process called somatic rearrangement to achieve a high variability of their receptors capable of responding to a variety of antigens [31]. Innate and adaptive immune systems don't act in vacuum but rather in close cross-talk, where the innate immune system is instrumental in a delayed activation of the components of the adaptive immune system, while at the same time the adaptive immune system utilises the effector mechanisms of the innate immune system [31].

The inception of the T-cells happens in the bone marrow from hematopoietic stem cells through the common myeloid progenitor cells after which they migrate to thymus where they go through selection steps before being released into the circulation [31]. The maturation of T-cells in the thymus is necessary for fine-tuning their repertoire and for the elimination of T-cells that could trigger inappropriate immune responses to immunological self. T-cells possess antigen binding receptors - T-cell receptors (TCRs) (Figure 2). These are transmembrane heterodimeric proteins. One T-cell carries one TCR. All T-cells initially express alpha-betaTCR which subsequently differentiate into several types, being CD8+ T-cells and CD4+ T-cells. Each T-cell possess alpha-beta-TCR with particular range of specificities [31]. Both alpha and beta chains of the TCR are comprised of the V domain that determines antigen specificity, C domain which connects TCR to the membrane and a transmembrane domain [33]. The TCRs themselves are highly polymorphic due to the process or somatic gene rearrangement of junction regions between the genes encoding TCRs. The TCRs are constructed from variable (V), diversity (D) and joining (J) segments which generate ValphaJalpha chains and VbetaDbetaJbeta chains. During the assembly, the regions around these gene segments are cleaved by specific RAG1 and RAG2 enzymes which are then joined by recombinase enzymes [31]. Due to a variable cleavage around those segments a large variety of junctions emerge. In addition, the V, D and J segments are assembled seemingly randomly which drives the remarkable diversity of the TCRs. In addition to the TCR, the T-cells also express ancillary CD3 molecules. CD3 molecules are non-covalently bound to the TCR and consist of three subunits - CD3-delta-epsilon, CD3-gamma-epsilon and zeta-zeta [33]. All of those subunits incorporate immunoreceptor tyrosine-based activation motifs (ITAMs) in their cytoplasmic domains which is necessary for both the TCR expression and transfer of an activation signal downstream. CD8 molecule is required for the initiation and the enhancement of the signal. In addition, CD8+

T-cells contain CD45 glycoprotein which is a positive regulator of the TCR signalling.

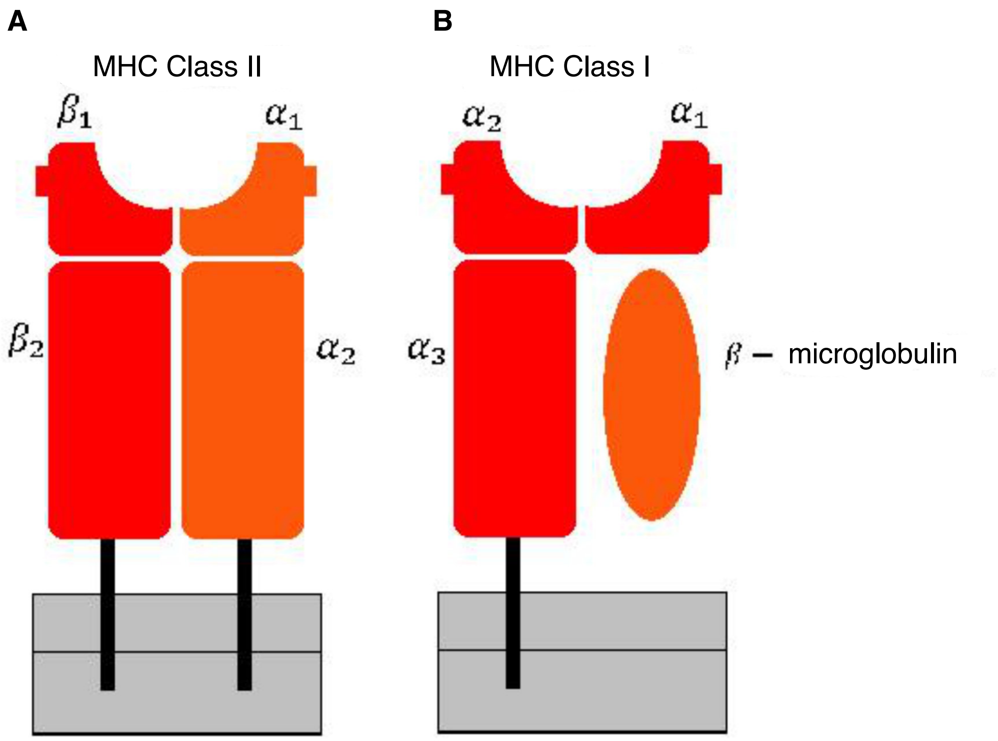


FIGURE 3: The schematic representation of the structure of the MHC class I and II molecules
The figure was retrieved from https://en.wikipedia.org/wiki/Major_histocompatibility_complex

The activation of the T-cells is only possible through the interactions with APCs (dendritic cells, macrophages and B-cells) which display the signatures of the antigens and the appropriate co-stimulatory signals. Depending on the identity of the type of T-cell, they interact with two distinct types of peptide-receptor complexes displayed on the surface of APCs - Major Histocompatibility Complex (MHC) molecules (Figure 3). The MHC molecules are glycoproteins that are displayed on a cell surface and in turn present short peptides derived from antigens - cellular, abnormal or non-self proteins for the recognition by the adaptive immune system [31]. In human, MHC molecules are known as Human Leukocyte Antigens (HLAs). The MHC molecules are broadly split into two classes - class II and class I (Figure 3A and Figure 3B). Between the two over nine thousand five hundred and forty-six polymorphisms of the HLA gene region have been documented [34]. MHC class I molecules (MHCIs) present cytosolic peptides generated as a result of the degradation of the proteins of intracellular pathogens (viruses and some bacteria) and tumors to CD8+ T-cells [30]. The MHC class I molecules are comprised of a highly variable 44 kDa transmembrane component (alpha chain, class I heavy chain) and a conserved 12 kDa beta2 microglobulin light chain [31, 35] (Figure 3B). The alpha-chain consists of three extracellular domains (alpha1, alpha2 and alpha2), the transmembrane domain and a short intracellular domain that attaches MHCI to the cell surface. They present short peptide sequences of 8-15 amino acids in length to CD8+ T-cells [35]. These peptides are released from the cytosolic proteins which are primarily disassembled by ubiquitin proteasome system (UPS) [31]. Despite a great diversity of the MHCI alleles that can be expressed in the human cells, not all of them are presented on the cell surface. It's believed that although up to 6 MHCI molecules can be produced in a given cell and a given individual, only about 4 will be presenting antigenic peptides [36]. In human, the MHCI molecules are represented by HLA -A,-B and -C [30]. There is a large diversity of HLA alleles - 650 for HLA-A, 1000 for HLA-B and 350 for HLA-C. The TCRs of CD8+ T-cells interact with both the exposed molecular surface of the antigenic peptide and the flanking regions of the MHCI molecules. The TCRs have very low affinity for floating antigenic peptides or empty MHCI molecules [31]. The alpha3 domain of the class I heavy chain simultaneously interacts with CD8 molecules of the CD8+ T-cells [31]. Both the TCR and the CD3 molecule have to interact with MHCI-peptide complex for the T-cell to be activated. In addition, it's paramount, that the TCR also interacts with co-stimulatory molecule, CD28 expressed by APCs for the full activation [31]. In absence of CD28 co-stimulation, CD8+ T-cells become anergic and enter apoptosis (Figure 4A).

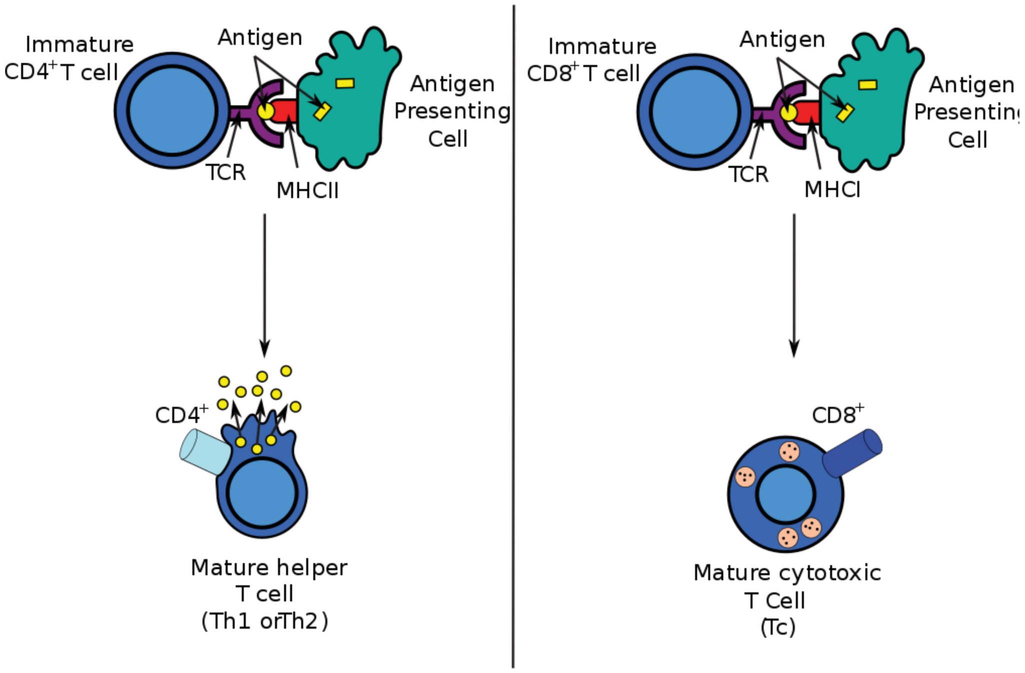


FIGURE 4: The mode of action of CD4+ and CD8+ T-cells
The figure was retrieved from https://en.wikipedia.org/wiki/Cytotoxic_T_cell

By contrast, MHC class II molecules present extracellular peptides to CD4+ T-cells derived from extracellular pathogens such as bacteria and the pathogens that are yet to enter the cells. In human they are represented by HLA DP, DQ and DR molecules (Figure 3A). T-cells are activated when they encounter APC presented antigenic peptides with the optima binding properties to the TCR [30, 31]. T-cells also secrete cytokines upon the recognition of MHC-peptide complex which are necessary for the modulation of their response. The two types of T-cells are CD8+ T-cells (cytotoxic T lymphocytes) and CD4+ T-helper cells (Th cells) (Figure 4B). CD8+ T-cells recognise peptides from the intracellular pathogens and tumors. They are necessary for elimination of pathogens that were not stopped by the antibodies secreted by B-cells and that entered the cells. Once activated, these T-cells expand and induce apoptosis of the target cells. Some of these cells remain in circulation and act as memory T-cells which can quickly re-expand in response to the infections with familiar pathogens. By contrast, Th cells don't directly kill the infected cells but rather direct other cells to perform the immune response. Th cells are activated after the recognition of MHCII presented peptides. They differentiate and secrete cytokines driving the immune response [30, 31]. There exist three primary types of T helper cells - Th1, Th2 and Th17 and Tregs, although there are other subpopulations of T helper cells. Th1 cells secrete IFN-gamma which activates macrophages and elicits anti-microbial immunity. They also assist in the differentiation of B-cells. Th2 cells secrete IL-4, IL-5 and IL-13 which drive the expansion of B cells secreting immunoglobulin E (IgE) and IgG. They also attract mast cell and eosinophils to the sites of infection. IgE in particular are associated with allergic responses and thus the strict regulation of their production Th2 is necessary. In general Th1 drive cell-mediated immune responses while Th2 are mainly responsible for the humoral responses and allergic reactions [31]. Th17 cells mainly secrete IL-17 which modulates the inflammatory responses particularly during chronic infections. T regs modulate and suppress immune responses when necessary. The defects in this population of Th may lead to autoimmune diseases.

B-cells also originate in the bone marrow from hematopoietic stem cells but compared to the T-cells they are not activated by APCs. They generate protein receptors - immunoglobulins (antibodies) that recognize pathogens that have not yet entered the host cells and allergens. Upon activation by Th cells via for example IL-6, B-cell proliferate and differentiate either into plasma cells that secrete antibodies or memory cells. Like memory T-cells, the memory B-cells can remain in the circulation for a long time after the infection and can be rapidly activated for antibody production if there is a reinfection with same or similar pathogens [30, 31]. The antibodies that are produced by plasma cells bind to the pathogens which promotes their destruction through complement activation, phagocytosis and activation of the effector cells. Different antibody types are structurally distinct and respond to specific subsets of antigens. Defects in the regulation of these branches of the adaptive immune system may result in the autoimmunity - an abnormal immune response to the immunological self via self-reactive T-cells and auto-antibodies, inappropriately long or intense inflammatory reactions and immunodeficiencies, in which the capacity to respond to the pathogens is curtailed [30]

2.2 CD8+ T-lymphocytes and their role in the response to viral infections and cancer

During a process of positive and negative selection, double positive alpha-beta T-cells differentiate into two distinct types - CD4+CD8- T-cells and CD8+CD4- T-cells. CD4+ T-cells only interact with MHCII molecules, while CD8+ T-cells - with MHCI molecules. About 60-70% of T-cells are CD4+, while 30-40% are CD8+ [31]. CD8+ T-cells display a strong cytotoxic activity towards host cells infected with intracellular pathogens or abnormal cells. Effector molecules that are produced by CD8+ T-cell include granzymes, perforins, cathepsin C and granulysin [33, 37]. CD8+ T-cells also express the Fas Ligand (FasL) which upon interaction with Fas receptors on the target cells further contributes to their death. They also secrete cytokines such are TNF alpha and IFN gamma necessary for the defence against intracellular pathogens. They also produce chemokines that recruit additional immune cells to the infection sites.

If there is an overactivation of CD8+ T-cells, it can cause unnecessary tissue damage and autoimmunity. To prevent this CD8+ T-cells express inhibitory receptors to modulate their immune responses. These include programmed cell death receptor 1 (PD-1) and CTLA-4, lymphocyte- activation gene 3 (LAG-3), T-cell immunoglobulin and mucin domain-3 (TIM-3), T-cell immunoreceptor with Ig and ITIM domains (TIGIT) and inducible T-cell co-stimulatory receptor (ICOS). The production of these proteins is used to monitor the effects of the immunotherapies. PD-1 elicits it's activities through binding to PD-L1 and PD-L2 ligands on the surface of the target cells and interferes with CD8+ T-cell signal transduction. This results in the decrease in cytokine secretion and the cell cycle arrest. CTLA-4 receptors bind the same targets as CD28 co-stimulatory molecules but with significantly higher avidity. The activation of CTLA-4 decreases the interactions of CD8+ T-cells with APCs. PD-1 mainly acts during the activated, effector phase of CD8+ T-cells, while CTLA-4 is activated during the priming of the CD8+ T-cells [31, 33, 37].

Different pathogens can cause both acute (*e.g.* influenza virus) and chronic infections (*e.g.* herpes virus). CD8+ T-cells either promptly remove all traces of the pathogens or keep their replication suppressed, which is contingent on the type of virus the immune system has to deal with. It's worth considering the response to several representative viruses in more detail. Lymphocytic choriomeningitis virus (LCMV) is a virus that mainly infects mice macrophages, lymphocytes, dendritic and glial cells [38]. It causes acute viral infections that induce strong CTL responses. In mice it was shown that CD8+ T-cells primarily respond to epitopes derived from viral glycoprotein and nucleoprotein. The abundance of these epitopes strongly correlated with the magnitude of CTL responses, however the degree of immunogenic response wasn't necessarily causal with the capacity to clear the infection. The absence of CD28 co-stimulatory signals interestingly didn't completely prevent the CD8+ T-cell mediated responses. Memory CD8+ T-cells were shown to persist in mice for the long time after the infection and require lower level of stimulation in order to be activated. Interestingly, the number of such memory CD8+ T-cells was affected by the infection with other pathogens. Remarkably LCMV-specific CD8+ T-cells were demonstrated to be corss-reactive with antigens from unrelated pathogens further suggesting the relevance of TCR degeneracy [38].

Respiratory viruses that infect humans are numerous and include respiratory syncytial virus (RSV), influenza virus, human metapneumovirus, rhinovirus, coronavirus, and parainfluenza virus. Depending on the virus and the individual susceptibility, they can cause a number of symptoms ranging from mild to life threatening [39]. So far, most of the vaccination strategies attempt to elicit the anti-viral response via broadly neutrolizing antibodies. However, in order to trigger a long lasting and effective immune response, other aspects of adaptive immunity are necessary. Increasingly, the role of CD8+ T-cells has been appreciated. Generally, after a respiratory infection, the dendritic cells present viral antigens to CD8+ T-cells in lung draining lymph nodes which leads to their activation and anti-viral response. Once activated, CD8+ T-cells, secrete inflammatory cytokines as well as elicit effector responses via production of granzymes A and B and perforins to lyse the infected cells. CD8+ T-cells can also eliminate the infected cells via cell-cell contacts via the interactions of cell surface proteins Fas and FasL. Memory CD8+ T-cells were detected in the peripheral blood for months after the respiratory infection in both mice and humans, although their numbers drop with age. These memory CD8+ T-cells can be quickly re-activated following reinfection and produce anti-viral cytokines such as TNF-alpha and IFN-gamma. Several attempts to generate CD8+ T-cells based vaccines were made including recombinant Listeria monocytogenes- (DC-LM) or vaccinia virus-boost (DC-VV) strategies, based on the presentation of immunodominant epitopes by dendritic cells to naive CD8+ T-cells. They showed promising results in mice studies [39].

Influenza are respiratory viruses that infect the airways and lung cells via two surface proteins - hemagglutinin and neuraminidase (HN) [38]. These two proteins quickly evolve which leads to the immune evasion of the new strains of influenza. CD8+ T-cells were shown to be the crucial line of defence against influenza. The memory CD8+ T-cells are capable of the rapid response to the infections with influenza. The magnitude of response was demonstrated to be collectively determined by a thymic selection of CD8+ T-cells, the assortment of TCRs, the kinetics of influenza antigen presentation and the composition of the antigenic peptides. The immunoproteasome was shown to be instrumental in the generation of antigenic peptides most readily capable of eliciting strong immunogenic responses. The strongest responses were observed for HA and influenza virus polymerase 2 protein. Interestingly, there were substantial discrepancies in the extent of the TCRs specificity and cytokine production, particularly IFN-gamma, depending on the epitope's origin. Presence of co-stimulatory CD28 molecules was crucial for the activation of CD8+ T-cells and the cytotoxic response. The memory CD8+ T-cells were demonstrated to persist in the tissues long after the infections.

Respiratory Syncytial Virus (RSV) is single stranded RNA virus that primarily infects lower respiratory tracts in young children and the elderly [38]. RSV was shown to suppress CD8+ T-cell activity in the lungs by interfering with with TCR signalling in the lungs. Most of CD8+ T-cells (50%) were responding to a single epitope from matrix 2 protein of RSV and a single epitope from fusion glycoprotein of RSV. Due to the interference with TCR singling, the cytokine production of CD8+ T-cells was dampened and the immunity provided by these CD8+ T-cells was relatively short lived.

Herpes simplex virus (HSV) is lytic virus that depending on the circumstances can cause a variety of outcomes. HSV-1 causes encephalitis and oral infections, while HSV-2 infects the genitals [38]. It primarily targets skin and neurones. Unlike the viruses that cause acute infections, HSV can persist and cause latent infections. HSV is known to perform the immune evasion from MHCI presentation by the expression of ICP47 which inhibits TAP. CD8+ T-cells were shown to control HSV replication and keep it in latent state by expression of antiviral cytokines and granzyme A. Just like in most cases, co-stimulation with CD28 is essential. CD8+ T-cells mainly respond to the epitope from HSV glycoprotein B (70-90%), ribonucleotide reductase and the immediate-early protein ICP27. It was shown that the germline-encoded TCR elements coupled to VDJ recombination were crucial in equipping T-cells with the ability to provide potent viral responses.

Hepatitis B virus (HBV) is a DNA virus infecting hepatocytes in the liver. It may lead to hepatitis [38]. Interestingly, CD8+ T-cells were shown to degrade HBV transcripts without the inhibition of transcription of HBV genes. IFN-gamma and TFN-alpha expression was shown to be crucial in the process of viral clearance. Remarkably, the direct contact of CD8+ T-cells wasn't necessary for the removal of HBV and was mainly achieved by cytokine secretion. In addition, the expression of certain chemokines by CD8+ T-cells attracts additional mononuclear cells to inflammation regions.

Gamma-Herpesvirus (gammaHV) initially cause acute infections in mucosal tissues as well as B cells and dendritic cells, but can persist in latent state for a long time to come by for example inducing latency in B cell responses. Two viruses from that family that should be highlighted are Epstein-Barr virus (EBV) and Kaposis sarcoma herpesvirus (KSHV/human herpesvirus 8). They are thought to be linked to multiple disorders such as cancer, multiple sclerosis and type 1 diabetes (T1D) in case of EBV [38, 40, 41]. In mice models of gammaHV, CD8+T cells were demonstrated to be crucial for the viral removal during the acute phase. This is mainly achieved via granzyme and perforin secretion by CD8+ T-cells. However, CD8+ T-cells while necessary can't control the viral infection. CD4+ T-cell responses are likewise necessary. In absence of Th cells there was a decreased expression of IFN-gamma and TNF-alpha expression by CD8+ T-cells. Primarily, CD8+ T-cells respond to epitopes derived from single-stranded DNA-binding protein (p56) and ribonucleotide reductase (p79). gammaHV can perform an efficient immune evasion. In mice gammaHV - gammaHV68, This is achieved via for example K3 gene encoding a zinc-finger-containing protein that decreases a half-life of MHCI molecules. M3 protein of gammaHV68, binds to chemokines and interferes with the migration of CD8+ T-cells to the infection sites.

While the majority of intracellular pathogens are viruses there are some examples of bacteria that reside within the cells, making them a target for CD8+ T-cells [38]. One prominent example of such bacteria is Listeria monocytogenes which enters the cytosol. This makes L.monocytogenes a prime target for a proteasomal degradation and MHCI presentation to CD8+ T-cells. L.monocytogenes is a gram-positive bacteria that infects both mice and humans. It can enter a variety of cell-types including epithelial cells, hepatocytes and bone-marrow cells. Broadly, CD8+ T-cells responding to L.monocytogenes can be split into two groups. One group responds to peptides derived from secreted bacterial proteins while the other interacts with peptides from nonsecreted proteins that contain N-formyl methionine at the amino-terminus. The antigens to which CD8+ T-cells rapidly respond include bacterial secreted proteins, such as p60, ActA and Listeriolysin O (LLO) - the membranolytic virulence factor and bacterial phospholipase PlcB. Interestingly, it was shown that the most efficient activation of CD8+ T-cells was achieved with the peptides derived from the secreted antigens compared to non-secreted ones. The other interesting observation was that CD8+ T-cells were the most responsive to the inefficiently presented epitopes compared to more numerous and efficiently presented ones, demonstrating that while the rate of antigen presentation is important, it's not the only factor that determines the efficiency of the immune responses. The memory CD8+ T-cells remained in circulation for a long time after the infection and the most effectively re-stimulated CD8+ T-cells were those with the highest affinities to the presented epitopes. The secretion of IFN-gamma was shown to be crucial for granting the protective effect of CD8+ T-cells against bacteria. Platteel *et al.* introduced a reverse-immunology based approach for the identification of the novel bacterial epitopes and identified the two proteasome generated spliced peptides produced as the result of the degradation of PlcB, that were demonstrated to elicit unique CD8+ T-cell responses non-cross-reactive with the TCRs recognising canonical peptides [20]

The other crucial role of CD8+ T-cells is the surveying of cells for signs of the abnormal cellular processes such as those that arise during tumorogenesis and elimination of such cells [33, 42]. CD8+ T-cells are the most important players in the anti-cancer immunity. It has been known that the presence of tumor infiltrating CD8+ T-cells in tumor tissues is indicative of the better survival of patients in a variety of cancers, such as melanomas, colorectal cancer (CRC) and ovarian cancer. This was demonstrated by infusions of CD8+ T-cells which were previously expanded *in vivo*. Naturally, most of the antigens that are displayed by APCs in tumor are self and thus most of the T-cells that could respond to them were removed during the process of the negative selection in the thymus. Those T-cells that are potentially self-reactive may leak into the periphery, however they possess lower affinities to antigens compared to the virus-specific CD8+ T-cells [42]. The matter is further complicated by the suppression of the T-cells and their responses by the tumor micorenvironemnt (TME). This can happen via the downregulation of the MHCI molecules, production of proteins degrading the enzymes, secreted by the activated CD8+ T-cells and upregulation of the immune-checkpoint molecules [33]. Tumor cells express ligands for immune-checkpoint receptors such as PD-1 ligand (PD-1L) and CD80 and CD86 - CTLA-4 ligands. Tumors are capable of inhibiting the expansion of CD8+ T-cells. Tumors can also suppress the maturation of APCs. The degree of infiltration of CD8+ T-cells is an important predictor of anti-tumor immune responses. TME is also known to create metabolic conditions incompatible with CD8+ T-cells survival such hypoxia and acidity [33].

The blockages of T-cell cytotoxic T lymphocyte-associated antigen 4 (CTLA4) and programmed cell death protein 1 (PD1) resulted in both induction of naive T-cells in case of inhibition of CTLA4 and enhancement of the response of the existing CD8+ T-cells [42]. However, the success of such interventions has been limited due to a large subset of patients being resistant to anti-PD1/anti-CTLA4 therapies - so called immune checkpoint blockade therapies (ICBs). These therapies employ monoclonal antibodies that target PD-1 and CTLA-4 and inhibits their activities. While such treatments were very effective in some patients, others have shown a significant resistance to such interventions [33]. Typically, the activated CTLs are found at their infiltration sites while naive and memory CD8+ T-cells are present in the circulation. It was observed that among all of the T-cells in the TME, only a fraction is capable or recognising and responding to tumor associated antigens. The other CD8+ T-cells are call bystander T-cells and may include CD8+ T-cells that were previously primed by the vital infection and are not directly relevant to TME. One way to distinguish the relevant CD8+ T-cells is by assessing their degree of expansion and the responses to anti-PD1 therapies. Moreover, tumor reactive CD8+ T-cells were shown to share the recurrent TCRs, in addition to these three primary repertoires, one more pool of anti-tumor CD8+ T-cells was identified - cells that exhibit the traits of dysfunction/exhaustion [42, 43]. The dysfunctional CD8+ T-cells are characterised by the expression PD1 molecules and CTLA4 limiting their ability to expand and elicit their protective effects. This is manifested via a progressive decrease in the expression of the key cytokines - IL-2, TNF-alpha and IFN-gamma. The fraction of such dysfunctional T-cells varies between 5 and 80% depending on the tumor, which could explain the variable responses of CD8+ T-cell in a number of cancers. The large number of CD8+ T-cells that were in the late dysfunctional state was correlated with the poor anti-tumor activity [43]. Interestingly, cytotoxic CD8+ T-cells and dysfunctional CD8+ T-cells were found to share overlapping TCRs suggesting cross-talk between these two states. The main cause of CD8+ T-cell dysfunction was shown to be the continuous exposure of CD8+ T-cells to the tumor antigens that increased the expression of the inhibitory receptors. Counterintuitively, the increased expression of the genes associated with T-cell activation was observed which could serve as evidence of a continuous exposure of the CD8+ T-cells to tumor associated antigens. It was speculated that the entry into dysfunctional state is linked to the anti-tumor reactivity of CD8+ T-cells due to the fact that they continuously interact with tumor derived epitopes. The other CD8+ T-cells, however could be classified as bystander cells since there are not stimulated by the tumor derived antigens. It is currently being investigated if targeting of such dysfunctional cells by PD1 and CTLA4 inhibition could enhance their reactivity and whether such treatments would have variable effects of CD8+ T-cells that have entered the dysfunctional state recently or not [42]. Currently, it is hypothesised that the CD8+ T-cells with low levels of dysfunction are the best equipped group of cells to respond to tumors, while terminally exhausted CD8+ T-cells will fail to adequately respond to tumors. Further studies will be crucial for the identification of CD8+ T-cell states, primarily activated CD8+ T-cells and CD8+ T-cells in the early dysfunctional states, that are predictive of the desired T-cell repertoire and thus the successful anti-tumor CD8+ T-cell based therapies.

The primary goal of the cancer immunotherapies is to induce sustained anti-tumor responses while avoiding autoimmunity [43]. Such approaches include aforementioned immune checkpoint blockades. Many of the currently ongoing trails of the adoptive T-cell transfer (ATT) and tumor-infiltrating lymphocyte (TIL) therapy focus on targeting of the tumor-associated antigens which are nevertheless not mutated. The downside of these approaches is that forced activation of previously exhausted CD8+ T-cells could lead to a severe autoimmune reactions with lethal consequences. It's thus, necessary to identify and isolate CD8+ T-cells which react specifically to the mutated tumor antigens, such as those containing single-point amino acid substitutions or those antigens, which were transcribed from altered mRNA transcripts due to for example exon-skipping, gene fusions and extension of the coding sequence beyond the wild type stop codon [29, 43, 44, 45, 46, 47]. These targeted approaches rely on an exploitation of tumor specific antigens and antigenic epitopes containing point mutations. Such CD8+ T-cells would be required only target the tumor cells without damaging the periphery. For example, Tran *et al.* identified several neo-eptiopes, derived from mutant KRAS G12D protein and presented by HLA-C*08:02 molecules. They were able to isolate CD8+ T-cells from a patient with metastatic colorectal cancer, reactive to the antigenic peptides from the mutated KRAS, containing the position with the mutation. These specific CD8+ T-cells were expanded and infused into them into the patient. The CD8+ T-cells produced IFN-gamma, TNF-alpha and IL-2. The infusion resulted in the regression of the metastatic lung tumors and a remission of the tumors 4 months after the treatment. The TCRs were reactive only to KRAS G12D derived epitopes and didn't recognise wild type analogous [45]. Adoptive T-cell transfers proved to be effective in management of the metastatic melanoma [48].

It's a well known fact that all tumors carry antigens with somatic mutations that in large part enable their uncontrolled growth and apoptosis evasion. About 1% of somatic mutations occur in coding regions which result in altered proteins. The majority of these mutations lead to the single point amino acid substitutions [46]. Some of this mutations can lead to oncogenicity. In some cases, frameshift mutations or chromosomal translocations (*e.g.* BCR-ABL, ETV6-AML1 or TEL-AML) can occur which could increase the number of potential tumor-specific epitopes. In some cases, such mutations result in the enhanced binding to the MHCI molecules, while in others the wild type version of the peptide could also bind to

MHCI but wasn't recognised by CD8+ T-cells compared to the mutated epitope [47].Many of single point mutations are highly individual while some tend to be present in several types of cancer. Some of such re-occurring mutations are known to emerge in BRAF and KRAS proteins in melanomas and pancreatic and solid cancers, respectively [48]. It was estimated there are about 140 of such mutations. The probability of the correct epitopes being presented by APCs increases with the number of mutations. There is a number of variables that can significantly affect the rate of presentation of such epitopes such as antigen expression level, it's proteolytic processing, transport to ER and binding to MHCI molecules. When it comes to canonical, non-spliced epitopes, only a small fraction of those is known to be produced and presented. For example, in vaccinia virus, 100 peptides have measured binding affinities to a frequently encountered HLA-A*02:01 molecule of 100 nM or less. Of those, 15% were processed by the proteasome and presented on the cell surface and of these, only 11% could elicit CD8+ T-cell responses [48].

The matter is further complicated by a varying preference of different epitopes for different HLA molecules [46]. The location of the mutation can also affect the presentation of mutated epitopes compared to wild type versions. If the mutations occur at one of the anchor sites, such epitopes could be presented both more or less efficiently. Alternatively, if the mutation is at one of the TCR contact residues, this could affect it's recognition by the CD8+ T-cells. In addition, these conceptually attractive epitopes could be inefficiently generated and transported into ER.

Traditionally, the most sought after epitopes have very high affinities to the respective HLAs [46]. However there are also examples of TCRs targeting neo-epitopes with binding affinities above 100 nM which still led to the destruction of the tumor. Thus, it appears clear that both the MHCI binding affinity and the strong interaction of the epitope with TCRs are instrumental in determining the efficacy of the immune response. The CD8+ T-cells with the desirable TCRs may be isolated from TILs, Plasmablastic Lymphoma (PBL) and peripheral blood mononuclear cells (PBMCs). The ability of the T-cells to induce immunogenic responses can be tested *in vitro* against for example peptide-HLA tetramers and clonally expanded if necessary [47, 48]. Single cell PCR was proposed to be a robust method for screening for the optimal TCRs. In addition, TCRs have to be screened for binding to self-antigens due to for example T-cell cross-reactivity [47].

One of the promising methods for the identification of tumor-specific mutations that is extensively used by our group in collaboration with Mishto *et al.*, involves chemical synthesis of 20-40 amino acid long polypeptides recapitulating the sequence of the mutated antigen of interest which contain a position with a mutated residue flanked by 10-20 residues. It's then checked *in silico* if the peptides derived from such polypeptides contain the mutation and are good binders. Their generation could be checked *in vitro* in the proteasomal digestions and finally, the peptides which production was confirmed can be chemically synthesised. The

TILs or T-cells from PBMCs could then be surveyed for their ability to recognise such synthetic peptides and elicit immunogenic responses [48].

In general, there are several caveats to these targeted approaches. It's still a question, whether the mutated epitopes are patient specific, or could also be recurrent. The issue of obtaining sufficient number of epitope-specific CD8+ T-cells and the tumor heterogeneity also need to be addressed [46]. Tumor heterogeneity is an issue in particular as most cancers are comprised of diverse populations of cancer cells with unique signatures. Since, the requirement is that most of the cancer cells should present one shared mutated epitope. Alternatively, ATT approaches should be designed to target multiple different tumor-associated mutations.

In addition to human mutated antigens, the other promising avenue of research are antigens of viral origins that could be used for treatment of cancers. These antigens are derived rom viruses demonstrated to be linked with a subset of human tumours such as cervical carcinoma, hepatocarcinoma, nasopharyngeal carcinoma and adult T cell leukaemia [47].

A transfer of the CD8+ T-cells into patients occurs through the variety of approaches including whole tumor cells, MHC-specific peptides, whole or partial proteins encoded by RNA or DNA, or in recombinant viral or bacterial vectors expressed in dendritic cells. Attempts to implement such vaccination strategies have produced inconsistent results in large part due to the immunosuppression by TME [43, 45]. The most effective therapies will ideally combine personalised tumor vaccinations with checkpoint blockade therapies, chemotherapy and approaches targeting proteins involved in tumor growth such as mitogen activated protein kinase (MAPK) pathways to achieve the best patient outcomes.

In addition, to ATT approaches, inoculations of patients with peptide based vaccines comprised of prospective neo-eptiopes that could activate CD8+ T-cells, however so far the success of such vaccinations have been modest due to insufficient activation of CD8+ T-cells and infiltration into the tumors [47].

The other strategy to target tumors with CD8+ T-cells is an adoptive transfer of CD8+ T-cells expressing genetically modified receptors (chimeric antigen receptors (CARs)) [33, 48]. CARs are synthetic receptors made by fusion of single-chain variable fragment (scFv) of an antigen-specific immunoglobulin with an intracellular signal- ling domain. Once these constructs were assembled and introduced in the T-cells, they are expanded. CARs can be engineered to recognise a variety of cancer molecules, including antigens, carbohydrates and glyolipids. Co-stimulatory signalling domains derived from CD28 are also commonly inserted to enhance the CD8+ T-cell activation [33]. In addition to that, domains for cytokines receptors to IL-12 and IL-18 are inserted into CARs to further boost their effectiveness. Despite some initial promising results, there were also significant downsides such as induction of controllable immune responses (cytokine storms) and an inconsistent outcomes for different tumors and patients.

2.3 Immunological tolerance

It's crucial that the immune system due to it's large destructive potential, is strictly controlled to avoid damage to the host's own cells and tissues. These control systems are called self-tolerance. Both the innate and the adaptive immunity have systems in place to prevent such autoimmune responses. Particularly, this control is important for CD8+ T-cells which recognise antigenic peptides derived from the proteomes of intracellular pathogens because they have to be able to distinguish the peptides derived from self proteome from non-self [31]. The selection of T-cells happens in a thymus - a specialised immune organ. The thymus possesses three distinct compartments. In a subcapsular zone, the emerging thymocytes rearrange their beta chains. Next, they move to a cortex of the thymus where the alpha chain rearrangement happens which results in the double positive alpha-beta-TCR. It is checked whether this emerging TCRs possess sufficient binding affinity to self-peptide-MHCI molecules in a process of positive selection (Figure 5). If T-cells don't pass this step, they undergo apoptosis and are removed by thymus resident macrophages [31]. The process of positive selection is directed by cortical Thymic Epithelial Cells - cTECs. cTECs present self antigenic peptides to emerging T-cells via both MHCI and MHCII molecules. cTECs express unique beta5 subunit - beta5t and proteasomes that incorporate it are called thymoproteasomes [37, 49]. It was shown that mice that don't have this subunit aren't capable of completing the positive selection of thymocytes properly. During this step, the T-cells that respond to self-peptides only weakly are given the preference [49]. Following that, the T-cells proceed to a medulla of the thymus where the T-cells are surveyed for a potential autoreactivity, The major driver of the negative selection are specialised APCs - meduallry Thymic Epithelial Cells (mTECs). mTECs express a vast array of tissue-specific proteins which is driven by a transcription factor AIRE (autoimmune regulate). Interestingly, it was shown that despite the fact that mTECs express the majority of human proteome and represent it on their cell surface, there are subpopulations of mTECs each of which only express a small fraction of all antigens (1-3%) [31]. In addition, self antigens are presented by thymic dendritic cells, although they are not a primary vehicle for antigen presebtation [31]. If T-cells recognise peptides presented by mTECs with high binding affinities, they are removed by apoptosis, made unresponsive or converted into Tregs [49] (Figure 5). The cells that pass the both steps, go on into circulation. The majority of T-cells (95%) are removed during the positive and negative selection.

The defects in the immunological tolerance can lead to unprovoked tissue inflammation and autoimmunity. The primary cause of this are thought to be the defects in T-cell differentiation and/or inappropriate activation [31]. These defects can affect both T-helper cells and cytotoxic T lymphocytes. The absence of T regulatory cells was suggested to be a major factor in the emergence of the autoimmune diseases. During negative selection of thymocytes in the thymes there can be a fraction of self proteome that isn't represented during central tolerance induction. This may be explained by low expression of some self antigens in mTECs which would decrease the probability of their presentation. As a result,

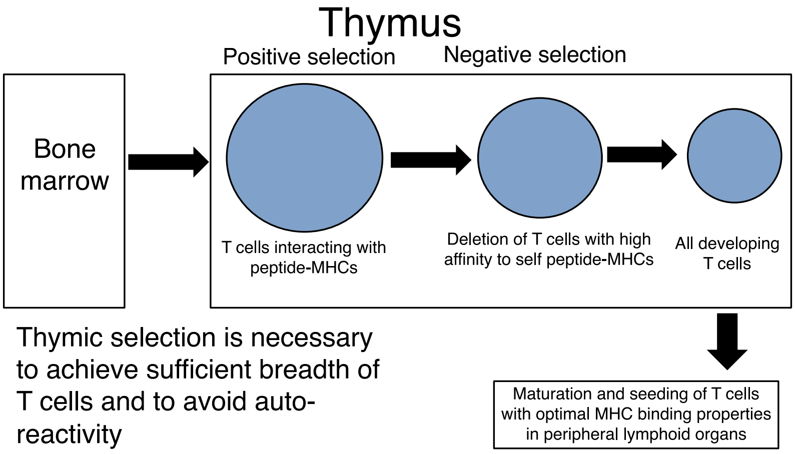

Thymus
Positive selection
Negative selection
Bone marrow
T cells interacting with peptide-MHCs
Deletion of T cells with high affinity to self peptide-MHCs
All developing T cells
Thymic selection is necessary to achieve sufficient breadth of T cells and to avoid auto-reactivity
Maturation and seeding of T cells with optimal MHC binding properties in peripheral lymphoid organs

FIGURE 5: The schematic representation of the thymic selection of T-lymphocytes

some epitopes may only be present in peripheral tissues but not in the thymus. In addition, extensive PTMs and peptide splicing could affect the recognition of such epitopes by T-cells that didn't encounter such versions of epitopes in the thymus [49].

Due to the blind spots in the central tolerance it was crucial to evolve an additional level of selection that is applied to the circulating T-lymphocytes. It ensures that the self-reactive T-cells that could be non-self. This can be done by quiescence, ignorance, anergy, and tolerance-induced cell death which are triggered upon different levels of activation of effector T-cells from naive to effector states [50]. In quiescence naive T-cells are held at low metabolic rates and are kept at G0 stage of the cell cycle which prevents their clonal expansion. There is a threshold for activation of T-cells kept in quiescence which is defined by the binding affinity strength. Ignorance refers to the failure of activation of self-reactive naive T-cells even when a specific self-epitope is present and is thought to be determined by an abundance of the self-epitope (*i.e.* being presented at low levels) and a localisation to specific tissues. During organ specific inflammation (*e.g.* in beta islets during T1D) or viral infections (*e.g.* with Epstein-Barr Virus (EBV) during MS) the ignorance may be overruled and the self-reactive T-cells are activated anyways leading to autoimmunity. When the T-cells are primed, a next stage of non-deletional tolerance is activated, known as anergy. If a given self-reactive T-cells is continuously stimulated in absence of co-stimulation (*e.g.* via CD28), this may render such T-cells hypo-responsive. In absence of self-reactive epitopes T-cells can become responsive again. When T-cells enter effector stage, the next line of defence is exhaustion. If an antigen stimulation continues for an extended period, T-cells become 'exhausted' and exhibit truncated responses to the antigens. Unlike anergy, during exhaustion stimulation the necessary co-stimulatory signals are present. Another tolerance mechanism activated during effector stage of T-cells is senescence which leads to to an inability of the chronically stimulated T-cells to proliferate and replicate. This state is actively maintained. Finally, self-reactive T-cells can be pruned in the periphery via the induction of programmed cell death during effector stage upon repeated re-stimulation. In addition, self-reactive T-cells can be reigned in by other cell populations such as regulatory T-cells [50].

Presence of self-reactive CD8+ T-cell however doesn't guarantee an autoimmune response. On the other hand, the presence of tissue-infiltrating lymphocytes is strongly correlated with self-reactive responses. For example, islet-destructing naive CD8+ T-cells self-reactive to ZnT8 antigen implicated in Type 1 Diabetes (T1D) were identified in both healthy and diseased subjects in similar frequencies in blood [51]. What distinguished T1D patients, were self-reactive CD8+ T-cells present in the pancreas. Thus, active immune-mediated destruction was hypothesised to be caused by defective peripheral tolerance mechanisms. Interestingly, ZnT8 has a low expression level in mTECs landing credence to the notion that antigen's expression level during negative selection is associated with a likelihood of T-cells being tolerized in the thymus. On the other hand, other beta-islet cell associated antigens such as preproinsulin (PPI), insulinoma-associated protein (IA-2), lucose-6-phosphatase catalytic subunit-related protein (IGRP) and glutamic acid decarboxylase (GAD65) are expressed in mTECs, but that didn't seem to affect frequencies of circulating CD8+ T-cells specific to these antigens. This suggets on the contrary a limited role of antigens expression in mTECs in determining the frequencies of circulating CD8+ T-cells. Other factors such as efficiency of presentation of a given epitope and a probability of encountering such epitope by TCRs also play crucial roles in determining a repertoire of T-cells [51].

In the context of the immune tolerance, it has long being postulated that pathogens such as viruses and bacteria could exploit the severely curtailed repertoire of circulating CD8+ T-cell via molecular mimicry of the antigenic peptides derived from the proteomes of viruses and bacteria to self antigens, against which CD8+ T-cells were already tolersed. It is also hypothesised to be the major factor in the emergence of a pathogen induced autoimmunity. This is due to the fact that a number of unique peptides of length 9 AAs is restricted, many of which could be shared by proteomes of viruses, bacteria and human [25, 52]. This is relevant because the majority of peptides presented by MHCI molecules tend to be of length 9 AAs and thus could be shared between viruses, bacteria and human. This opens prospect for both immune evasion or the autoimmunity caused by a stray pathogen derived epitopes that could activate previously dormant self-reactive T-cells. The matter is further complicated by peptide splicing, during which, two distant, non-continuous peptide fragments into the novel peptide sequence not encountered in an original protein [27]. This could potentially substantially increase the possibilities for the molecular mimicry. The additional layer to this complex topic is TCR degeneracy, meaning that one TCR could in principle recognise multiple different epitopes [25, 52].

2.4 Ubiquitin-proteasome system and Proteasome

Ubiquitin-proteasome system (UPS) is a crucial component of a cellular homeostasis that provides a vehicle for cellular proteins metabolism and turnover. Protein level regulation is necessary for normal cellular functions, such as cell cycle progression and cellular differentiation [3, 4] The primary function of UPS is the degradation of short-lived, regulatory, redundant, damaged and mis-folder proteins. There are many sources of protein damage such as ultraviolet and ionising radiation. Internally, the proteins can be damaged by reactive oxygen species generated during metabolism or immune responses [53]. The damage may also occur due to errors during transcription, translation or incorrect post-translational modifications (PTMs) [53]. The degradation of proteins and polypeptides occurs in 3 basic steps: attachment of several ubiquitin molecules to the degraded proteins, ATP dependent unfolding of the protein/polypeptide by 26S proteasome regulatory subunits and it's degradation by the core catalytically active subunits of 26S proteasome [3]. Ubiquitin is a 76 amino acid (AA) long polypeptide that is attached to the degraded protein in 3 primary steps: activation of the C-terminal residue of ubiquitin by E1 enzyme, transfer of the activated ubiquitin from E1 to E2 ubiquitin-conjugating enzymes and to protein/polypeptide also bound to E3 ligase and finally a covalent binding of the activated ubiquitin to the protein/polypeptide from E2 by E3. These steps are repeated to form a polyubiquitin chain [3, 54] (Figure 6). Following that process, the polyubiquitinated protein/polypeptide (substrate) is unfolded and degraded by 26S proteasome to short peptides between 3 and 22 amino acids in length and eventually to the individual amino acids by the cytosolic peptidases [55]. Meanwhile E2 and E3 ligases are released from the substrate and the ubiquitin moieties are released from the substrate by deubiquitinating enzymes (DUBs). Interestingly, UPS is involved in recycling of not just cytoplasmic but also nuclear and membrane proteins. In addition, UPS also has some non-proteolytic functions such as kinases activation. The defects in the UPS degradation can result in a variety of ailments such as cancers and neurodegenerative diseases [55].

The key element of UPS is a 26S proteasome - a large multi-subunit enzymatic complex containing several catalytic subunits [1, 2, 5].It is comprised of roughly 700 kDa 20S particle, which performs proteolytic reactions, and regulatory subunits such as 900 kDa 19S complex, cytosolic PA28alphabeta or nuclear PA28gamma, which secures the tagged substrates, unfolds them in an ATP dependent manner and feeds them into the 20S particle [4, 5, 56, 57] (Figure 7). The 19S unit is composed of the two parts - base and lid. The base consists of six ATPases and 2 subunits that bind to 20S particle and the lid which is composed of the 8 non ATPase subunits. These ATPases of the base are responsible for a substrate unfolding for sending a substrate into proteasome channel [5, 55]. The regulatory subunits typically associate with the 20S particle on both ends. In addition, 26S proteasomes are known to harbour C-terminal hydrolases that have a role in processing of ubiquitin [3]. The association of the 20S proteasome appears to be dependent on a cellular environment. For example, PA28gammabeta regulatory subunit is upregulated in immune associated tissues

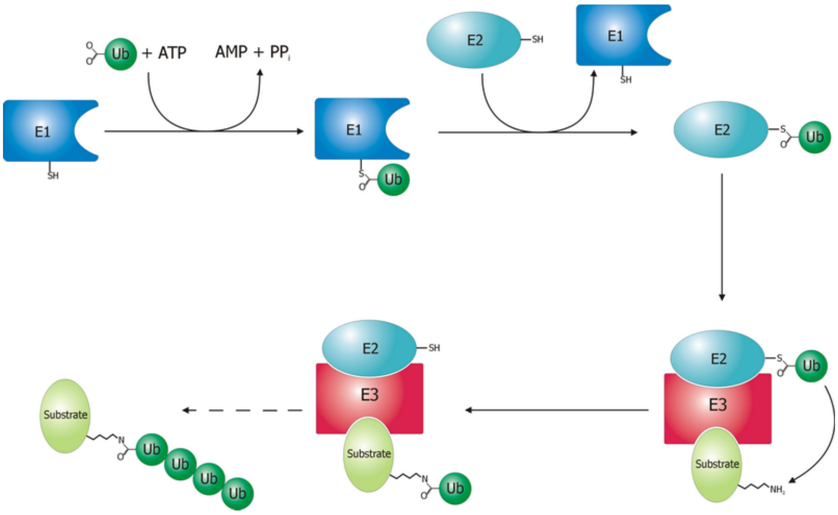
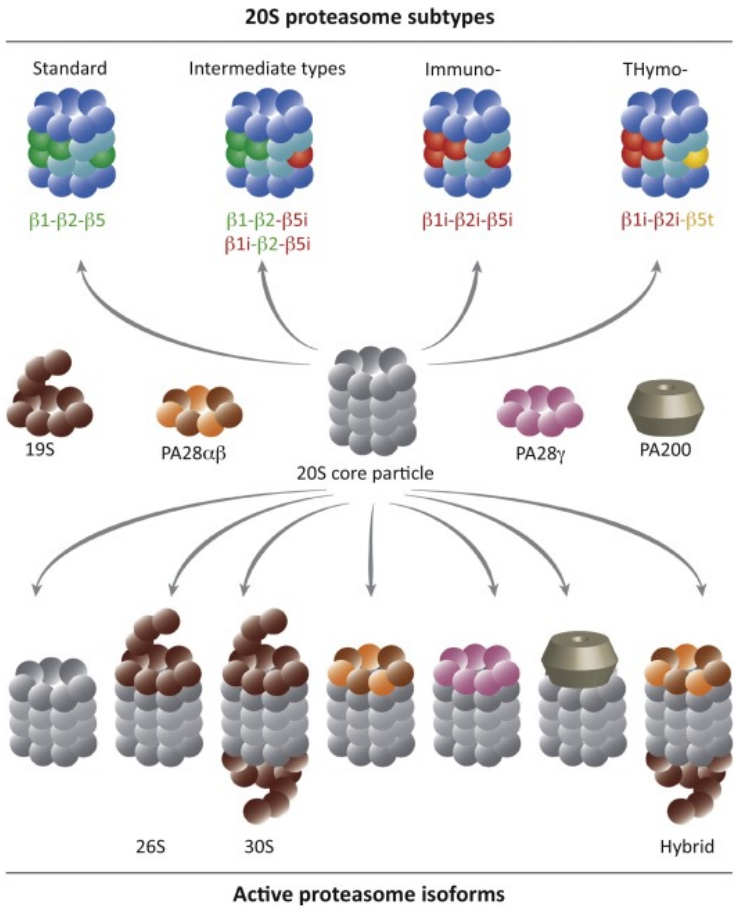


FIGURE 6: The schematic representation of the polyubiquitination
The figure was retrieved from https://en.wikipedia.org/wiki/Ubiquitin

and typically forms complexes with immunoproteasome. PA28gamma binds and activates 20S proteasome in a nucleus.

20S particle is a central part of 26S proteasome responsible for the catalytic activities. It is a hollow barrel shaped structure which is comprised of four stacked heptameric rings - 28 subunits in total [4] (Figure 8). Each ring is made up of seven subunits. Outer two rings are made from alpha subunits while internal two rings are composed of beta subunits and are responsible for the catalytic activity of the proteasome. The alpha subunits are necessary to maintain the structure of the 20S proteasome and to form a gate through which the substrates marked for degradation travel to a catalytically active subunit. Proteasome in turn performs proteolysis by three catalytically active beta subunits - beta1, beta2 and beta5. The catalytically inactive beta subunits are in turn thought to be important for the formation of the substrate binding channels. The substrate is bound N-terminally and C-terminally to a catalytically active threonine 1 of the proteasome in the specialised non-primed and primed binding sites, respectively. Amino terminal threonine (Thr1) in their active sites acts as a catalytic nucleophile and catalyses a breakage of a peptide bond between two amino acids which results in a formation of an acyl-enzyme intermediate between the N-terminal part of the substrate and the catalytic Thr1 residue of the proteasome while a C-terminal part is released. The process is completed by a nucleophilic attack by a molecule of water on an acyl-enzyme intermediate and a subsequent release of the N-terminal piece [1, 57]. While different beta subunits can in principle cleave after every amino acid residue in the substrate, they are known to possess certain specificities. That is, beta5 subunit typically cleaves after hydrophobic residues, beta2 - after basic residues and beta1 - after acidic residues [1, 4]. It's important to point out that despite these preferences, each beta subunit can, in principle, cleave after any amino acid and there is a certain overlapping specificity of subunits [55]. The



Trends in Immunology

FIGURE 7: The different proteasome isoforms and regulatory subunits

The figure was retrieved from [7] upon authors authorisation

specificity of the different subunits is thought in large part to be driven by a structure of substrate binding primed and non-primed sites. This is manifested in an affinity of a distinct subset of substrate sequences (sequence motifs) for the non-primed and primed binding sites. In fact, amino acids located next or distant to a cleavage position (flanking sequences) can be important determining factors in the outcome of the reactions. The time spent in the bound and stabilized state is thought to be a crucial determinant of the outcome of the proteolytic reaction [57]. The hydrophobic and polar interactions between the substrate binding site and substrate residues is an important auxiliary factor in the establishment of a substrate-proteasome interaction. The degradation of the substrates occurs in a processive manner and the efficiency of this degradation is thought to depend on an amino acid composition of the substrate, it's abundance, an interaction of the substrate with the substrate binding sites of the proteasome and a rate of translocation of the substrate and products in and out of the proteasome [56]. Sharon *et al* showed that an occupancy of the catalytic chambers of the proteasomes was a function of a length of the substrate, but typically all of the catalytically active subunits were occupied. In addition, if the rate of hydrolysis was slower than the passage of substrates through proteasomal chambers there was a potential to store the substrates in antachambers for their continual degradation [56]. It was demonstrated that the movement of the substrates along the proteasome chamber into the catalytic site and the overall accessibility of the proteasome for the substrates are the important limiting factors in the proteasome's digestion efficiency [57].

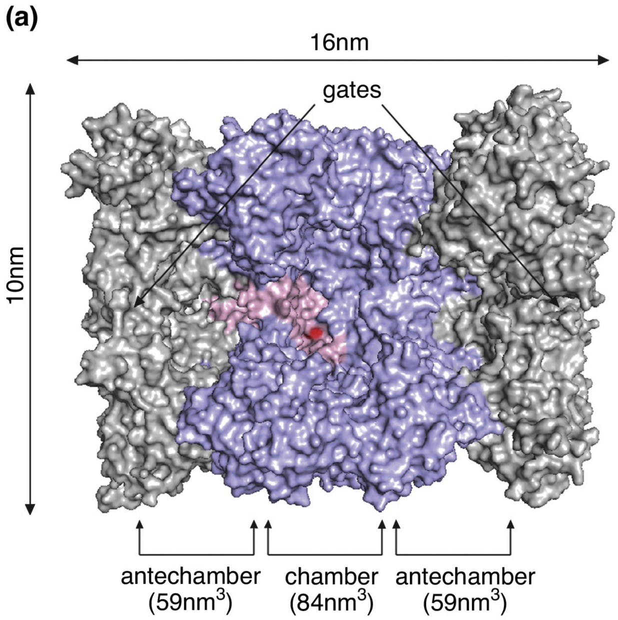


FIGURE 8: The crystal structure of the proteasome. Gray color corresponds to the alpha subunits, while blue - to beta subunits
The figure was retrieved from [58] upon authors authorisation

Moreover, there is a significant degree of variation in the catalytically active beta subunit composition of 20S proteasome depending on the type of tissue where it resides (Figure 7). The most basic variety of proteasome is a standard 20S proteasome containing conventional beta1, beta2 and beta5 subunits. The second most common type of proteasome in the 20S immunoproteasome which is encountered in immune and lymphatic tissues. Primarily it's expressed in cells such as T cells, B cells, monocytes, macrophages, dendritic cells [59]. 20S immunoproteasome is upregulated during tissue inflammation and viral/bacterial/fungal infections. In this isoform, the regular beta subunits are replaced with LMP2 (b1i), MECL-1 (b2i) and LMP7 (b5i) upon IFN-gamma and TNF-alpha induction [4, 25, 47, 57, 59, 60]. The primary functions of this proteasome are thought to be the removal of damaged proteins similarly to a standard proteasome and a generation of antigenic peptides for MHC class I antigen presentation pathway (APP), although the standard 20S proteasome is also capable of that. The immunoproteasomes however are thought to be somewhat more efficient in terms of the production of a hight affinity MHCI peptides [55]. This is not uniformly true and there are multiple examples of antigenic peptides more efficiently generated by a standard proteasomes [55]. An interesting finding showed that the immunoprotesome is uniquely well suited for the degradation of polyubiquitinated protein conjugates more efficient than standard proteasomes, however the data so far has been contradictory and requires additional verification [59]. In addition to the catalytically active beta subunits, the immunoproteasomes seem to preferentially associate with PA28alphabeta regulary particles which are themselves induced by INF-gamma. This regulatory subunit seem to contribute to a preferential presentation of the antigenic peptides [5]. Interestingly, the malignant cells from hematopoietic non-solid tumors were shown to be enriched in immunoproteasomes [60]. Moreover, immunoproteasomes were shown to be expressed in a range of normal tissues such as intenstinal epithelial cells, colon and liver in absence of continuous secretion of IFN-gamma [60]. It remains to be seen what the purpose of the immunoproteasome in the normal tissues is. Finally, there exists a thymoproteasome. It resides in the thymus and contains a specialized beta5t subunit but otherwise is closely related to the immmunoproteasome [1]. The thymoproteasome is thought to be expressed in cortical thymic epithelial cells and is crucial in the process of thymic positive selection of T lymphocytes as it generates antigenic peptides in the locally residing antigen presenting cells (APCs) [59]. On the contrary, during the process of thymic negative selection driven by medullary thymic epithelial cells (mTECs) the bulk of MHCI presented peptides is thought to be produced by the immunoproteasomes [55]. It was demonstrated in a series of *in vitro* and *in silico* experiments that these diverse proteasome types differ not in the quality of the generated product peptides (*i.e.* unique peptide product sequences, attributed to the specific proteasome types) but rather in the frequency of usage of different cleavage sites owing to their unique substrate specificities which results in a dramatically different quantities of the same peptides [4]. This is hypothesised to be associated with the differing substrate sequence preferences exhibited by different proteasome catalytically active subunits and an overall structural changes that occur in the 20S proteasome [5]. This is critical because when it comes to the MHCI presentation, the quantity of the generated peptide rapidly translates into quality - *i.e.* only the peptides presented in sufficient qualities are capable of triggering immunogenic responses [25].

The activity of proteasome is known to be strongly regulated by post-translational modifications (PTMs) [4]. These include phosphorylation, acetylation, glutathiolation, ubiquitination, oxidation and glycosylation Depending on the location of these PTMs such as phosphorylation, this could inhibit the proteasome's activity or instead increase it [54]. The other important way of regulation is a sub-cellular localisation of the proteasome. The majority of proteasomes reside in the cytoplasm and moreover some are linked to specific cellular components such as cytoskeleton. Proteasomes can also reside in the nucleus and mitochondria. In response to environmental stimuli, the proteasomes can migrate to different compartments. For example, the lack of glucose can lead to the recruitment of the proteasomes to the nucleus [4].

While traditionally, 26S proteasome have been thought to be the primary drivers of the protein processing, 20S proteasome was demonstrated to be degrading certain unfolded proteins and polypeptides [4]. When it comes to the substrate degradation by 20S proteasomes, it's typically partially and completely unfolded proteins as well as proteins with intrinsically disordered regions (IDRs), intrinsically disordered proteins (IDPs) and the proteins subjected to the oxidative damage [4, 55]. 26S and 20S proteasomes were suggested to generate overlapping but nevertheless qualitatively distinct sets of peptide products due to a varying usage of different cleavage sites [61]. It's important to point out that the majority of the mammalian proteasomes are 20S and not 26S, which constitute only 1/3rd of all biologically active proteasomes. In addition, a considerable fraction of mammalian proteins (20%) was shown to be degraded specifically by 20S proteasomes [55]. The entry of the substrate into the 20S proteasome chamber is thought to be regulated by the substrate itself as well as heat shock protein 90 (HSP90) and certain histones [55]

Interestingly, not just 20S but also 26S proteasome was also revealed to be able to perform degradations without the marking of substrates with ubiquitin - *i.e.* in an ubiquitin-independent manner. This is particularly relevant for substrates containing unstructured regions [4, 62]. This enables us and others to study the biochemical mechanisms of the substrates degradation of substrates by the proteasomes (both 20S and 26S) in the close *in vitro* systems containing only the proteasome and a substrate of interest.

2.5 Proteasome catalysed peptide splicing

It was demonstrated that N-terminal residue (P1 primed (P1)) of a previously released non-continuous C-terminal fragment is capable of attacking a C-terminal P1 residue (non-primed) of a N-terminal fragment in a transpeptidation reaction. This results in a formation of a novel peptide sequence not encountered in the original substrate. This novel activity of the proteasome is called Proteasome-Catalysed Peptide Splicing (PCPS) [6, 58] (Figure 9).

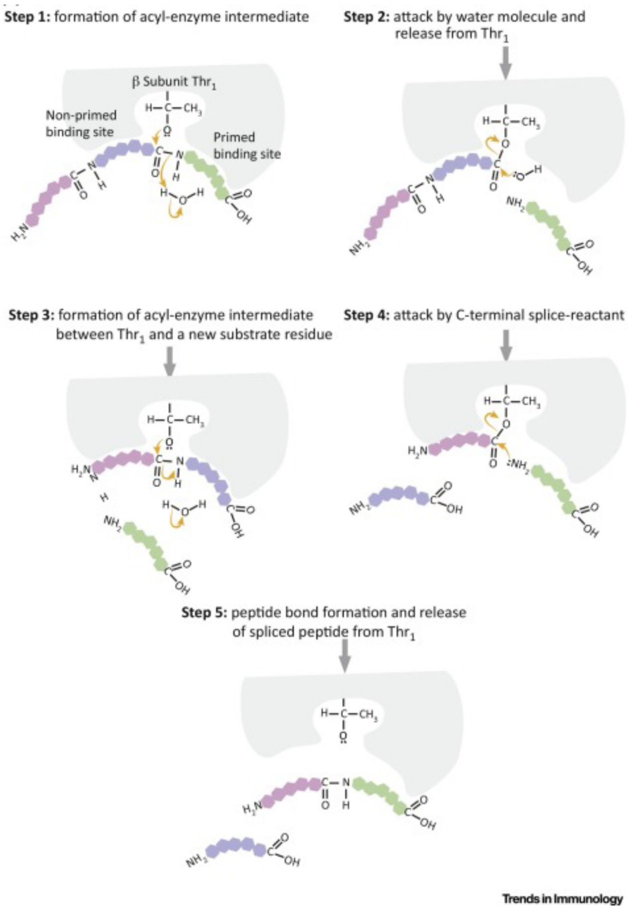

Step 1: formation of acyl-enzyme intermediate
β Subunit Thr$_1$
Non-primed binding site
Primed binding site
Step 2: attack by water molecule and release from Thr$_1$
Step 3: formation of acyl-enzyme intermediate between Thr$_1$ and a new substrate residue
Step 4: attack by C-terminal splice-reactant
Step 5: peptide bond formation and release of spliced peptide from Thr$_1$
H–C–CH$_3$
H$_2$N
NH$_2$
OH
Trends in Immunology

FIGURE 9: The transpeptidation reaction catalysed by the proteasome

The figure was retrieved from [7] upon authors authorisation

The first step in PCSP is a hydrolysis of a peptide bond of a bound polypeptide, that results in a release of a C-terminal fragment, while a N-terminal peptide remains bound to the catalytically active threonine as an acyl-enzyme (acyl-ester) intermediate. This is followed by the nucleophilic attack of the acyl-enzyme intermediate by the N-terminus of the non-continuous C-terminal reactant present in the proteasome's catalytic chamber [55]. This model of the splicing is called the transpeptidation model and is favoured by the proteasome due to a confined space in the catalytic chamber, which prevents a diffusion of the peptide fragments and promotes their accumulation at higher concentrations [9, 27]. The stability of an acyl-enzyme intermediate and a time it spends in bound state, is thought to be crucial for an outcome of a splicing reaction [57]. Mishto M. *et al* suggested that the longer N-terminal fragment stays as the acyl-enzyme intermediate, the more likely it will be that the splicing reaction will occur [27]. In contrast to the non-primed binding site, the identity of the primed binding site remains elusive. Mishto M. *et al* suggested that there exists a binding site for the C-terminal reactant distinct from the primed binding site which speeds up the splicing reaction [27] (Figure 10). Only then will the prospective non-continuous C-terminal fragments be able to compete with molecules of water to complete the splicing reaction. Just as for a regular peptide bond cleavage, the structures and amino acid compositions of the primed and non-primed binding sites in large part determine the sequence preferences of the proteasome for the splicing reactions. Interestingly, it appears that proteasome has very different preferences for sites used for cleavage and used for splicing. In fact, the splice reactants that are utilised for splicing reactions are generated using cleavage sites that are not typically used for regular peptide-bond hydrolysis that results in non-spliced peptide [27]. In addition, a concentration of splice reactants as well as a retention time in the vicinity of the catalytic Thr1 appear to be the important determinants for the success of a splicing reaction [63]. Another important factor in the production of a cis-spliced peptide is a proximity of spliced reactants to each other in the substrate, *i.e.* intervening sequence length [63]. Reducing the length of intervening sequence was shown to increase efficiency of transpeptidation reaction.

Non-spliced (cleavage) (proteasomal cleavage peptides (PCP)) peptides are peptides that are derived from a polypeptide/protein molecule via regular peptide-bond hydrolysis and is considered the primary activity of the proteasome. When it comes to PCPS, there are two distinct types of spliced peptides (proteasomal spliced peptides (PSP)) - cis and trans. In cis splicing two shorter splice-reactants (N-terminal portion and C-terminal portion in a ligated sequence) originate from a same polypeptide and are separate by an intervening sequence which can vary in length. If the ligation of cis-reactants occurs in a normal order (N-terminal splice reactant in the original polypeptide sequence remains as such in ligated peptide) such splicing is called normal cis. If the ligation order is reversed such splicing is known as reverse cis. In trans-splicing two reactants are derived from two separate molecules of the same polypeptide or two distinct polypeptides (Figure 11) [6, 58].

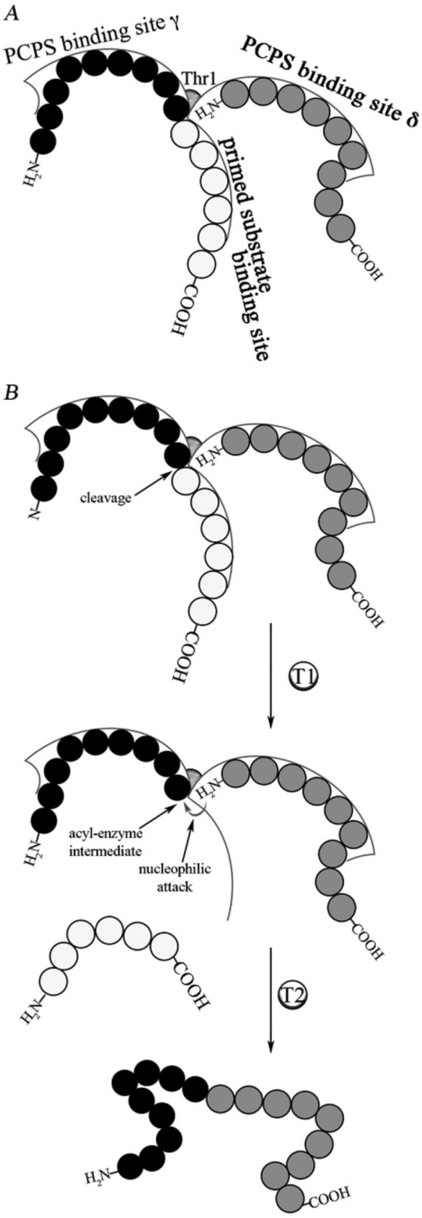


FIGURE 10: The model of PCPS binding sites

According to the model, the N-terminal splice reactant (black circles) binds in the non-primed binding site. On the other hand, the C-terminal splice reactant (grey circles) could be bound either at the primed binding site or at a specialised binding site for C-terminal splice reactants. The binding pockets can fit 5-6 AA long peptides with the N or C-terminal extensions of variable length. During the reaction, the acyl-enzyme intermediate of the catalytically active Thr1 is formed with the N-terminal splice reactant while the intervening sequence (white circles) is released. Following that, the C-terminal splice reactant performs a nucleophilic attack on the acyl-enzyme intermediate which results in the creation of the spliced peptide. The figure was retrieved from [27] upon authors authorisation

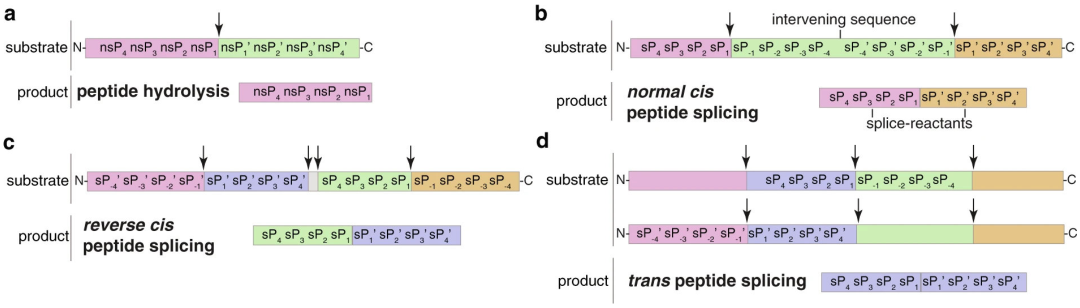

a
substrate
N-
nsP4 nsP3 nsP2 nsP1 nsP1' nsP2' nsP3' nsP4'
-C
product
peptide hydrolysis
nsP4 nsP3 nsP2 nsP1
b
intervening sequence
substrate
N-
sP4 sP3 sP2 sP1 sP-1 sP-2 sP-3 sP-4 sP-4' sP-3' sP-2' sP-1' sP1' sP2' sP3' sP4'
-C
product
normal cis
peptide splicing
sP4 sP3 sP2 sP1 sP1' sP2' sP3' sP4'
splice-reactants
c
substrate
N-
sP-4' sP-3' sP-2' sP-1' sP1' sP2' sP3' sP4' sP4 sP3 sP2 sP1 sP-1 sP-2 sP-3 sP-4
-C
product
reverse cis
peptide splicing
sP4 sP3 sP2 sP1 sP1' sP2' sP3' sP4'
d
substrate
N-
sP4 sP3 sP2 sP1 sP-1 sP-2 sP-3 sP-4
-C
N-
sP-4' sP-3' sP-2' sP-1' sP1' sP2' sP3' sP4'
-C
product
trans peptide splicing
sP4 sP3 sP2 sP1 sP1' sP2' sP3' sP4'

FIGURE 11: The types of proteasome generated spliced peptides
The figure was retrieved from [17] upon authors authorisation

Several groups provided an experimental evidence of splicing reactions catalysed by proteasome. The very first proteasome generated spliced peptide was identified by Hanada *et al* in 2004 [8]. They showed that a 9-mer spliced peptide was derived from Fibroblast Growth Factor-5 (FGF-5), was presented by HLA-A*03:01 molecules and recognised by CD8+ T-cells. To identify it, HLA-A*03:01 expressing COS-7 cells were transfected with plasmids expressing fragments of FGF-5 and analysed an induction of CD8+ T-cells. They then identified a minimal 60 AA long fragment processing of which could induce CD8+ T-cells. None of the linear 8,9 or 10-mers from that fragment could stimulate CD8+ T-cell responses. By analysing potential 9-mer and 10-mer peptides from the narrowed down 49-mer fragment they identified a 9-mer peptide - NTYASPRFK which could induce strong immunogenic responses. The peptide consisted of five N-terminal amino acids and four C-terminal amino acids demonstrating that an intervening sequence length of the peptide was 40 amino acids. RNA-splicing and ribosome skipping couldn't explain the generation of this peptide. The synthetic peptide was pulsed into HLA-A*03:01 positive and negative EBV transformed B cells. The activation of CD8+ T-cells only occurred in the HLA-A*03:01 expressing B cells. Moreover, the presentation of the antigenic peptide was blocked by proteasome and TAP inhibitors.

In a follow-up study, Dalet *et al.* confirmed that the antigenic peptide was indeed produced by the proteasome catalysed peptide splicing via an excision of the 40 AA intervening sequence [9]. This further verified the transpeptidation model. Curiously, the decrease in the intervening sequence length seemed to increase the efficiency of the production of the spliced peptides by the proteasome. In the same study they also demonstrated a possibility of trans-splicing *in vitro*. They incubated the 49-mer precursor peptide with 20S proteasome and loaded the target cells with the digest. These cells were capable of stimulating CD8+ T-cell clones. The production of the spliced peptide was confirmed by a tandem-Mass-spectrometry (MS/MS). However, the efficiency of the splicing reaction was low judging by the kinetic of the peptide generation. The fragment NTYAS formed the acyl-enzyme intermediate with the catalytically active Thr1 while the PRFK fragment would perform the nucleophilic attack. The cleavage of the C-terminal peptide bond of NTYAS - *i.e.* the release of the intervening sequence was crucial for the reaction. The blockage of the N-terminal group of PRFK fragment also stopped the reaction. By shortening the intervening sequence length, the efficiency of the reaction increased dramatically. It's known that the inner channel of the proteasome has a diameter of 13 angstrom which means that it can accommodate more than one polypeptide. In addition, the proteasome could bind polypeptide in each of its catalytic chamber simultaneously. The trans-splicing was demonstrated by transfecting full length FGF-5 proteins each of which would contain the mutation in either N-terminal or C-terminal spliced reactant into COS-7 cells expressing HLA-A*03:01. The only way, the correct antigenic peptide could be produced was by ligation of fragments from two separate proteins with two different point mutations. Indeed the simultaneous transfection resulted in a stimulation of the CD8+ T-cells. Moreover, the *in vitro* digestions of the two mutated precursors by 20S proteasome confirmed the generation of the trans-spliced peptide. Nevertheless, the efficiency of such reaction was lower compared to cis-splicing.

Vingeron *et al.* in characterised a fusion peptide assembled from two noncontiguous fragments of melanocytic glycoprotein gp100(PMEL17) presented by the melanoma cells [10]. The generation of the peptide was confirmed *in vitro* by incubation of the 13-mer precursor peptide with purified 20S proteasomes. It was the first study that showed that the reaction occurs by transpeptidation involving an acyl-enzyme intermediate. First, they isolated the CD8+ T-cell clones from the melanoma patient that recognized an antigen presented by HLA-A*32. They then transfected HLA-A*32 expressing COS-7 cells with plasmids encoding variable fragments from gp100. The observed that following the transfection of the 13-mer peptide RTKAWNRQLYPEW covering position 40 to 52 of gp100 after into EBV transformed B cells, these cells could stimulate CD8+ T-cells. Following that, they synthesised peptides that could be obtained from the 13-mer and revealed tha residues located at the N- and C-terminus of the 13-mer were crucial for the immunogenic responses. By deleting the middle positions of the 13-mer (43-46), they identified a 9-mer peptide RTKQLYPEW that could stimulate CD8+ T-cells, suggesting that RTK and QLYPEW were spliced following the removal of the 4 amino acid long intervening sequence length. They then eluted the peptides from HLA-A*32 molecules of the tumor cells and separated them by HPLC. The fractions of the peptide that stimulated CD8+ T-cells were the same as the synthetic 9-mer passed through HPLC under the same conditions. An application of proteasome inhibitors abrogated the recognition of the target cells by CTLs. The 20S proteasomal digests of the 13-mer peptide could strongly stimulate the CD8+ T-cells. The digests were passed through HPLC and the same fraction as the synthetic peptide could again stimulate CTLs. The subsequent analysis with MS/MS confirmed the identity of the peptide. They proposed a model, in which following intervening sequence hydrolysis, RTK formed and acyl-enzyme intermediate with the catalytically active Thr1 which then interacts with QLYPEW fragment present in the proteasome to form the final peptide. The energy necessary to form the new peptide bond was generated by the cleavage of the intervening sequence fragment WNR. Additionally, the identified one more spliced peptide derived from the 13-mer - RTKAQLYPEW.

Warren *et al*. characterised a spliced peptide generated by reverse cis-splicing derived from SP110 protein and produced by 20S proteasome [11]. They isolated CD8+ T-cells from cancer patient receiving hematopoietic cell transplantation (HCT). One of the CD8+ T-cell clones was found to recognise a peptide presented by HLA-A*03:01 molecules. They generated cDNA library of possible peptides contained in 60 AA polypeptide containing a stretch of sequence of SP110 and screened the produced peptides for their capacity to induce the isolated CD8+ T-cells and identified a peptide containing A to G substitution in SP110 protein. They then synthesised a polypeptide covering 20 AA stretch of the 60 AA polypeptide and transfected into COS-7 cells. These cells were recognised by the isolated CD8+ T-cell clones. Moreover, both ends of the 20 AA polypeptide were required to induce the immunogenic response. They then generated a series of non-overlapping short peptides derived from the 20-mer and identified two non-continuous peptide fragments which together formed a fusion antigenic peptide. They later managed to identify minimal fragments required to induce the stimulation of CD8+ T-cells - SLPRGTAS and STPK, forming SLPRGT-STPK. Moreover the peptide could only bind to HLA-A*03:01 when the fragments were ligated in a reverse order. Finally, they eluted the peptides presented by HLA-A*03:01 positive EBV transformed B cells and performed high- performance liquid chromatography (HPLC). The fraction of peptides that stimulated CD8+ T-cells was the same as a chromatographed synthetic peptide, confirming its identity. Moreover, they performed *in vitro* proteasomal digests of a synthetic 20-mer peptide by 20S proteasome. The target peptide was identified by the tandem-Mass-spectrometry(MS/MS) and the digests could stimulate CD8+ T-cells. The STPK portion of the peptide was thus first liberated by the peptide bond hydrolysis while SLPRGT formed acyl-enzyme intermediate. N-terminal group of STPK then attacked the acyl-enzyme intermediate formed by SLPRGT. The cleavage of the intervening sequence from SLPRGTASSR - ASSR was necessary to generate a sufficient energy required to finalise the reaction [11].

Dalet *et al*. described a production of an antigenic spliced peptide from tyrosinase that was recognised by TILs from melanoma patient [12]. One of the CD8+ T-cell clones from the melanoma patient was found to recognise an antigen presented by HLA-A*24:02 molecules. The authors identified a 1-377 long sequence from tyrosinase transfection of which into COS-7 led to CD8+ T-cell stimulation. Interestingly, the sequence included a signal peptide necessary for a translocation into ER due to a requirement for ER associated degradation (ERAD) of substrates. If the cells were transfected without the signal peptide, they failed to elicit the immune response. They further showed that Asn deamidation to Asp at positions 337 and 371 was necessary for the stimulation of CTLs. They then chemically synthesised 180 different peptide sequences covering positions from the the tyrosinase regions 332-342 and 367-377 containing two Asp substitutions that were joined in the reverse order. Two of the product peptides - IYMDG-TADFSF and IYMDGAADFSF could be efficiently recognised by CD8+ T-cells. The authors then eluted MHCI bound peptides from the eluted cells that were recognised by CD8+ T-cells and fractionated them by HPLC. The eluted peaks corresponded to the synthetic peptide IYMDGTADFSF. A normal cis-spliced version of the peptide ADFSFIYMDGT failed to elicit the CTL response, on the other hand. Finally, the identity o the peptide was confirmed by MS/MS. Treatment of the melanoma cells with proteasome inhibitors significantly decreased the recognition of the cells by CD8+ T-cells. A 25 AAs precursor peptide was than incubated with 20S standard and immuno-proteasome. The digests were loaded onto the the tumor cells and only the digests by 20S standard proteasome could elicit the responses. Moreover, the antigenic peptide of interest was only identified in the 20S standard proteasomal digestions. Both a removal of the intervening sequence of the N-terminal spliced reactant and a blockage of the N-terminus of the C-terminal spliced reactant failed to generate the spliced peptides. Translation of tyrosinase into the ER was required for the generation of the spliced peptide which was followed by the proteasomal degradation. The conversion of Asn to Asp via deamidation was necessary for the peptide to elicit immunogenic responses. In addition, the presentation of an antigenic peptide was blocked by the administration of TAP inhibiton.

Michaux *et al.* identified a spliced antigenic 9-mer peptide derived from gp100 and presented by HLA-A*03:01 molecules [64]. Interestingly, in contrast to a previously described peptides, that peptide was produced by a ligation of the 8 amino acid long fragment with a single arginine residue. First, CD8+ T-cells recognising peptides presented by HLA-A*03:01 expressing melanoma cells were identified. In order to identify a region of gp100 that included the N- and C-terminal fragments of the peptides, the authors transfected vectors encoding different fragments of gp100 into COS-7 cells expressing HLA-A*03:01. The peptide was located in a 28 AA long segment of gp100 covering residues 184 to 211. Linear peptides from that regions failed to elicit any CD8+ T-cell responses. They then synthesised 96 spliced peptides that could cover that 28 AA long region, that could be generated by reverse cis-splicing and split them into 15 groups. The groups that could elicit the immune responses were fractionated by HPLC and the fractions were surveyed for their ability to induce CTL responses. A most potent peptide was RSYVPLAHR that was formed by reverse-splicing of RSYVPLAH to single R. They then eluted the peptides from HLA-A*03:01 expressing melanoma cells, fractionated the peptides and compared the fractions with the one corresponding to the peptide of interest. They manage to find one such MHCI eluted fraction which corresponded retention time (RT) wise to the peptide of interest. Nevertheless, the peptide splicing occurred through proteasome by transpeptidation. However, the peptide performing nucleophilic attack on the acyl-enzyme intermediate has to be at least 3 amino acid long suggesting that C-terminus of the generated peptide had to be additionally trimmed by the proteasome to produce the final peptide. To check this, they digested a precursor peptide RSYVPLAHSSSAFT, containing both fragments with 20S proteasome which didn't generate the product peptide. On the other hand, incubation of RSYVPLAHSSSAFT with short fragments (3 and 4 AA long) containing R could on the other hand generate C-terminally extended precursor of the RSYVPLAHR peptide. This drove the authors conjecture that a C-terminal spliced reactant has to be at least 3 AA long. They also showed that the proteasome could subsequently perform trimming of the C-terminally extended precursor that results in the peptide, which was characterised.

In general, trans-spliced peptides were demonstrated to be generated in different contexts both *in vitro* and *in cellulo* [17, 27, 63]. *in cellulo* this was demonstrated to occur with low efficiency by co-transfecting the cell with plasmids encoding two parental proteins from which two splice reactants would be derived and then testing for an ability to elicit CTL responses [9].

In 2015, Ebstein *et al* demonstrated a production of spliced antigenic peptides from gp100 via both transpeptidation and condensation reactions [2]. Gp100 derived 13-mer was digested *in vitro* by human 20S proteasomes. Three spliced peptides were identified - RTKQLYPEW, QLYPEWRTKAWNR and QLYPEWRTK. The analysis was focused on the reverse cis-spliced peptide QLYPEWRTK due to its high predicted binding affinity values for HLA-A*03:01 and HLA-A*01:01. Interestingly, RTKQLYPEW formed by normal cis-splicing was generated less efficiently that the reverse-cis spliced peptide. Due to the fact that QLYPEW is the N-terminal spliced reactant which ends in Trp - the C-terminus of the original 13-mer, the authors suggested that the peptide could be generated by the condensation reaction as opposed to transpeptidation. To test this, the two splice-reactants were incubated with 20S standard and immunoproteasomes. Despite the absence of the intervening sequence length, the proteasome could ligate the two fragments, which was done more efficiently by the immunoproteasome. Moreover, the reaction was only efficiently conducted by the beta5 subunit of both 20S i- and s- proteasomes. The condensation reaction was also demonstrated to take place not just *in vitro* but also *in cellulo*. The peptide was moreover, presented by the melanoma cells and recognised by CD8+ T-cells specific to this reverse-cis spliced peptide.

It's important to be able to predict which potential non-spliced and spliced epitope candidates will be produced by proteasome in order to decrease the potential search space. Despite the fact that there is evidence that suggests that PCPS is not a random process but is rather highly controlled by the proteasome, presently, there exists no reliable predictors of cleavage/splicing preferences making such predictions difficult [27]. A few attempts were made to determine the general peptide splicing rules - to pinpoint the P1 and P1' residues and their pairing most likely to be used in splicing reactions [18, 19]. In both cases, it was suggested that the sequence characteristics of the N-terminal splice reactant (SR1) was the major driving force behind splicing reactions further corroborating the results of Mishto M. *et al* [27]. The authors also showed that as Mishto M. *et al* had demonstrated, that spliced peptides display sequence preferences that are distinct from proteasome generated cleavage peptides non-spliced peptides - *i.e.* where proteasome cleaves it doesn't necessarily splices. In both studies it was concluded, despite some differences, that small, hydrophobic and acidic residues were the preferred residues in P1 positions for the splicing reactions. In addition a N-terminal residue of C-terminal splice reactant (P1') and the residues surrounding the P1 and P1' residues were also demonstrated to be important in determining proteasome sequence specificity for splicing. In addition, as a component of the analysis of polypeptides digestions provided in our large database of *in vitro* proteasomal digestions we looked at the amino acid frequencies of peptide products in position P1, we observed that standard 20S proteasomes preferentially cleave after polar uncharged residues while small uncharged amino acids were preferred in splicing reactions [17].

The other important matter regarding the mechanisms of proteasomal catalysis that has been explored in the recent years are the differences in cleavage reactions catalysed by different proteasome isoforms. Mishto *et al*. studied the differences in cleavage site usage between standard and immunoproteasomes by performing *in vitro* digestions of four synthetic polypeptides [65]. They found that despite significant differences in usage of different sites in the polypeptides for hydrolysis among proteasome isoforms there were no residues that were not used by either of the proteasome types. Further analysis showed a difference in usage of different P1 residues for cleavage reactions - particularly higher efficiency of cleavages after hydrophobic residues by immunoproteasomes. They also discovered that these differences in cleavage site usage coupled with unequal substrate degradation rates resulted in marked differences in the production efficiency of a number of potential MHCI epitopes of interest. Finally, it was further corroborated that cleavage-site usage was not only determined by amino acids forming a peptide bond, being cleaved but also by their surrounding residues [65]. This was further shown by Liepe *et al*. by combining *in vitro* proteasomal digestions, Mass Spectrometry and mathematical modelling to demonstrate that proteasomal dynamics, *i.e*. the frequency of peptide-bond cleavage was dependent on the particular sequence motifs [66]. Not only that, but it was also demonstrated that the differences in polypeptide processing could also be attributed to the peptide transport to the catalytic chamber of the proteasome and the differences of the transport regulation between proteasome types [66]. Kucklekorn *et al*. explored digestion dynamics of human 20S thymoproteasome (20t) and compared it to other ptroteasome isoforms - 20s and 20i [67]. Similarly to 20s and 20i proteasome there were considerable differences not just in cleavage site usage of P1 residues but also polypeptide transport to the proteasomes' catalytic chamber between the three isoforms. Despite the overall comparable catalytic activity between the isoforms there were marked differences in hydrolysis efficiency when looking at specific cleavage sites in specific polypeptides which was reflected in the generation of peptide products. The difference in the ability to generate self epitopes between 20t and 20s/20i proteasomes could thus further explain the unique utility of the thymoproteasome in the thymic positive selection during antigen presentation by cortical thymic epithelial cells (cTECs). These collective observations of purely quantitative differences in proteasome catalysis were echoed by Winter B. *et al*. who also looked at the catalytic activity of standard and immuno-proteasomes [68]. Despite having overall similar substrate specificity, there were some considerable differences, in which P1 residues were mostly used by 20s and 20i proteasomes. The immunoproteasome was shown to mostly favour cleavages after hydrophobic residues as also observed by Mishto

M. *et al*. while the standard proteasome was mostly performing hydrolysis after basic residues, also indicating an unequal potency of different beta subunits of proteasome as it's known that beta5 subunits favour peptide bond hydrolysis after hydrophobic residues, beta2 - after basic residues and beta1 - after acidic residues [68]. The differences in hydrolysis subsequently impinged upon quantities of MHCI presented epitopes produced by different proteasomes.

2.6 MHC class I Antigen Presentation Pathway and the potential role of spliced peptides

A crucial aspect of an adaptive immune response is a targeted elimination of the cells infected with pathogens and abnormal cell by activated CD8+ T-cells (cytotoxic T lymphocytes (CTLs)). In addition to a protein turn-over, one of a principle functions of the proteasome is a generation of short peptides derived from proteins (antigens) for the presentation by MHCI molecules for the recognition by receptors of CD8+ T-lymphocytes [55, 57]. Often, these peptide fragments are then trimmed or degraded by cytosolic aminopeptidases or alternatively translocated to an endoplasmic reticulum (ER) [36]. In ER the N-terminally extended precursors are additionally processed by ER-resident N-terminal peptidases. Alternatively, such precursors could be preemptively trimmed in the cytosol by tripeptidyl peptidases (TPPs) [55]. The typical length or MHCI presented peptides is 8-10 AAs, although it can be as long as 15 AAs for some HLA molecules. The binding of the antigenic peptides to the MHC-I binding pocket primarily occurs through peptide anchor residues which are located at a C-terminus of the peptides and near their N-terminus [5]. The MHC-I bound peptides often have hydrophobic or basic residues at a C-terminus [5, 57]. It is known that only a small share of all peptides generated by proteasome, spliced and non-spliced end up being presented on the cell surface [5, 36]. The peptides that are generated can come from a variety of different positions in a substrate including N- and C-terminal residues of the substrate. A majority of the antigenic peptides is derived from the cytosolic proteins - both self and non-self. A large fraction of newly generated proteins (40%) are short lived and rapidly degraded [36]. Some of those nascent proteins underwent errors in the translation or improper PTMs making them targets for the UPS. In this context, it's worth mentioning so called defective ribosomal products (DRiPs) which are rapidly destroyed by the proteasome. DRiPs are proteins that are not properly folded and are rapidly degraded by the proteasomes [35]. The peptides derived from a hydrolysis of these DRiPs could thus be important source of the MHCI peptides. Interestingly, other proteases were implicated in the generation of the antigenic peptides such as tripeptidyl peptidase II and metallopeptidases, although they are not main contributors of antigenic peptides [55]

Following the degradation, the newly generated antigenic peptides are bound to a channel protein transporter associated with antigen processing (TAP) which shuttles the bound peptides from the cytosol to a lumen of ER. If necessary, the peptides are processed by ER resident aminopeptidases (ERAPs) to a preferred length - ERAP1 and ERAP2 [69]. Such peptides are normally processed to the correct C-terminus by the proteasome, however the N-terminus often contains the additional residues that have to be removed, which is why ERAP activity is crucial. An inhibition of ERAPs was shown to significantly alter the composition of MHCI presented peptides [69]. Following that, the peptides are bound to a protein tapasin that assists in a formation of an MHCI-peptide bound complex. TAP stabilises the vacant MHCI molecules and enables the loading of the antigenic peptides into the MHCI cleft [69]. TAP contains two ATP-binding subunits - TAP-1 and TAP-2 [31]. TAP uses two ATP molecules to first perform peptide translocation and then to release peptide in the ER lumen [36]. Moreover, MHCI molecules can bound precursor peptides, in which case they extend past the binding interface of MHCI. Such bound precursors can also be trimmed by ERAPs to a correct length [69]. N-terminally extended precursors could however under certain circumstances elicit immune responses. The loading of the peptide onto MHCI molecule is typically performed by 3-4 copies of tapasin [36]. Tapasin also binds to ER resident chaperones which are necessary for a stabilisation of the MHCI molecules during loading [36]. The loaded MHCI molecules are then translocated via Golgi Apparatus via a standard secretory pathway to the cell surface where the presented peptides are recognised by TCRs of CD8+ T-cells. The sensitivity of different TCRs varies dramatically and as little as 10 molecules of an antigenic peptide displayed by a single APCs could be sufficient to trigger an immune response [55]. It's important to point out that a substantial fraction of the peptides that are generated by the proteasome are filtered out in the subsequent steps either due to length (too short or too long) or sequence incompatibility (Figure 12). After the presentation, the MHCI molecules are internalised and recycled [36]. It is crucial that the antigenic peptides (epitopes) are presented by the antigen presenting cells (APCs) such as dendritic cells and macrophages. This is because they express co-stimulatory molecules and secrete molecules necessary for an activation of the naive CD8+ T-cells. APCs primarily express immuno- and intermediate proteasome. Once CD8+ T-cells are activated, they can respond to the epitopes presented by non-APC cells.

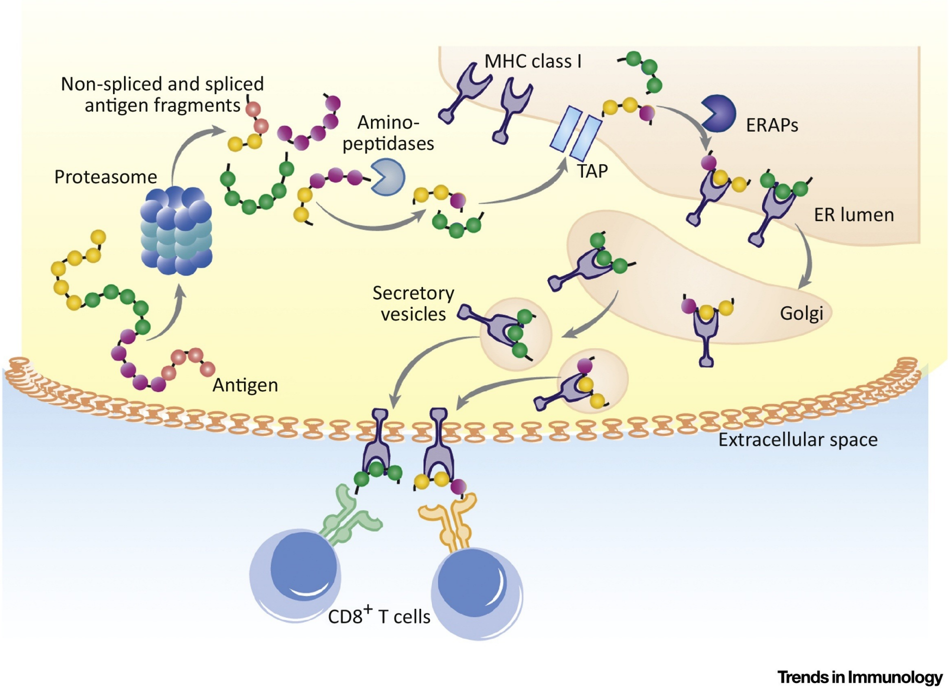


FIGURE 12: The MHC class I antigen presentation pathway
The figure was retrieved from [7] upon authors authorisation

Novel approaches utilizing bioinformatics and Mass Spectrometry (MS) proved that PCPS is much more frequent among MHCI presented peptides, than originally proposed. For example, Liepe *et al* showed that PCPS produced peptides used for MHCI presentation comprises 30% of the diversity and 20% of the abundance of the cells imunno-peptidome suggesting that the spliced peptides play an important role in CTL mediated immune response [13]. Recently, Liepe *et al.* have examined the MHCI immunopeptidome of colon and breast carcinoma cell lines. Using their in-house method for a spliced peptides identification in MHC-I immunopetidomes measured by mass-spectrometry, it was shown that spliced peptides comprise a large fraction of the immunopeptidome of cancer cell lines [14]. According to the analysis, they constituted roughly 23.6% of the variety and 19% of the abundance of the MHC-I immunopeptidomes of the cancer cell lines. In fact, some of the antigens were only represented by spliced peptides. Interestingly, they seem to be derived from so called sequence 'hot-spots' within the antigens [14]. It was also demonstrated that spliced peptides possessed similar sequence motifs to non-spliced peptides, reflecting shared characteristics determining MHCI presentation despite presence of certain sequence differences. In addition, a number of spliced peptides is positively correlated with antigens length, abundance, hydrophobicity and isoelectric point (IP)[14]. Faridi *et al.* applied their in house bioinformatic workflow to characterise HLA-I immunopeptidomes of a variety of monoallelic cell lines and found that trans-spliced peptides are produced and HLA-I presented as frequently as cis-spliced peptides and possess sequence features making them good HLA-I binders [70]. Despite this, their prevalence in the immunopeptidome and the role in the immune response still remains to be elucidated [58]. In a recent study by Faridi *et al.* found that a contribution of cis-spliced peptides to an entire immunopeptidome of melanoma cells was 6-8% [15]. Some of those cis-spliced peptides were demonstrated to trigger specific CD8+ T-cell responses. We have recently published a large database of *in vitro* digestions of a variety of synthetic polypeptides by 20S and 26S proteasomes in which we showed that spliced peptides are produced at frequencies that are larger than previous estimates and qualitatively constitute a significant portion of the digestion products [17].

The spliced epitopes could elicit responses of CD8+ T-cells already primed by non-spliced peptides or alternatively trigger unique CD8+ T-cell populations. For example, phospholipases PlcA and PlcB from Listeria monocytogenes were surveyed for presence of cis-spliced epitopes [20]. Two cis-spliced epitopes derived from PlcB were shown to be produced by 20S proteasome *in vitro* and elicit CD8+ T-cell responses non overlapping with CD8+ T-cells reactive to non-spliced peptides. In another antigen of L. monocytogenes, Listeriolysin O a cis-spliced epitope candidate was identified that by contrast was only recognised by CD8+ T-cells that were originally primed by non-spliced peptides that was partially overlapping with the cis-spliced peptide [71]. Similarly, cis-spliced and non-spliced peptides were identified in HLA-I immunopeptidomes of HIV-I infected cells that were triggering same CD8+ T-cell populations cross-reactive to those peptides [21]. We have recently described a novel spliced peptide derived from a cancer associated antigen KRAS G12V that is produced by 20S proteasome, efficiently binds to HLA-A*02:01 molecule and which structural properties in bound state make it an attractive target for T-cell receptors [72]. Cis-spliced peptide was obtained from melanoma gp100 antigen, demonstrated to be generated both *in vivo* and *in cellulo* and was shown to elicit CD8+ T-cell responses specific to that cis-spliced peptide in peripheral blood of melanoma patients.

In recent years a number of studies have been conducted in which the authors attempted to estimate the frequency of spliced peptides in the immunopeptidome. Depending on a material used, a choice of HLA for the elusion, experimental techniques used and the utilised algorithms for identification of spliced peptides, very different conclusions were made regarding the relative frequency of spliced peptides in the immunopeptidome, which ranged from 1 to 34% [13, 14, 15, 21, 70, 73, 74].

2.7 Type 1 Diabetes

As outlined above, defects in a process of negative selection of thymocyes and/or peripheral tolerance mechanism could lead to autoimmune diseases. One such prominent auto-immune disease is type 1 diabetes (T1D) [34, 75, 76]. The key feature of T1D is T-cells mediated destruction of pancreatic beta cells that produce insulin (Langerhans islands). As the result, there is a severe lack of insulin which leads to a hyperglycaemia due to an overproduction of glucose (Figure 13).

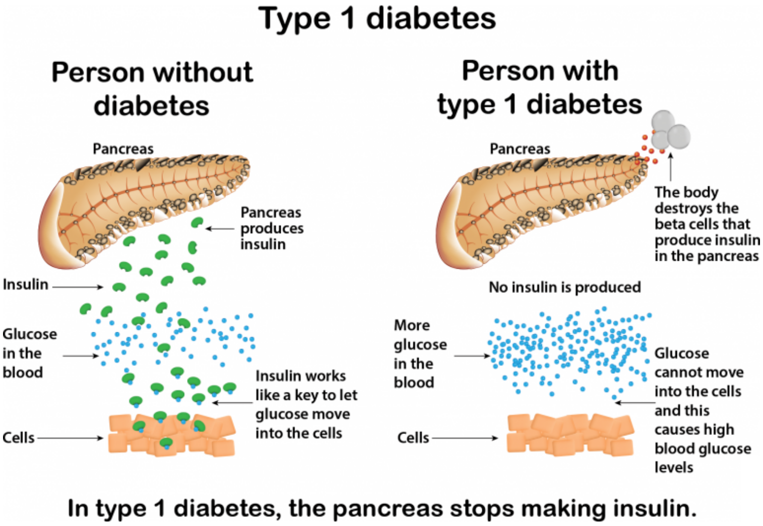

Type 1 diabetes
Person without diabetes
Person with type 1 diabetes
Pancreas
Pancreas produces insulin
Insulin
Glucose in the blood
Insulin works like a key to let glucose move into the cells
Cells
Pancreas
The body destroys the beta cells that produce insulin in the pancreas
No insulin is produced
More glucose in the blood
Glucose cannot move into the cells and this causes high blood glucose levels
Cells
In type 1 diabetes, the pancreas stops making insulin.

FIGURE 13: The difference between healthy and T1D pancreas
The figure was retrieved from https://www.ndss.com.au/living-with-diabetes/about-you/young-people/type-1-diabetes/

In addition, there is an increased rate of fat breakdown and fatty acid oxidation which results in an increased production of ketones [34, 75]. The only effective treatment of T1D is a lifelong insulin injection. T1D typically manifests itself during childhood and adolescence, however the onset can happen later in life [75]. The degree of beta islets destruction varies from patient to patient. One of the pieces of evidence of this includes an isolation of T-cells which are specific for beta cell specific antigens including pre-proinsulin, pro-insulin, insulin, glutamic acid decarboxylase (GAD), protein tyrosine phosphatase (IA-2) and zinc transporter 8 (ZnT8) from patients and their recorded ability to destroy beta islets *in cellulo* [34, 75]. Both CD4+ and CD8+ T-cell were implicated in a development of T1D [75]. One crucial factor that distinguishes healthy individuals from T1D patients is a presence of tissue infiltrating lymphocytes in a pancreas. In addition to T-cells, autoantibodies reactive to pancreatic specific antigens were implicated and could serve as markers of an immune destruction and can arise before the clinical manifestations of the disease. Genetic and environmental factors all determine a probability of the progression of a disease. Studies conducted with identical twins showed that even if there is a genetic susceptibility to T1D, the disease may not necessary develop suggesting crucial contribution of the environment. Nevertheless, the risk of T1D is significantly higher in monozygotic twins compared to dizygotic twins [34]. Environmental factors under the most scrutiny include viral infection, diet, prior vaccinations, toxins and the climate. The environment is thought to trigger the disease progression in at risk individuals. An effect of these factors is also positively correlated with age. Epigentics including DNA methylation and Histone deacetylation was also found to play an important role in determining T1D susceptibility as some of the epigenetic changes were found to enhance the expression of previously suppressed inflammatory response genes [76].

Infections with a number of pathogens, in particular, including rubella virus, rotavirus, mumps virus, cytomegalovirus (CMV), coxsackie B viruses and EBV were shown to be linked with an increased risk of the T1D [34, 75]. It's currently debated whether viruses only enhance a progression of T1D or if they by themselves could trigger its development. It was observed that coxsackievirus and adenovirus possess receptors that are unique to beta-cells [76]. One of a primary mechanisms of the viral effect on the T1D is a molecular mimicry - *i.e.* sequence similarity of the antigenic peptides derived from viral and self antigens [34]. A sequence identity or similarity of the viral antigens to self could trigger previously dormant auto-reactive T-cells that were missed by the central and/or peripheral tolerance [34]. For example, cross-reactivity of T-cells was found between a non-structural P2-C protein of coxsackievirus and a autoantigen GAD65, between a VP1-protein of enteroviruses and a beta cell autoantigen tyrosine phosphatase IA-2 and between GAD65 and human CMV [34]. However the homology between the antigenic peptides and even their immunogenicity don't necessarily lead to any detrimental outcomes.

A number of genetic factors is known to be strongly linked to T1D. For example, certain gene loci such as IDDM1 (the HLA gene region) and IDDM2 (the insulin gene region) were described to be associated with T1D. IDMM1 is the locus that contains genes encoding MHC molecules. It encodes three groups of genes. A first group (class I) encodes alpha chains which bind to beta microglobulin of MHCI. A second group (class II) encodes genes that encode alpha and beta chains that form MHCII molecules - HLA-DR, HLA-DQ and HLA-DP. Finally, the class III encode the complement genes, heat shock protein 70 (Hsp70) and TNF alpha. Due to a great variety of HLA molecules, the identification of such HLAs that are the most relevant for T1D has been challenging [75]. Nevertheless, a number of MHCI and MHCII molecules was implicated, primarily HLA-DRB1, HLA-DQB1, HLA-DQA1 and HLA-DPB1 and several HLA-B alleles. While some HLAII molecules confer high protection against the disease (DQ6.2), other make an individual more susceptible (DR3.DQ2 and DR4.DQ8). It was suggested that major factors determining susceptibility to T1D are chemical and steric features of HLA molecules. Those determine a peptide binding affinity, a strength of interaction with TCRs and a stability of HLA-peptide complexes.Several suggestions about the interplay between MHCII-peptide-TCR complexes and the vulnerability were made. Firstly, the HLA molecules conferring the increased risk of T1D may strongly bind to the beta cell relevant antigenic peptides which leads to an increased presentation and a higher chance of T-cell responses. Secondly, the

HLA molecules which form less stable HLA-peptide complexes in a thymus may increase a probability of T-cells avoiding a removal. Lastly, some individuals may have a TCR repertoire which is associated with a stronger interactions of potentially autoreactive T-cells with HLA-peptide complexes [75]. In recent years however, it has been argued that many of the HLA presented self antigens have low binding affinities to HLA and that that the avidity of TCRs for HLA-peptide complexes is weak, and that one of a main traits of T-cell mediated autoimmune response is an abnormal T-cell signalling [76].

Aside from the HLA T1D predisposing genes, polymorphisms in the IDDM2 locus that encode insulin were implicated. Depending on polymorphism, these variants could grant both protective effects of increase the risk of T1D. It was suggested that a vulnerability to T1D in context of IDDM2 locus is dictated by a regulation of expression of two downstream genes - insulin and insulin-like growth factor 2 (IGF-2). Protective IDMM2 locus variants are associated with an increased expression of insulin the thymus which would increase a presentation of insulin derived peptides by mTECs. On the other hand, a decreased expression of insulin coupled to its increased expression in pancreas may increase the risk of the autoimmune responses. IGF-2 on the other hand, was implicated due to its role in T-cell development and thymic negative selection [75].

Moreover, other genes were also hypothesised to increase susceptibility to T1D. Those include PTPN22, CTLA4, IF1H1, CLEC161 and PTPN2. CTLA4 is receptor on CD4+ and CD8+ T-cells that binds to B7 ligands that in turn activate CD28 co-stimulatory molecule. CTLA4 negatively controls T-cell expansion and activation of T-cells. It was demonstrated that alterations of CTLA4 expression may increase T-cell autoreactivity. Certain substitutions in CTLA4 were shown to lead to decrease of its presence on the cell surface, which could lead to an increased activity of T-cells and T1D association [34]. PTPN2 is a gene coding for lymphoid tyrosine phosphatase (LYP). LYP has an inhibiting effect on TCR signal transduction and mutations in these gene were likewise associated with T1D. Other risk associated genes worth mentioning are interleukin 2 receptor alpha (IL2RA), small ubiquitin modified 4 (SUMO4) and signal transducers and activators of transcription (STAT). IL2RA encodes for CD25 which is expressed on regulatory naive T-cells and memory T-cells. It binds to IL-2 and is important for the proliferation of Tregs which regulate an activity of activated T-cells. SUMO4 regulates an activity of NF-kB which controls an intensity of the immune responses. STAT4 is expressed in PBMCs, dendritic cells and macrophages.

When it comes to autoreactive CD4+ and CD8+ T-cells, several antigens were implicated as potential targets of these cells, including GAD65, IA-2 [195], ZnT8, non-specific islet cell autoantigens (ICAs), imogen 38, pancreatic duodenal homeobox factor 1 (PDX1), chromogranin A (CHGA), islet specific glucose-6-phosphatase catalytic subunit-related protein (IGRP), heat shock protein 60 (hsp60), carboxypeptidase H (CPH), and islet cell antigen 69 (ICA69) [34]. Several methods were used to detect islet reactive T-cells which allowed to identify CD4+ T-cells reactive to ICA69, IA-2, CPH, insulin, GAD, IA- 2, and ZnT8. T1D is characterised by an infiltration of beta islets by mononuclear immune cells, including dendritic cells, macrophages, and T cells which directly contribute to a beta cell destruction [34]. Auto-reactive CD4+ T-cells secrete cytokines and chemokines that in turn activate CD8+ T-cells which in turn exhibit cytotoxic activities and attract other T-cells and macrophages to beta islets. One of the conjectures regarding an escape of self-reactive T-cells from deletion is an abundance of PTMs of antigenic peptides unique to beta islets and not represented in the thymus, as well as products of alternative mRNA splicing and mistakes in translation [34, 76] Despite a great number of antigens implicated into T1D, a primary antigen is thought to be insulin and pro-insulin due to their specific expression in beta cells. Several groups have described a phenomenon of a formation of hybrid antigenic peptides consisting of the fragment of insulin or pro-insulin as a N-terminal splice reactant and a fragment of the beta islet associated antigens from the secretory granules as a C-terminal part of the hybrid peptide. These peptides are produced via ligation which is distinct from the transpeptidation reaction catalysed by the proteasome. These peptides are thought to be generated in crynosomes in beta cells and lysosomes in APCs present in the pancreas and are recognised by CD4+ T-cells. PCPS was also implicated in the production of antigenic peptides triggering autoimmune responses. Transpeptidation reactions were described between IAPP and PTPRN, SLC30A8 and PCSK2, and PIK3R3 and PIK3R1 [76].

2.8 Mass Spectrometry

Mass Spectrometry is an analytical technique which allows to investigate a precise chemical composition of substances. It is achieved by an ionization of the analysed molecules. Each molecule obtains a mass/charge ratio that characterise it. Mass Spectrometers are comprised of three components - ion source, mass analyser and detector. When it comes to ion source, ions are generated via a variety of techniques, one of the most frequently being used is electrospray ionisation (ESI) [77, 78, 79, 80]. It is used to transition substances from the liquid phase to gaseous phase. In case of proteomics, the sample containing the peptides is pumped through a micrometre-sized orficce at a high voltage. The liquid than dissipates into charged droplets that evaporate which leaves the peptides in the gas phase. Depending on an ionisation technique, the ions are either generated by bombardment with electrons or positively charged protons. In many cases, protonation is an ionisation approach of choice. Following that, a mass analyser separates the generated ions via their mass/charge rations (m/z). In the current mass spectrometers most frequently quadruples are combined with orbitrap analysers. Quadrupole mass analyser separates the generated ions via an oscillating electrical field between four cylindrical rods in a parallel arrangement, where each pair of rods produces an electrical field. This procedure allows to distribute the ions according to a specific m/z range. Orbirap mass analysers detect different ions based on their oscillation frequencies [81]. After injection, ions are trapped in the orbitrap while they move across an axis of a kernel metal spindle. Orbitraps are mass analysers of choice due to a high resolution and a low mass deviation [81]. Mass Spectrometry is an invaluable tool for a large scale qualitative and quantitative proteome and peptidome analysis [79, 80]. In bottom up approaches, the proteins are digested with a specific protease, such as trypsin which cleaves proteins into peptides after K and R at P1 position [79, 80]. The generated peptides are then measured by Mass Spectrometry and by using matching their spectra to the theoretical spectrum, are assigned to the the most matching proteins. When it comes to proteomics and peptidomics, one of the most frequently used strategies is a tandem Mass Spectroemtry (MS/MS) which allows to additionally split the peptide precursor ions into fragment ions. The combination of these fragment ions in most cases allow to decipher the precise amino acid sequence of a peptide. To facilitate this, quadruple element is succeeded by a collision cell, where the fragmentation takes place. Orbitraps measure the image current which is quantiative representation of the ions. The current is then converted to a frequency using Fourier transform. The precursor ions that have not yet been fragmented are referred to as MS1 while the fragmented ions are referred to MS2 ions. The MS1 precursor ions are first isolated by quadruple and then fragmented. There are several fragmentation strategies. The two most commonly used are electron transfer dissociation (ECD) and collisional dissociation (CID) [79]. In CID, the fragmentation of precursor ions occurs by collisions with inert gas molecules such as nitrogen, helium or argon in a collision cell. A more efficient variation of CID is higher energy collisional induced dissociation (HCD) which allows for a better ion fragmentation (Figure 14) [13]. Upon collision, the precursor ions break at peptide bonds. The given fragment is only detected if it carries at least one charge. If this charge is retained on a N terminal fragment, the ion is classed as either a, b or c. If the charge is retained on a C terminal fragment, the ion type is either x, y or z [78]. In HCD, b and y ions and to an extent a ions are generated the most frequently [78, 80].

Typically mass spectrometers are coupled to High Performance Liquid Chromatography (HPLC) for additional separation of the peptides based on the hydrophobicity which provides mobile phase compatibility with ESI and maximisation of the identification of peptides [13, 77, 80]. In addition, chromatographic separation allows to characterise ions via a retention time (RT) of the corresponding peptides. The generated MS data consists of a precursor ion mass and charge and a tandem MS spectrum which consists of pairs of observed m/z values and intensities for the detected fragments resulting from the precursor ion. Mass resolution is a dimensionless ratio of a mass of the peak divided by its width. The m/z values that are produced are most frequently monoisotopic, meaning that m/z is calculated for the first isotope peak [78]. Necessary conditions for obtaining the monoisotopic peaks is a sufficient mass resolution to resolve an isotopic distribution and a sufficient signal to noise to be able to isolate and pinpoint a first peak of the isotopic distribution. By contrast, average masses imply merging of the whole isotope distribution into a single peak which suggests a very low resolution. The majority of protemics/peptidomics data is analysed via data-dependent acquisition (DDA), meaning that the mass spectrometer follows a set of user-defined rules (such as m/z, charge, intensity and cross-section). The precursor ions are selected based on the specified characteristics such as charge or intensity of the precursor ions (for example, top N most abundant ions) for acquiring MS/MS spectra [77, 80]. By contrast, in data independent acquisition (DIA), the quadruple surveys the entire mass ranges with large m/z values (*e.g.* via 20-40 m/z steps) [77, 80]. The big challenge in DIA approaches is a correct assignment of a large number of fragment spectra and a detection of low abundant proteins/peptides. To translate the experimentally obtained MS/MS spectrum, it is searched against a theoretical database of *in silico* digested peptides that correspond to a specific mass and MS/MS spectrum. The experimental spectra are then scored against the theoretical spectra [77, 79]. In Mascot search engine, the quality of match is frequently expressed as ion scores (Figure 14) [13]. The higher the ion score, the more likely the experimental MS/MS spectrum is correctly assigned [13, 80]. The additional metric of a quality of a match is q-value which could be used as a measure of a false discovery rate (FDR). FDR is a fraction of correctly assigned spectra. One way to reliably estimate FDR is to implement target-decoy approaches [13]. The entries in the database that was used for searching are reversed and randomised. A rationale is that if such erroneous sequences match the proper database search results, such results are likely incorrect. It's important that the correct and decoy databases are not overlapping. The important consideration in the database search are PTMs [78, 79]. They can significantly impact the m/z of the ions. The PTMs can be either fixed or variable. If the PTM is fixed, it's applied to each corresponding residue in the protein/peptide which simply shifts it's mass. If the PTM is variable it means that it can be applied to only some residues. Due to this, database search using many variable PTMs is avoided due to a large increase in a search space. Thus, typically only the most abundant PTMs that are expected to occur are selected. Yet, other databases such as PEAKS use completely different strategies in which a peptide sequence is assembled *de novo*. To perform a quantitative analysis of the data, the most straightforward approach is label-free quantification (LFQ) [13, 77, 79]. In LFQ, an MS ion current peak area for each identified precursor ion is extracted. A downside to this approach is a large variance in signal intensities attributed to an amino acid composition of each peptide. By contrast, label-based approaches utilise stable isotopes to mark different conditions [79, 82]. The isotopes can be introduced either via metabolic labelling or via isobaric labelling. Upon fragmentation, distributions of isotopes can be deduced.

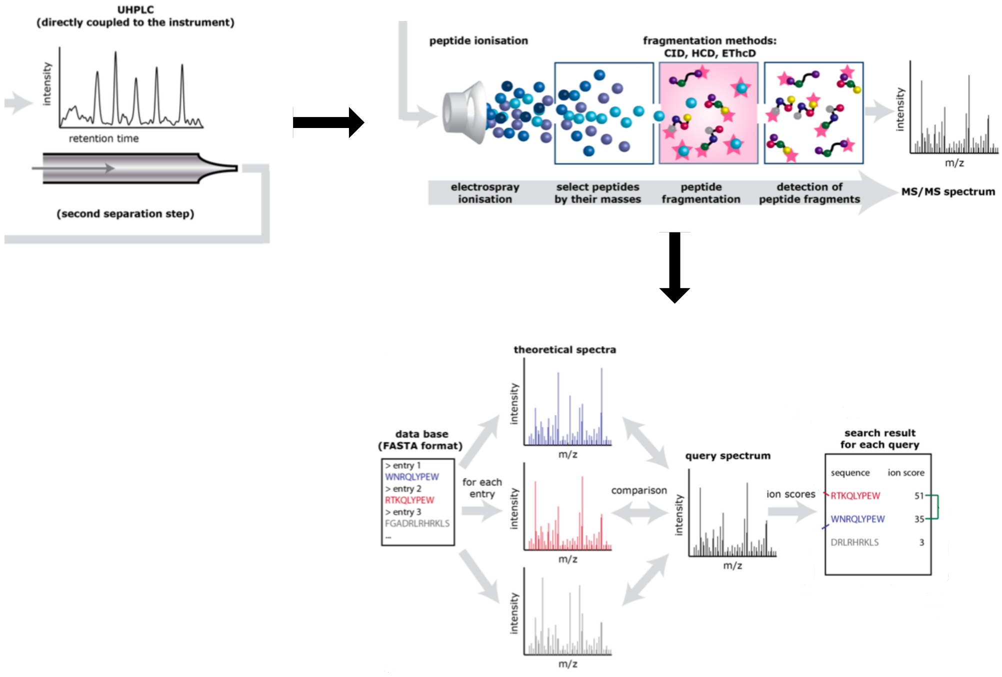


FIGURE 14: The typically used workflow in the shotgun proteomics/peptidomics experiments

The figure was adapted from [13] upon authors authorization

Chapter 3

The role of proteasome generated spliced peptides in the immune evasion by viruses

The results of this chapter were published in [82].

3.1 Introduction

In order for T-cells to be activated, they first have to recognise non-self derived antigenic peptides (epitopes) which are presented on the surface of antigen presenting cells (APCs) by MHCs. MHC class I (MHCI) presents these epitopes to CD8+ T cells. MHCI molecules bind peptides of approximately 8-15 amino acids (AAs) [83]. CD8+ T cells are the primary force against intracellular pathogens such as viruses and some bacteria. Activation of CD8+ T cells triggers a cytotoxic response against the infected cells (*e.g.* secretion of granzymes and perforins by CD8+ T cells) which destroys them along with the pathogen. In order to respond to antigens from different viruses and bacteria, there has to be a wide variety of TCRs with different sequences and binding affinities. In order to achieve this, thymocytes go through the process of somatic recombination (gene-rearrangment) of Variable (V), Joining (J) and Diversity (D) gene segments which are germline encoded, that leads to great variability in the sequences and structures of antigen binding regions of the TCRs that enable them to recognize antigens from a variety of pathogens [83, 84].

This occurs though random selection and shuffling of those gene segments along a chromosome. The process is such that at the junctions of VDJ segments (in CDR3 regions) additional nucleotides are added or removed. In contrast to VDJ gene segments, those junction regions are not germline encoded. The fusion of those segments paired with addition or removal of nucleotides in between them leads to the emergence of a unique antigen binding site. It is estimated that around 10**18 unique TCR could theoretically be generated by this process [32]. This large variety of TCR variants arises during CD8+ T Cell maturation in the thymic cortex. During positive selection in the cortex of the thymus, the emerging thymocytes are selected on the basis of their capacity to bind peptide-MHC complexes present in the thymus with an optimal affinity [83, 85]. They are also selected based on the ability to recognise some conserved regions of MHC proteins. The resulting TCRs are specific for a particular epitopes bound to MHC classes of alleles.

MHCI present both self and non-self derived peptides and thus its necessary that non-self peptide-MHC (pMHC) complexes are sufficiently different from self pMHC to avoid triggering of the immune response while recognising the majority of viral epitopes. To achieve this, during the negative selection of nascent double-negative (DN) T-cells (T cell clones that would become CD8+TCRalphabeta in case of MHCI presentation pathway) in the medulla of the thymus, those that recognize self pepetide-MHC complexes with high affinity are deleted from the pool of circulating T-cells or rendered unresponsive [83]. It was demonstrated that it's the high avidity of TCRs for peptide-MHCs that determines whether a given T-cell would be tolerized or spared [83, 85]. It was demonstrated that changes in the binding properties of thymocytes could have a significant impact on which thymocytes are tolerized. The negative selection occurs due to the presentation of a wide range of self peptide-MHCI complexes on the surface of Antigen Presenting Cells (APCs) in the medulla of the thymus. These APCs, such as medullary Thymic Epithelial Cells (mTECs) and thymic Dendritic Cells (DCs), express transcription factors such as autoimmune regulator (AIRE) in case of mTECs that drives the expression of a very large variety of self-antigens. Interestingly, different subpopulations of mTECs express only a small portion of antigens (1-3%) [85]. The direct presentation of antigens by mTECs is responsible for the bulk of negative selection that occurs in the thymus. Antigen presentation in the medulla promotes the identification of potentially autoreactive CD8+ TCR T cell clones and their elimination [85]. Alternatively, these cells are directed to a different fate and become regulatory T-cells [85]. As a result, only such T cells that don't recognize self-peptide HLA-I complexes transform into naive CD8+ T-cells and join the overall circulating pull of T-cells present in the periphery. Interestingly, it's the negative selection that in large part determines the specificity of T-cells for non-self peptides and to self MHC molecules [85]. The defects in the process of the negative selection results in autoimmune diseases such as Type 1 Diabetes (T1D) and Multiple Sclerosis (MS) [31, 49]

Despite the presence of central tolerance induction, some auto-reactive T-cells still survive and are present in the periphery [83, 85]. It was demonstrated that only about 60-70% of self reactive T-cells are pruned during central tolerance induction [50]. For example, It was demonstrated in the transgenic mice expression TCR-beta chain of an anti-H-Y T cell clone recognising male-specific Db-Smcy3 complex that as many as 25-40% of self-reactive T-cell could escape torlerance in both thymus and the periphery [86]. This ability of certain T-cells to escape clonal deletion is thought to be in part associated with an antigen expressiohn from which self-reactive epitope is derived.

In the other study, it was shown via peptide-MHC tetramer enrichment that T-cells specific to a range of epitopes from self-antigens were present in the periphery in abundances similar to T-cells reactive for peptides derived from non-self antigens in peripheral blood mononuclear cells (PBMCs) of healthy donor [84]. Despite this similarity, non-self and self specific T-cells exhibited different gene expression profiles suggesting a correlation with their subsequent fates. Self-reactive T-cells were also resistant to an ex vivo activation. Along these lines, It was demonstrated that T-cells specific to Y chromosome encoded SMCY antigen were present not just in females but also in males only at 3 times lower frequency [84].

TCRs recognize peptide-MHC complexes via primarily the complimentary determining region 3 (CDR3) [86, 87]. It has the highest genetic variability which translates into high sequence diversity. As such length of CDR3 loops and their amino acid composition were suggested to be the predictors of clonal deletion [86].

Nevertheless, in the majority of cases, the deletion happens in the medulla of the thymus and when tolerance induction mechanisms work as intended, potentially self-reactive T-cells are deleted or rendered unresponsive [83, 85]. On the one hand, it allows to prevent autoimmune responses but on the other, potentially considerably limits the diversity of non-self peptides that could be utilized by the immune system. If some of the self-epitopes recognised by the potentially auto-reactive CD8+ T cells are identical to non-self-epitopes which could be generated from viral antigens, one would expect an impaired CD8+ T cell response against viruses, since these potentially autoreactive CD8+ T cell clones would have been eliminated in the thymus or pruned in the periphery. Previously, it was shown that the immune system is likely not to react to a moderate number of the non-spliced viral and bacterial epitopes based on the cross-reactivity of T cell receptors (TCRs) of CD8+ T cells recognizing the self peptides. Molecular mimicry of the non-self epitopes to self was suggested to be a viable strategy for the immune evasion in several experimental studies. This could create so called "holes" in the T-cell repertoire.

It was assessed in a number of studies whether either immune evasion of the pathogens via T-cell holes or their triggering of the autoimmune responses could be explained by the sequence overlaps of self and non-self peptides produced by the regular peptide-bond breakage by the proteasome. In one study it was investigated whether there are short peptide sequences of length of 5 amino acids (AAs) embedded in the human proteome that are shared with selection of viral proteomes. In total 36103 unique human proteins were compared against 717 viral proteins from 30 different viruses with human tropism. All of the 30 viral proteomes had multiple instances of the sequence overlap of peptides of length 5 to human proteomes (over 90% of peptides were identical between viral and human proteomes). There was a strong correlation between the extent of the sequence overlap and the length of any given virus [88].

Similarly, a particularly high sequence overlap was determined for peptides derived from hepatitis C (HCV) virus and human proteome on the level of pentamers suggesting that the extensive sequence sharing could explain high infection rates by HCV in human [89]. In hepatitis B virus (HBV) it was shown that its proteome shares 65 peptides of length 7, one of length 8 and one peptide of length 9 [90]. These similarities in their proteomes were suggested to be associated with autoimmune reactions caused by HBV.

Significant sequence overlaps of peptides of length 9 amino acids were also found between 40 bacterial proteomes [91]. One third of all human proteins was found to share at least one peptide with at least one bacterial species regardless of their pathogenicity towards human. As for viruses, the number of overlapping peptides was strongly correlated with the size of bacterial proteomes. This was suggested to aid in the immune evasion of the bacterial pathogens. In the follow-up study by the same group, extensive overlaps were found between human and bacterial peptides of lengths 5, 6, 7 and 8 across 40 different bacterial species [92]. In fact, all of the human protein contained at least one peptide of length 5 overlapping with bacterial proteome.

The matter of the immune evasion via molecular mimicry was investigated in detail by Calis J., *et al.* who looked at the extent of the self/non-self sequence overlap of cleavage peptides presented by MHCI molecules for a variety of bacterial and viral proteomes [24]. The authors considered both exact and degenerate overlap based on T-cell cross-reactivity by implementing HLAI binding predictions and T-cell recognition model. They showed that with the increase in the peptide length the overlap drops drastically. The majority of all theoretically possible peptides with the length of 9 amino acids (9mers) were distinct from self-peptides - only 0.2% were identical to self peptides for viruses and 0.19% for bacteria, on average, which couldn't explain the immune evasion of pathogens via T-cell holes. Upon incorporation of the T-cell cross-reactivity model, the self/non-self overlap increased drastically. On average, the self-non-self overlap increased up to 0.7% for 9mers when assuming that TCRs could interact with any position in a 9mer and 29% based only on the middle 6 positions of the 9mers, respectively where TCRs are known to interact with the epitopes presented by MHCI molecules, when considering degenerate recognition model [24]. The authors concluded that even when considering cleavage peptides, a large fraction of non-self peptides would be invisible to the immune system thus creating gaps in T-cell mediated immunity - T cell holes.

Prior to Calis J., Burroughs *et al* studied the extent of the ability of the immune system to distinguish self from non-self using distinct cleavage of 9mers derived from human proteins and showed similar overlap of non-self epitopes to self [93].

The potential practical implications of homology of epitopes to self in the immune system were demonstrated in the study conducted by Rolland M., *et al.* in which they investigated the viral mimicry of HIV-1 to human [94]. They focused on HIV-1 specific Cytotoxic T lymphocytes (CTLs) and discovered that the higher the similarity of nonself peptide to the proteome was, the lower the observed ELISpot responses was hence the lower the number of patients responding to the peptide was.

Assarsson E., *et al* examined the repertoire of T-cells capable of recognizing Vaccinia Virus derived epitopes presented by HLA-A*02:01 molecules [95]. They identified 1679 potentially immunogenic epitopes. Only 630 out of those candidate epitopes were immunogenic in vaccinia-immunised mice. In another study, 170 unique peptides derived from vaccinia virus were identified in the H-2b immunopeptidomes of infected mouse cells. About 80% of those peptides elicited an immunogenic response in at least one of the infected mice and 39% elicited the response in more than half of mouse [96].

The immune escape of Hepatitis C virus (HCV) was also suggested to possibly occur due to similarity of nonself derived Hepatitis C peptide to self by Wolf *et al*, even though infection with it causes virus-specific immune responses [97]. HCV manages to avoid destruction in the majority of infected individuals. Among potential mechanisms of immune evasion proposed was the creation of T cell holes. The authors demonstrated for an HCV derived peptide that a mutation altering TCR contact residue significantly curtailed TCR recognition and failed to activate CD8+ T cells without affecting peptide processing or MHC binding. The lack of reactivity to this variant could be explained by a lack of appropriate T cells.

We named these matching peptides, zwitter peptides [82]. Zwitter is the German word for hybrid, hermaphrodites from zwi-, duplex. If CD8+T cells specific for zwitter epitopes were eliminated in thymus, they could no longer detect the viral epitope during an infection, which could result in a "hole" in T cell repertoire. If, on the contrary, CD8+ T-cells were not eliminated by the central tolerance, they might be primed during viral infection and, since they recognize also a self-antigen, this might trigger an autoimmune response.

Given the tentative importance of molecular mimicry of non-self peptides to self-peptides for the immune evasion by pathogens, one question related to the relevance of the spliced peptides for immune response that we set out to answer was how this new found breadth of potential epitopes would affect the immune screening and the size of T-cell holes.
We aimed to answer the following research questions:

- What is the sequence coverage of human proteome by all possible non-spliced and spliced viral peptides of length 9 amino acids (9mers) - total number of viral-human zwitter peptides?
- How many of viral-human zwitter peptides are HLA-A*02:01 binders (a theoretical size of T-cell holes)?
- How does the antigens expression in the thymus affect the size of T-cell holes?
- What is the size of T-cell holes considering frequencies of spliced peptides in the human cells immunopeptidome?

3.2 Materials and methods

3.2.1 Estimation of viral-human zwitter peptides (Figure 15A)

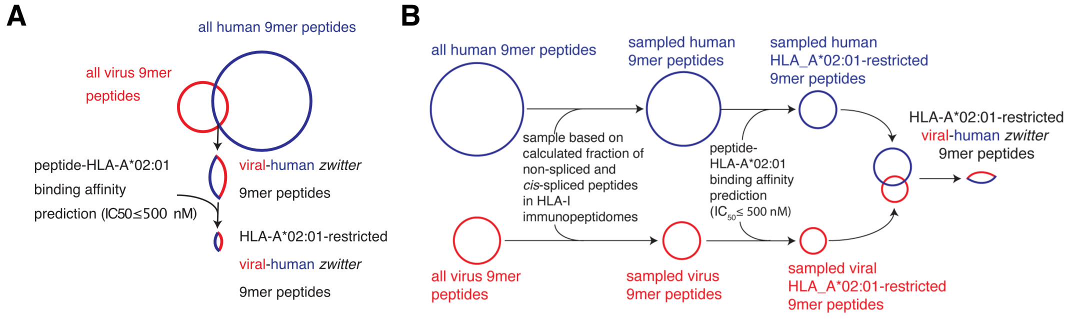

A
all human 9mer peptides
all virus 9mer peptides
peptide-HLA-A*02:01 binding affinity prediction (IC50≤500 nM)
viral-human *zwitter* 9mer peptides
HLA-A*02:01-restricted viral-human *zwitter* 9mer peptides
B
all human 9mer peptides
sampled human 9mer peptides
sampled human HLA_A*02:01-restricted 9mer peptides
sample based on calculated fraction of non-spliced and *cis*-spliced peptides in HLA-I immunopeptidomes
peptide-HLA-A*02:01 binding affinity prediction (IC50≤ 500 nM)
HLA-A*02:01-restricted viral-human *zwitter* 9mer peptides
all virus 9mer peptides
sampled virus 9mer peptides
sampled viral HLA_A*02:01-restricted 9mer peptides

FIGURE 15: An overview of the general approach used to estimate viral-human zwitter peptides

Schematic representation of *in silico* pipelines to estimate the frequency of zwitter peptides predicted to bind HLA-A*02:01 complexes not accounting (A) or accounting (B) for non-spliced and cis-spliced peptide frequency in HLA-I immunopeptidomes. For (A) all possible 9mer sequences of either non-spliced or spliced peptides are computed for viral and human proteomes. First, all theoretically possible viral and human 9mers are aligned which results zwitter estimate among all possible 9mers. Following HLA-A*02:01 binding prediction, a subset of identical peptides is derived (roughly 4%), which constitutes the zwitter HLA-A*02:01 binders.

Viral proteomes were obtained via ViralZone and Human proteome via from Swiss-Prot [98, 99]. The Human proteome database contained 20191 protein entries with a total of 11323862 amino acid residues. Per viral species, one strain was selected. Viruses were chosen depending on whether human is a known host for those viruses. 109 viruses were selected (Table S1).

In this study, we computed all theoretically possible viral-human zwitter peptides of length 9 amino acids. The reason we focused on peptides of length 9 is that it is known that in human the majority of peptides binding to MHCI molecules have length of 9 amino acids [13, 14, 70]. We define a given peptide as zwitter if it has the exact match of all positions (1-9) of the peptides of length 9 amino acids (AAs) (from here on referred to as 9mers) between a given human and a viral 9mer. Another words, zwitter 9mer peptide was any 9mer peptide that had a sequence, that could be obtained by either peptide hydrolysis or cis peptide splicing both from self proteins and from viral proteins. This was estimated for all 9mers. For viral and human proteomes, we computed all possible 9mer sequences of non-spliced peptides by cutting proteins into fragments of length 9 amino acids and of all possible normal and reverse cis-spliced peptides sequences by computing combinations of any N- and C-terminal splice reactants of any length such that the resulting spliced peptide sequence had a length of 9 amino acids. We had to impose a number of restrictions on the computation of spliced peptide database to avoid unfeasibly large databases. The maximum intervening sequence length of <= 25 AAs between splice reactants was set and trans-spliced peptides weren't considered in this study due to unfeasibility of computing all possible trans-spliced peptides without knowing precise driving forces (amino acid and peptide sequence preferences) behind their generation

[14, 27]. Following that, a sequence matching of computed 9mer sequences was performed between viral and human derived peptides.

We define 2 types of zwitter peptides - non-spliced and cis-spliced. Cis-spliced peptides include non-spliced viral peptides overlapping to cis spliced human peptides, cis-spliced viral peptides overlapping to non-spliced human peptides and cis-spliced viral peptides overlapping to cis-spliced human peptides.

The frequency (share) of viral-human zwitter peptides was calculated as a fraction of viral peptides that are zwitter relative to all viral 9mers. For the entire viral proteome, the share of zwitter peptides was calculated by dividing the number of zwitter peptides by all unique computed viral peptides.

$$F_v = (z_v / p_v) * 100$$

Where F_v - The relative frequency of viral-human zwitter peptides; z_v - number of all vira-human zwitter 9mers; p_v - number of all possible 9mers for a given virus

The share of combined non-spliced and spliced zwitter peptides was calculated by summing up unique non-spliced and spliced zwitter peptides and dividing this number by the total number of all possible unique viral non-spliced and spliced peptides:

$$F_{v,all} = ((z_{v,i} + z_{v,j} + z_{v,k} + z_{v,l}) / p_{v,all}) * 100$$

Where $F_{v,all}$ - the share of combined non-spliced and spliced zwitter peptides; $z_{v,i}$ - number of all spliced viral 9mers overlapping to human spliced peptides; $z_{v,j}$ - number of spliced viral 9mers overlapping to human non-spliced peptides; $z_{v,k}$ - number of non-spliced viral 9mers overlapping to human spliced peptides; $z_{v,l}$ - number of non-spliced viral 9mers overlapping to human non-spliced peptides; $p_{v,all}$ - number of all possible non-spliced and spliced peptides for a given virus

Next, we narrowed our analysis of zwitter peptides to those which are predicted to be presented by MHCI molecules. MHCI binders are such peptides which have sufficiently high binding affinity to MHCI molecules and are presented on the cell surface for TCRs. Only these peptides would be relevant from immunological point of view and would constitute the actual T-cell holes. In human, MHCI molecules are also known as Human Leukocyte Antigens (HLAs). In this study we considered zwitter peptides, predicted to be HLA-A*02:01 binders. HLA-A*02:01 was chosen for the analysis because it is the most frequently encountered HLA haplotypes in caucasian population with the known binding affinity - 500 nM [24]. For peptides predicted to be HLA-A*02:01 binders, we determined how many of such binders are zwitter and calculated the frequency of such peptides

$$B_v = (z_b / b_v) * 100$$

Where B_v - frequency of HAL-A*02:01 zwitter peptides; z_b - number of viral-human zwitter 9mers predicted to be HLA-A*02:01 binders; b_v - number of all possible 9mer HLA-A*02:01 binders for a given virus.

Statistical test values for the comparisons between different groups of viral-human zwitter peptides are provided in Table S2 and Table S3.

3.2.2 Peptide-HLA-A*02:01 binding affinity prediction

Binding of computed non-spliced and spliced 9mers to HLA-A*02:01 molecule was predicted using Stabilized Matrix Method (SMM). The standalone version of binding tool was downloaded from IEDB Analysis Resource [100]. We set a binding affinity cut-off to <= 500 nM to distinguish binders from non-binders to HLA-A*02:01.

In order to assess whether overlapping peptides are more likely to be HLA-A*02:01 binders, on per virus basis we counted the number of non-zwitter viral non-binders, non-zwitter viral binders, zwitter viral non-binders and zwitter viral binders and then performed odds ratio test.

3.2.3 Estimation of viral-human zwitter peptides considering the potential antigen repertoire of human mTECs

In order to analyse the impact of antigens expression in medullary thymic epithelial cells (mTECs) on the number of zwitter peptides that could be relevant for the immune evasion, we obtained the data from studies by two independent groups - microarray gene expression values of mature medullary thymic epithelial cells (mTECs) (antigen presenting cells (APCs) involved in the negative selection in the thymus) that were used to study self-antigen diversity and single cell RNA sequencing data (scRNA-seq) of progenitor TECs used to investigate thymus organogenesis with mTEC properties[101, 102]. In the first study, the material was derived from patients that underwent corrective cardiac surgery. The material for the second data-set was derived from healthy human fetuses as a result of medically interrupted pregnancy at weeks 8, 9 and 10. We used the subset of data that ostensibly corresponded to TECs with progenitor property of mTECs (based on the expression of mTEC markers CLDN4 and JAG1). For microarray derived data, we calculated average gene expression values (reported as log2 transformed fluorescence intensities) for technical replicates of each mTEC subset provided in the dataset obtained with two versions of microarray and took the maximum average value. For scRNA-seq we performed log-normalization of gene expression values of individual cells (reported as copy number of transcripts per individual gene (number of distinct unique molecular identifiers (UMI)) to mitigate the relationship between sequencing depth and gene expression and then took an average of the normalized gene expression value between individual cells [103, 104]:

$$x_i = log((100000 * UMI_{ij} / \sum_{i=1}^{n} UMI_{ij}) + 1)$$

Where x_i - log-normalized expression of gene i of cell j; $100000 * UMI_{ij}$ - expression value of gene i of cell j prior to normalization expressed as UMI counts; $\sum_{i=1}^{n} UMI_{ij}$ - sum of UMI counts per cell j.

Since we were not only interested in understanding which antigens were most likely expressed, but also which antigens were presented in HLA-I immunopeptidomes of mTECs, we did not set a fixed cut-off for gene expression, but we defined a crude model for antigen presentation based on the gene expression values. We assumed that the chance of an antigen being presented in mTECs HLA-I immunopeptidomes was directly correlated with the gene expression of that antigen. Thus, using the gene expression values of the processed data, we normalized them for each gene to weights via min-max normalization approach such that the sum all weights would be equal 1 for both Microarray and scRNA-seq derived data:

$$w_i = (E_i - min(E))/(max(E) - min(E))/\sum_{i=1}^{n}(E_i - min(E))/(max(E) - min(E))$$

Where w_i - expression value of gene i normalized via min-max approach to weight; E_i - expression value of gene i prior to normalization; $max(E)$ - largest expression value in the dataset; $min(E)$ - smallest expression value in the dataset

Based on the outcome of sequence matching of 9mers described above, locations of zwitter non-spliced peptides and cis-substrings in each human antigen could be derived. If a given antigen was shown to be expressed in mTECs, we assigned probabilities of presentation of zwitter peptides from each antigen based on the expression values of their source antigens in mTECs.

Following that, we sampled unique zwitter peptides based on calculated probabilities of presentation in mTECs. The sampling size was set at 100% of the number of zwiiter peptides to reflect the odds of presentation of each given peptide. Sampling was performed with replacement based on the calculated probabilities 60 times for statistical power. We then recalculated the share of zwitter peptides.

The frequency of viral-human zwitter peptides based on gene expression in mTECs compared to all viral 9mer was computed as:

$$M_v = (z_{m,v}/p_v) * 100$$

Where M_v -The frequency of viral-human zwitter peptides based on gene expression in mTECs; $z_{m,v}$ - the number of sampled viral-human zwitter peptides with weights w_i and p_v - is the number of all possible 9mer peptides of virus v.

When we considered both predicted zwitter peptide-HLA-A*02:01 binding affinity and gene expression in mTECs, the viral-human zwitter peptide frequency was computed as:

$$MB_v = (z_{m,b,v}/b_v) * 100$$

Where MB_v - viral-human HLA-A*02:01 zwitter peptide frequency based on gene expression in mTECs; $z_{m,b,v}$ - is the number of sampled viral-human zwitter peptides restricted to HLA-A*02:01 binding with weights w_i and b_v is the number of all possible 9mer peptides restricted to HLA-A*02:01 binding of virus v.

3.2.4 Estimation of the frequency of viral-human zwitter peptides weighing PCPS frequency (Figure 15B)

Not all 9mer non-spliced and cis-spliced peptides that could derive from the human proteome are produced by proteasomes and presented on the cell surface by HLA-I molecules [7]. Therefore, we considered this crucial factor in our *in silico* analysis of zwitter peptides. We aimed to determine the fractions of non-spliced and cis-spliced peptides produced by the proteasome and presented on a cell surface, relative to all theoretically possible sequences and the subsequent impact on the frequency of both non-spliced and cis-spliced zwitter peptides (Figure 1B).

This is possible due to the fact that a large data-set of *in vitro* digestions of synthetic polypeptides is currently available to us and that a good correspondence was demonstrated between *in vitro* digestion and *in vivo*/*in cellulo* data [2, 10, 11, 12, 20, 64, 71, 72, 105, 106, 107, 108, 109, 110]. Moreover, we integrated the information on the currently available estimates of the frequency of spliced peptides constituting immunopeptidome that aided us in our exploration of the spliced peptides fraction produced by the proteasome and presented on the cell surface by MHC-I molecules. A wide range of frequencies of spliced peptides relative to all peptides in the cellular immuno-peptidomes was reported in the past and was highly dependent on the approach for spliced peptides identification, software that was used for the data analysis and database searching, and the material used for the analysis [13, 14, 15, 21, 70, 73, 74].

We have recently published a large database on *in vitro* digestions of a variety of synthetic polypeptides by 20S and 26S proteasomes [17]. From this *in vitro* digestion data, it can be derived that significant portion of all possible non-spliced peptides and a very small fraction of spliced peptides are produced. Based on the available *in vitro* digestions of a variety of synthetic substrates by the 20S proteasomes, we could further determine the average fraction of produced non-spliced peptides. In order to estimate the fraction of non-spliced peptides produced by proteasome we used 4 hours digestions in our database by 20S standard proteasomes as the bulk of the accumulated data came from digestions by this proteasome type (47 synthetic substrates in total). We chose 4 hours digestions in order to better represent the realistically expected diversity of products that could be generated by proteasome *in vivo*. Our published database contains only peptide sequences that passed the appropriate quality control steps (for details see [17]). This large database contains 2,429 unique non-spliced and 2,379 unique cis-spliced peptide products.

As stated above, an estimate of fraction of non-spliced peptides can be directly obtained from *in vitro* digestions of synthetic polypeptides with purified proteasomes. We first calculated the fraction of all produced non-spliced peptides (included in the published database) relative to all theoretically possible non-spliced peptides for each polypeptide in the database:

$$f_{non} = (n_{non}/N_{non}) * 100$$

Where f_{non} - the fraction of all produced non-spliced peptides; n_{non} - number of identified non-spliced peptides; N_{non} - number of all theoretically possible non-spliced peptides

and then took the median value between all polypeptides as the estimate of fraction of non-spliced peptides generated by proteasome. These calculations resulted in the value of 27%. We could have used the same strategy to compute the fraction of cis-spliced peptides produced by proteasomes compared to all theoretical cis-spliced peptide products as:

$$f_{cis} = (n_{cis}/N_{cis}) * 100$$

Where f_{cis} - the fraction of cis-spliced peptides produced by proteasomes compared to all theoretical cis-spliced peptide products; n_{cis} - number of identified non-spliced peptides; N_{cis} - number of all theoretically possible non-spliced peptides

However, cis-spliced peptides have been proven to be produced in significantly lower amount than non-spliced peptides. This would mean that cis-spliced peptides produced by proteasomes *in vitro* could not pass all APP steps and become antigenic as compared to non-spliced peptides. On the contrary, in HLA-I immunopeptidomes all the peptides that could not have passed the preceding APP steps have been filtered out and preserving only relevant peptides. We thus integrated the information available about cis-spliced peptide frequency in HLA-I immunopeptidomes measured through MS reported by various groups with the information of non-spliced peptide frequency in *in vitro* digestions.

Its thus possible based on the fractions of generated peptides obtained from *in vitro* digestions, to calculate frequencies (f) of spliced peptides in the immunopeptidome *in vivo*. Keeping these facts in mind, relative frequency (fraction) of cis-spliced peptides (f) as measured by Mass Spectrometry (MS) can be calculated as following:

$$f = 100 * n_{cis}/(n_{cis} + n_{non})$$

Where f - relative frequency (fraction) of cis-spliced peptides in HLA-I immunopeptidomes; n_{cis} - number of presented spliced peptides; n_{non} - number of presented non-spliced peptides.

f was estimated to be in the range of 1-34% according to various studies. We could then compute the number of cis-spliced peptides presented in HLA- I immunopeptidomes for a given estimate of f as:

$$n_{cis} = 100 * f * N_{non} * f_{non} / (100 - f)$$

Where n_{cis} - number of cis-spliced peptides presented in HLA- I immunopeptidomes for a given estimate of f; N_{non} - number of theoreticaly possible non-spliced peptides; f_{non} - non-spliced peptide frequency in *in vitro* digestions

The total number of all theoretical cis-spliced peptides can be computed as:

$$N_{cis} = \gamma * N_{non}$$

Where N_{cis} - The total number of all theoretical cis-spliced peptides; γ - a ratio of all possible spliced to all possible non-spliced peptides (in calculation we use 398 which was estimated for human antigens longer than 500 AAs); N_{non} - number of theoreticaly possible non-spliced peptides Finally, these transformations result in:

$$f_{cis} = 100 * f * f_{non} / \gamma * (100 - f)$$

Where f_{cis} - the fraction of cis-spliced peptides produced and presented on HLA-I immunopeptidomes

Importantly, numbers of all theoretically possible spliced and non-spliced peptides for a given polypeptide or antigen can be computed and are readily available. The fractions of non-spliced peptides produced by proteasome could be derived from the experimental *in vitro* digestions data, while the frequencies of spliced peptides in the immuneopeptidome were evaluated in multiple studies [13, 14, 15, 21, 70, 73, 74].

Thus, from the fraction of non-spliced and hypothesized frequencies of spliced peptides in the immunopeptidome, we can subsequently estimate the fractions of spliced peptides produced by the proteasome *in vitro*.

We used a range of potential frequencies of spliced peptides relative to all produced peptides to calculate fraction of spliced peptides produced by the proteasome - 1-34%. Using these frequencies and the calculated fraction (27%) produced non-spliced peptides, we could calculate the fractions of observed spliced peptides relative to all possible spliced peptides. Based on these derived fractions, we randomly sampled non-spliced and spliced peptides 600 times from all viral and human proteomes without replacement. Following that, viral-human zwitter peptides among all sampled peptides and sampled HLA-A*02:01 binders were computed as described and finally zwitter 9mer and zwitter HLA-A*02:01 binders were determined.

3.2.5 Statistical analysis

Unless stated otherwise, all statistical tests were done in R and differences in distributions were tested using two tailed Kolmogorov-Smirnov test (ks.test) with cut-off for significance set to p-value < 0.05. To determine the relationship between virus length and number of zwitter peptides, we calculated Pearson correlation coefficient. The strength of association of non-zwitter vs zwitter peptides being HLA-A*02:01 binders was computed via odds ratio and significance was tested using Fisher exact test, or alternatively chi square test if the sample size was too large for Fisher exact test to test significance of association.

3.2.6 HIV-Derived HLA-A*02:01-Restricted Non-immunogenic 9mer Peptides

As proof of principle, we selected a pool of HIV-derived HLA-A*02:01-restricted 9mer peptides, which were previously suggested to be non-immunogenic. This pool included non-spliced epitope candidates derived from HIV, which: (i) were investigated by Perez *et al*. through IFN-gamma ELISpot assay in HIV1 - infected donor peripheral blood mononuclear cells (PBMCs) pulsed/non-pulsed with synthetic epitope candidates. We considered as non-immunogenic those peptides that did not induce immune response after peptide stimulation [111]. (ii) were included in a database by Ogishi and Yotsuyanagi. This database collected outcomes of various T cell activation assays on HLA-I-restricted non-spliced peptide sequences (8-11 mer peptides). In this database, we selected HIV-derived HLA-A*02:01-restricted 9mer peptides, which were confirmed as non-immunogenic among all studies considered in the database [87]. (iii) were included in the EPIMHC database, which collected datasets of T cell response against epitope candidates. In this database, non-immunogenic peptides were selected by applying the following parameters: Allele, HLA A*02:01; Length,9mer; MHC source, Human; Peptide source organism, HIV1; Peptide Binding Level, all; T-cell activity, all; Immunogenicity level, all; Processing, all. The pool of peptide candidates derived from these three databases were then analyzed for peptide-HLA-I bind affinity predictionas described aboveand only peptides with predicted peptide-HLA-A*02:01 IC50 <= 500 nM were selected (Table 1) [112]. The identification of eptiope candidates reported in Table 1 was performed by Dr. Camila R. R. Barbosa.

3.2.7 Modeling of Protein 3D Structures

For visualization purpose, the structures of Gag-Pol polyprotein of the HIV strain MVP5180 and the human Major Vault protein (MVP) were predicted and visualized through the fully automated protein structure homology-modeling server, accessible via Expasy web server [113].

3.3 Results

3.3.1 Estimation of viral-human zwitter peptides

It is known that MHC class I molecules primarily present peptides of 9 amino acids long (AAs) (9mers). First, we wanted to investigate the sequence coverage of human proteome by viral proteome considering all possible 9mers. To do this, we computed all non-spliced peptides of length 9 amino acids and all computationally possible spliced peptides for both viral and human proteomes. We set a limit on intervening sequence length of 25 amino acids and only considered cis-spliced peptides - peptides which splice reactants originate from the same substrate molecules. Following that, we aligned the resulting 9mer sequences between virus and human and calculated the complete sequence overlap (*i.e* zwitter peptides). We identified 2340 and 9350135 theoretical viral-human zwitter non-spliced and cis-spliced 9mer peptides, respectively. On average they correspond to 0.06% and 2.89% of the share of zwitter non-spliced and spliced peptides to all viral non-spliced and spliced peptides per virus, respectively (Figure 16A). This small number of non-spliced zwitter peptides was expected based on the previous estimates. However, the inclusion of spliced peptides into the assessment substantially increased the number of zwitter peptides despite the significantly larger number of ways spliced peptides could be generated and even with the imposed restrictions on spliced database. Moreover, spliced peptides account for the majority of the combined number of zwitter peptides across all viruses.

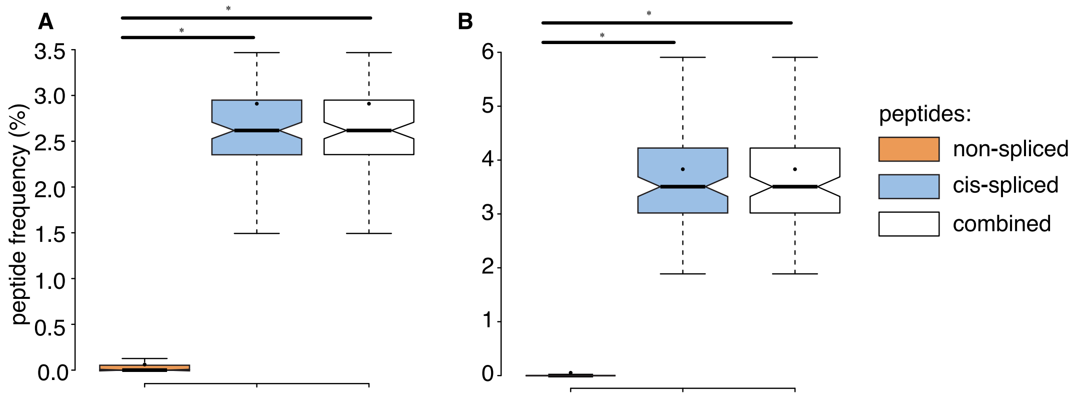


FIGURE 16: Viral-human zwitter peptides.

(A, B) Frequency of viral-human 9mer (non-spliced, cis-spliced and combined) (A) zwitter peptides and (B) HLA-A*02:01-restricted zwitter peptides, compared to their cognate viral peptide database and considering the whole human proteome database. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Dots represent the mean. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides. Significant difference between groups are labelled with * (p-value < 0.05).

3.3.2 Estimation of viral-human zwitter peptides based on peptide-MHC-I complexes

Only a minority of all peptides produced by proteasome would be presented by MHC-I molecules and recognized by TCRs [95]. We were especially interested in zwitter MHCI binders as only HLAI binders are relevant from the immunological point of view. We thus considered how the theoretical size of holes in CD8+ T-cell mediated immunity would change considering number of zwitter peptides predicted to be HLA-A*02:01 binders. To estimate the number of HLA presented zwitter peptides, we performed the prediction of binding of 9mers to a frequently encountered HLA-A*02:01 haplotype of MHCI. Finally, we checked which of the zwitter peptides found among all 9mers were among predicted binders. Interestingly, regardless of the fact that binding prediction algorithms were trained exclusively on non-spliced peptides, roughly 4% of peptides were predicted to be binders for both non-spliced and spliced peptides. This suggests that properties of anchor residues determining the binding to HLAI for spliced and non-spliced peptides are similar. The proportion of binders would depend on the algorithm for binding predictions, type of HLAI molecule as well as threshold for distinguishing binders from nonbinders which vary depending on HLAI type.

When we considered HLA-A*02:01 binders, the total number of zwitter non-spliced peptides was 87 which corresponds to 0.05% of the overall pool of viral HLA-A*02:01 binders, on average. This would suggest that purely based on 9mer sequences and considering only non-spliced peptides, the immune system would easily be able to distinguish between self and non-self, since only a minority of presented peptides would be zwitter. The number of zwitter spliced peptides on the other hand was 504339 which constitutes 3.816% of the total number of spliced viral HLA-A*02:01 binders, on average (Figure 16B). Among zwitter, HLA-A*02:01 binders constitute roughly 5% of all zwitter 9mers (Figure 17A).

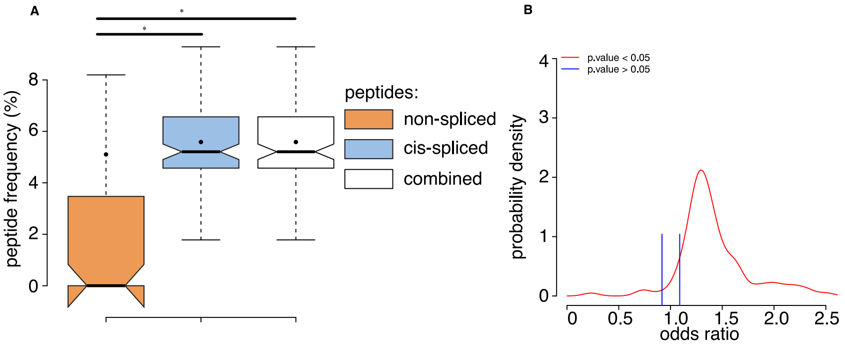


FIGURE 17: Viral-human zwitter binders.

(A) Frequency of HLA-A*02:01 binders among viral-human zwitter peptides. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Dots represent the mean. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides. Significant difference between groups are labelled with * (p-value < 0.05). (B) Probability densities showing distributions of odds ratios that zwitter peptides are more likely to be HLA-A*02:01 binders than non-zwitter peptides per virus.

To additionally investigate the relevance of zwitter viral 9mers presented for the immune evasion, for each virus we performed odds ratio test to determine if they were more likely to be HLA-A*02:01 binders compared to viral non-zwitter peptides. Zwitter peptides are more likely to be HLA-A*02:01 binders than non-zwitter peptides (Figure 17B).

As expected, when we relaxed IC50 cut-off, the number of both viral-human zwitter non-spliced and cis-spliced 9mer epitopes increases (Figure 18A). We then narrowed down our search to a very stringent IC50 cut-off of 50 nM, in order to determine the theoretical share of viral-human zwitter 9mer epitope candidates among putatively the most immunodominant peptides. In the past, a correlation between the immunogenicity of epitope, its binding affinity and stability in complex with MHCI had been shown. For instance, Platteel *et al.* demonstrated this for cis-spliced epitope candidates with predicted binding affiniti to H2-Kb of IC50 <= 2nM in a mouse model of Listeria monocytogenes infection [20]. Assarsson *et al.* used a transgenic mouse model to show that the binding affinity of all vaccinia derived immunodominant HLA-A*02:01 epitopes was below 50 nM [95]. Restricting IC50 cut-off to <= 50 nM left 11 non-spliced and 87,154 cis-spliced zwitter peptides peptides which correspond on average per virus to 0.06 and 4.19% of the pool of HLA-A*02:01-restricted viral peptides (Figure 18B).

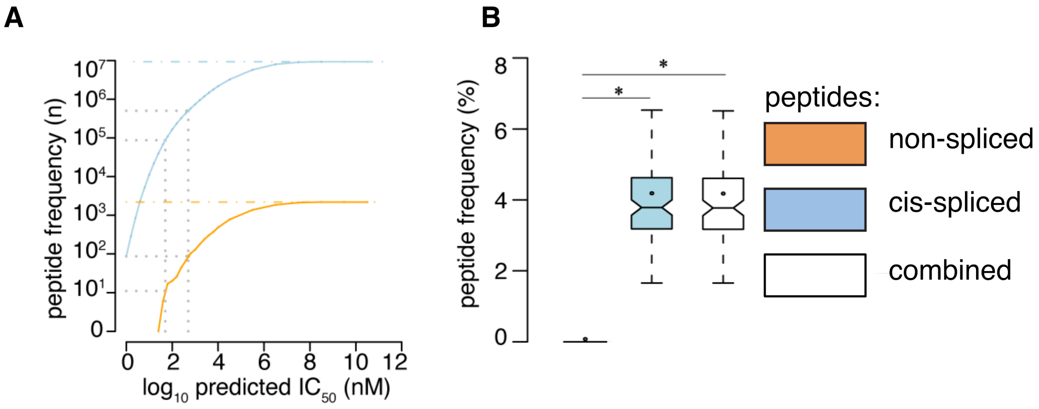


FIGURE 18: Frequency of viral-human zwitter binders depending on IC50 cut-off

(A) Number of viral-human zwitter non-spliced and cis-spliced epitope candidates depending on the peptide-HLA-A*02:01predicted IC50. Gray dot lines mark the predicted IC50 of 500 nM and 50 nM. The blue and orange dot lines depict the number of viral-human zwitter cis-spliced and non-spliced peptide without peptide-HLA-A02:01predicted IC50 cut-off. (B) Frequency of HLA-A*02:01-restricted (predicted IC50 <= 50 nM) viral-human zwitter 9mer epitope candidates, compared to their cognate viral peptide database and considering the whole human proteome database. White color denotes zwitter combined peptide. Significant difference between groups are labelled with * (p-value < 0.05).

3.3.3 Example of T Cell Tolerance Against Viral-Human zwitter Epitope Candidate

In order to demonstrate that the sequence overlap with self could serve as an explanation of the lack of immunogenicity, we obtained a list of HIV-derived HLA-A*02:01-restricted 9mer peptides, which were shown to fail to elicit a cytotoxic response and have a binding affinity of <= 500nM (Table 1). We then applied our pipeline to determine whether any of those experimentally verified epitopes could also be identical to human peptides as either non-spliced or cis-spliced peptide. With these restrictions, we found that the peptide QLAEVVQKV (IC50 = 50 nM) derived Gag-Pol polyprotein of the HIV strain MVP5180 (Gag-Pol$_{955-963}$) could also originate from the Major Vault protein as a cis-spliced peptide MVP$_{786-790/762-765}$ [QLAE][VVQKV] with intervening sequence smaller than 26 AAs (Figure 19). This may explain why this peptide failed to elicit an immunogenic PMBC response in any of the HIV patients [111]. What's more, MVPs was shown to be expressed in mTECs in both gene expression data-sets that we utilized [101, 102]. This may suggest that this peptide is presented on the surface of mTECs which then prunes CD8+ T-cells that could recognize it thus explaining a lack of the response to this epitope in HIV patients. 25 AAs is a rather strict cut-off for the intervening sequence length and upon removing the length restriction, we further identified six additional viral non-spliced peptides which sequence can be recapitulated in human (Table 1).

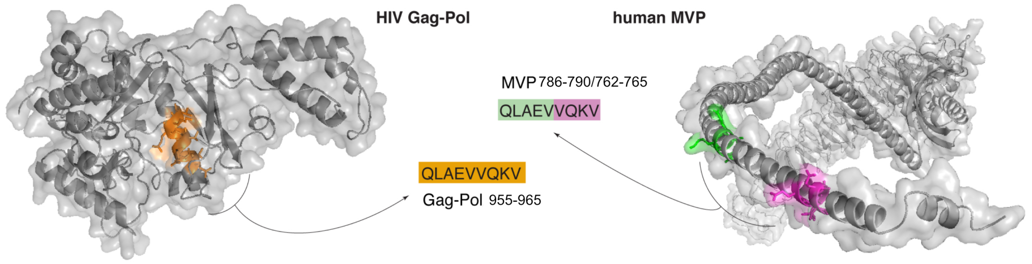


FIGURE 19: Viral-human zwitter epitope candidqte example
Example of HIV-human zwitter epitope candidate QLAEVVQKV, which may be derived from HIV Gag-Pol as non-spliced peptide, and from the human MVP as cis-spliced peptide. Both peptides are depicted in the cognate antigens

Table 1: List of HIV-derived HLA-A*02:01-restricted non-immunogenic 9mer peptides and their zwitter peptide pair

Peptide	IC50 (nM)	Rank	Ref	cis-spliced 25 intervening sequence	any cis-spliced
WLWYIKIFI	24.1	0.5	[112]		
MLQLTVWGI	34.8	0.6	[111]		
LTFGWCFEL	43.5	0.8	[111]		
SITNWLWYI	44.9	0.8	[111]		

LLNATAIAV	50.7	0.9	[112]		Q9P273|TEN3_HUMAN.1504-1506/1405-1410
QLAEVVQKV	50.5	0.9	[111]	Q14764|MVP_HUMAN.786-790/762-765	Q14764|MVP_HUMAN.786-790/762-765
ALQDSGLEV	56.3	1.1	[111]		Q13263|TIF1B_HUMAN.655-658/601-605
ALQDSGSEV	90.5	1.5	[111]		sp|Q8IZJ1|UNC5B_HUMAN.458-462/31-34
LLQYWSQEL	87.9	1.5	[112]		
IVGAETFYV	93.9	1.6	[112]		
QMHEDVISL	93.9	1.6	[112]		
QLQARILAV	109.4	1.8	[112]		Q9P2M7|CING_HUMAN.1138-1143/660-662
HLEGKIILV	150.6	2.3	[111]		
RMYSPISIL	162.9	2.3	[112]		Q9P225|DYH2_HUMAN.1840-1843/3935-3939
HLEGKVILV	177.4	2.5	[112]		Q8N2C7|UNC80_HUMAN.264-267/2799-2803
EMMTACQGV	210.4	2.9	[111]		
TLQEQIAWM	259.4	3.3	[111]		
FLQSRPEPT	371.6	4.1	[112]		
MTNNPPIPV	427.6	4.4	[87]		
QLTEVVQKI	424.7	4.4	[111]		

List of 9mer non-spliced peptides derived from various strains of HIV and predicted to bind HLA-A*02:01 complex with an IC50 500 nM. These peptides also failed to trigger a specific CD8+ T cell response in HIV-infected donors. The corresponding prediction of the peptide-HLA-A*02:01 binding affinity is reported as IC50 and rank, and it was computed by applying SMM algorithm. The potential human origin of the same sequences through peptide splicing by allowing either only cis-spliced peptides with intervening sequence 25 amino acid residues (cis-spliced 25, int. seq.) or any cis-spliced peptides is described through the UniprotKBs protein code and their location within the antigen.

3.3.4 Estimation of viral-human zwitter peptides considering the potential antigen repertoire of human mTECs

During the process of thymic negative selection selection, TCRs that can recognize pMHC complexes with reasonably high affinities are selected while TCRs with abnormally high affinities which can result in strong autoimmune response are removed [83]. This ideally would result in a repertoire of TCRs capable of responding to both invading pathogens and tumor neoantigens while avoiding autoimmunity all together. The antigen presenting cells residing in medulla of thymus are known as medullary thymic epithelial cells (mTECs) and are the drivers of the negative selection process. In principle, they would be expected to display peptides derived from the entire human proteome to achieve full protection from autoimmunity. However, the extent to which different antigens are presented is different. We reasoned that if a protein coding gene has a high expression in mTECs, it would lead to higher protein abundance and subsequently would have a higher probability of being represented on the cell surface. If a virus has zwitter peptides that could be derived from such antigen, those viruses would be more likely to avoid immune detection because it would be more likely that immune system was tolerized to such epitopes. Weinzier *et al*. assessed the correlation between levels of gene expression and the amount of pMHCs on the cell surface. in renal cell carcinomas and healthy kidney tissues [114]. The HLA-presented peptides were eluted from the cell surface and then peptides from carcinomas were isotopically labelled. Gene expression levels were analysed via microarray. Following that, a quantitative comparison was performed between peptides and source mRNAs. The authors observed a weak correlation between the two (r = 0.32). Later, Pearson *et al*. investigated which self antigens could generate MHC-I presented peptides and what protein features could predict the efficiency of presentation [115]. One of the aspects of the work was to determine whether gene expression of source antigens could explain their ability to generate epitopes. Indeed, RNA-sequencing analysis showed that the average gene expression was significantly higher in antigens that were sources of MHC-I epitopes. However, in some cases high expressing protein coding genes produced no epitopes or vice versa [115]. Thus, despite the fact that gene expression does not mirror the HLA-I immunopeptidomes, it appears, to some extent, to be a predictor of antigen presentation.

We thus considered the probability of presentation of the zwitter epitopes in the medulla of the thymus and subsequent deletion of autoreactive TCRs and repeated our analysis weighing gene expression values of human mTECs. To do this, we obtained Microarray gene expression values of human mTECs and single cell RNA sequencing data from mTEC progenitors in human embryos [101, 102]. For scRNA-seq data despite being more sensitive technique for estimation of expression, the cells come from early developmental stage prior to bulk formation of mature TECs. Due to the differences in data acquisition and the developmental stage at which cells were collected the direct comparison of data from Microarray and RNA-seq is difficult.

Since gene expression is to some extent a predictor of antigen presentation, we transformed gene expression values to probabilities of antigens being presented in HLA-I immunopeptidomes of mTECs which we used to sample zwitter peptides. We then re-computed the number of zwitter peptides based on those probabilities. Compared to the whole human proteome, incorporation of potential antigen repertoire based on mTEC transcriptome resulted in a decreased average number of both zwitter non-spliced and cis-spliced peptides. On average, 0.04% and 2.53% of the pool of HLA-A*02:01-restricted virus non-spliced and cis-spliced 9mer peptides, respectively, were zwitter peptides using Pintos RNA sequencing database (Figure 20A). Somewhat higher decrease was obtained using Zengs RNA sequencing database - 0.039% and 1.972%, respectively (Figure 20B).

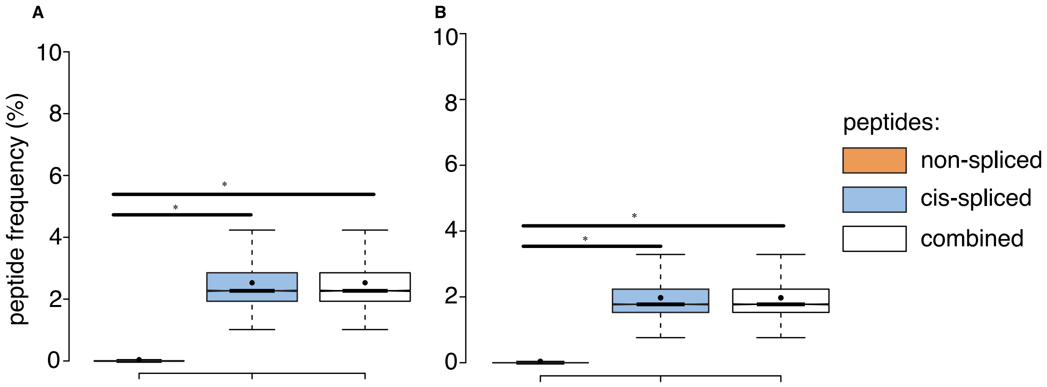


FIGURE 20: Viral-human zwitter peptides based on mTEC gene expression.

(A, B) Frequency of viral-human 9mers (non-spliced, cis-spliced and combined) HLA-A*02:01-restricted zwitter peptides compared to their cognate viral peptide databases considering the human mTEC transcriptome computed either (A) from Pinto *et al.* [101] or (B) from Zeng *et al.* [102]. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Dots represent the mean. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides. Significant difference between groups are labelled with * (p-value < 0.05).

3.3.5 Estimation of viral-human zwitter peptides weighing cis-PCPS frequency

The analysis done so far did not factor in how many of the theoretical possible non-spliced and cis-spliced peptides are generated by the proteasomes and presented on the cell surface and what the frequencies of non-spliced and cis-spliced peptides in HLA-I immunopeptidomes are. We wanted to know how those factors would affect the actual size of T-cell holes that could realistically be expected. The analysis of *in vitro* digestions of synthetic polypeptides by proteasomes showed that despite the fact they can in principle cleave and ligate after any amino-acid, these reactions are non-random and driven by a number of factors such as amino acid and sequence preferences displayed by different catalytically active beta subunits. It was shown in the *in vitro* digestions that cis-spliced peptides are produced in significantly smaller amount than non-spliced peptides by proteasomes [27, 63, 72]. Therefore, only small numbers of non-spliced and and even smaller of spliced peptides are likely generated by proteasomes in large enough quantities and go on to pass through the subsequent APP selection steps to be presented on the cell surface by MHCI molecules. For example, in *in vitro* digestions of a variety of synthetic peptides by 20S proteasomes, we assessed that roughly 25-30% of all possible non-spliced peptides of any length are produced [17]. Among spliced peptides an even smaller fraction is produced, below 1%. An ever smaller fraction of those produced non-spliced and spliced peptides will then be presented on a cell surface by MHCI molecules.

In order to assess the effect of proteasomal generation and presentation on the number of zwitter peptides, we integrated information from two data-sets - a large database of non-spliced and spliced peptides produced *in vitro* by purified proteasomes and HLA-I immunopeptidome elutions from human cell lines.

Different frequencies of spliced peptides in the immunopeptidome were reported [13, 14, 15, 21, 70, 73, 74]. Liepe *et al.* utilized an approach based on the computation of all theoretically possible cis-spliced and non-spliced peptides that could be derived from a given antigen and then performed matching of the experimental with the resulting theoretical MS/MS spectra. They estimated that cis-spliced peptides constituted roughly one third of the diversity of the HLA-I imunnopeptidomes of a variety of cancer cell lines [13, 14]. Faridi *et al.* used *de novo* sequencing in combination with their in house algorithm for determining the origins of the identified peptides and arrived to an estimate of 25% HLA-I immunopeptidomes on average being comprised of spliced peptides. Interestingly, a large fraction of those identified spliced peptides could be explained by trans-splicing according to Faridi *et al* [70]. They used this approach in a follow-up study to characterize the immunopeptidomes of melanoma cells which resulted in a more modest estimate of the contribution of spliced peptides into the immunopeptidome of 6% [15]. Likewise, Mylonas *et al.* employed a denovo sequencing approach paired with an alignment tool for characterization of *de novo* sequences as cis-spliced peptides and a variety of data-base search tools to confirm the identity of the high quality *de novo* sequences. According to their estimate, on the contrary the frequency of cis-spliced peptides in the immunopeptidome was only 2-6% or less [73]. Rolfs *et al.* developed an algorithm called New-Fusion for the spliced peptide identification which is based on separate searching of the experimental spectra against two separate datasets containing only N-terminal or C-terminal ions to identify N- and C-terminal splice-reactants, followed by *in silico* ligation of the two parent peptides. They employed it for the profiling of tandem mass spectrometry data from HLA-I, HLA-II immunopeptidomes and trypsin digested proteins from mouse beta islets. Their estimates were similarly low to Mylonas *et al.* - 1-4% of the peptides were assigned as spliced [74]. Paes *et al.* attempted to determine the contribution of cis-spliced peptides into HLA-I immunopeptidomes of HIV-infected cells. Their approach was based on *de novo* sequencing of high quality tandem MS spectra, similarly to Mylonas *et al.* and Rolfs *et al.* followed by *in silico* splitting of candidate cis-spliced peptides into fragments and matching of those fragments to the proteome separately. Like the two aforementioned studies, their conclusion was that spliced peptides constitute only 2-5% of HLA-I immunopeptidome [21].

Recently, we have generated a large database on *in vitro* digestions of 47 synthetic peptides by a variety of standard- and immune-proteasomes (both core

20S particle and complete 26S) in different experimental conditions[17]. The data we accumulated in the database could be used to estimate the fractions of non-spliced peptides with reasonable accuracy. In fact, in the past, we and others demonstrated a concordance between *in vitro* experiments carried out with purified 20S proteasomes and *in cellulo* and *in vivo* experiments were demonstrated in various studies investigating both viral and tumor epitopes [2, 10, 11, 12, 20, 64, 71, 72, 105, 106, 107, 108, 109, 110]. For example, multiple groups combined cellular (*e.g.* transfection of cancer cells with plasmids encoding the source proteins for the peptide of interest followed by CTL recognition assays) and *in vitro* approaches (*e.g.* proteasomal digestions of the precursor peptides sometimes followed by the loading of the purified peptides onto HLA molecules to test for CTL responses). This was done to independently validate the generation of the peptides of interest as well as their presentation on the cell surface as well as to compare peptides' production by different proteasome types [10, 11, 64, 105, 107, 108, 109, 110].

We estimated that 27% of all theoretical non-spliced 9mer peptides that could be produced by proteasomes are generated in a detectable amount. The frequency of cis-spliced peptides in HLA-I immunopeptidomes, on the other hand, is a contentious topic and depending on the algorithm of peptide identification a large range of frequencies of spliced peptides in HLA-I immunopeptidomes were reported - between 1% and 34% [13, 14, 15, 21, 70, 73, 74].

Based on the reported frequencies of spliced peptides, we could determine the fractions of produced cis-spliced peptides compared to all cis-spliced peptides and finally assess what the expected number of zwtitter peptides would be. Using a range of frequencies of presented spliced peptides and fraction of generated non-spliced peptides we calculated the fractions of observed spliced peptides in viral and human peptides. We then randomly selected peptides 600 times for statistical power, based on those fractions and re-assessed the number of zwitter peptides.

If we assumed a 15% cis spliced peptide frequency in HLA-I immunopeptidomes, over all 600 randomly drawn samples, we identified, on average of sampling, a total of 7 HLA-A*02:01-restricted viral-human zwitter non-spliced 9mer peptides (Figure 21A) which correspond to 0.00406% of all of HLA-A*02:01-restricted virus non-spliced 9mers. On average of sampling, 6 viruses had HLA-A*02:01-restricted viral-human zwitter non-spliced 9mer peptides and no more than 5 peptides per virus were estimated. When it comes to cis-spliced peptides, we identified, on average, a total of 0.3 HLA-A*02:01-restricted viral-human zwitter cis-spliced 9mer peptides. They correspond to 0.00000262% of the pool of HLA-A*02:01-restricted virus cis- spliced 9mer peptides. On average, of sampling, only 1 virus had HLA-A*02:01-restricted viral-human zwitter cis-spliced 9mer peptides and no more than 2 peptides per virus were estimated.

We then repeated the analysis considering a wide range of reported frequencies of cis-spliced peptides in the HLAI immunopeptidome (Figure 21B). Overall, regardless of the frequency of sampled spliced peptides and despite repeating sampling multiple times, only the minority of sampled peptides were found to be zwitter. When zwitter peptides were encountered, only a handful of peptides were zwitter and the majority of zwitter peptides were non-spliced in absolute numbers, compared to spliced peptides regardless of considered frequencies of spliced peptides. With the increase in frequency, the contribution of spliced peptides into the overall pool of zwitter peptides remained similarly small. In absolute terms, even among outlier values with few exceptions, there is almost no increase in number of zwitter peptides as the frequency of spliced peptides increases.

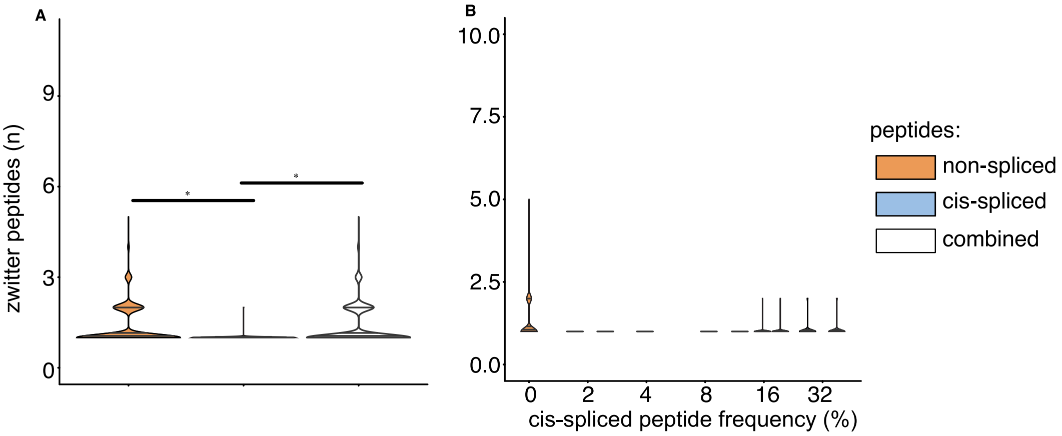


FIGURE 21: Viral-human zwitter peptides considering cis-spliced peptide frequency in HLA-I immunopeptidomes.

Charts are presented as violin plots to show probability densities and to highlight outlier values. (A) Distribution of the number of viral-human 9mer HLA-A*02:01-restricted (non-spliced, cis-spliced and combined) zwitter peptides per virus across all 600 random samples presented as violin plots (rotated densities). Significant difference between groups are labelled with * (p-value < 0.05). This analysis was carried out by hypothesizing that cis-spliced peptides represent 15% of peptides in HLA-I immunopeptidomes and by using the whole human proteome as database. The distribution of the number of viral-human zwitter cis-spliced peptides has been displayed among viruses that had at least one zwitter peptide. (B) Number of HLA- A*02:01-restricted viral-human zwitter non-spliced and cis-spliced 9mer peptides per virus per sampling iteration, depending on a broad range of theoretical cis-spliced peptide frequencies in HLA-I immunopeptidomes. Here, viral proteomes are compared to the whole human proteome database. The number of viral-human zwitter non-spliced and cis-spliced 9mer peptides per virus per iteration has been computed among viruses that had at least one zwitter peptide. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides. Significant difference between groups are labelled with * (p-value < 0.05). Horizontal lines represent 1st (25th percentile), 2nd (50th percentile) and 3rd (75th percentile) quartiles.

Moreover, most of the zwitter HLA-A*02:01 binders are similarly non-spliced and only a small number of spliced peptides are zwitter. This is however expected considering that even the maximum frequency of spliced peptides results in a very small fractions of produced spliced peptides. This is in contrast to non-spliced peptides where almost one third would be expected to be produced by proteasome and presented on the cell surface.

To additionally determine how the increase in frequency of spliced peptides in the immunopeptidome would affect the number of zwitter peptides, we counted how many viruses had non-zero number of zwitter peptides. Based on HLA-A*02:01 binders only 5% of viruses on average had zwitter peptides with the exception of the frequencies of cis-spliced peptides larger than 30% (Figure 22). The share of viruses with non-zero number of zwitter peptides based on all 9mers was, as expected significantly higher.

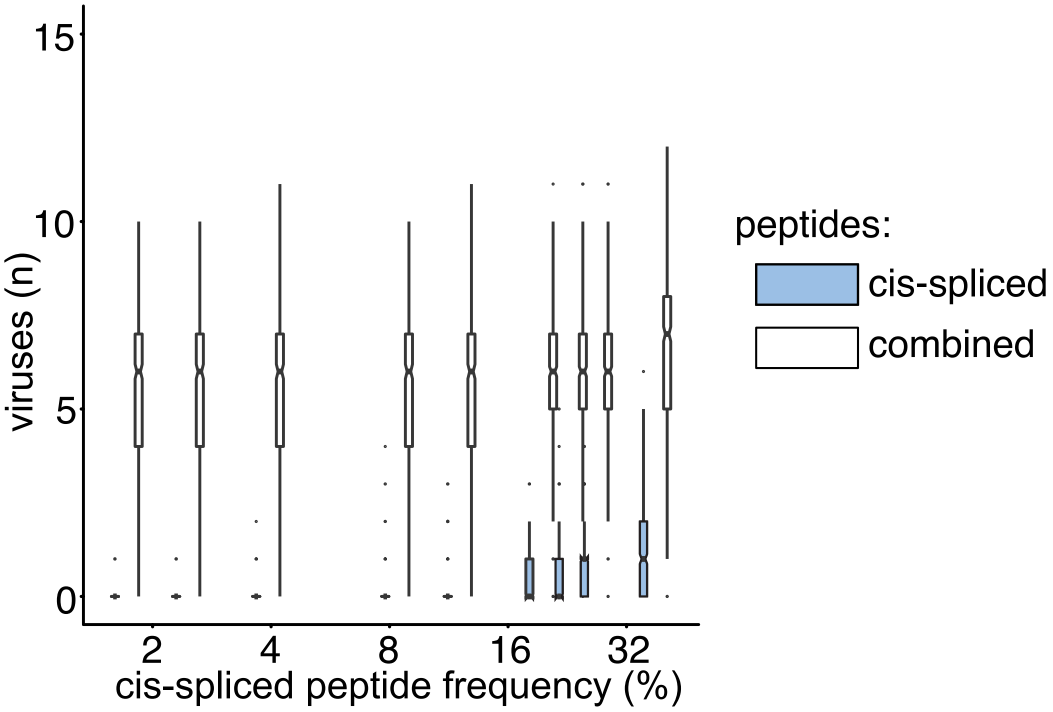


FIGURE 22: Viruses that have zwittter peptides depending on cis-spliced peptide frequency in HLA-I immunopeptidomes.

Average number of viruses that contain at least one HLA-A*02:01-restricted viral-human zwitter 9mer peptide per iteration, depending on a broad range of theoretical cis-spliced peptide frequencies in HLA-I immunopeptidomes. Viral proteomes are compared to the whole human proteome database. The boxplots of the combined peptides have been slightly shifted on the x axis for representation purpose. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides.

Thus, based on the expected frequencies of produced and HLA-A*02:01 presented 9mers, the number of zwitter peptides drops considerably with contribution of spliced peptides into the T-cell holes being minor at most.

There are various factors that can predict the number of potential viral-human zwitter peptides for each given virus. One such factor is the total number of amino acids in the viral proteomes. We observed a strong correlation between the number of viral-human zwitter peptides and the size of virus proteome databases when we considered the entire human proteome (Table S3). The correlation was however weak when taking the frequencies of cis-spliced peptide in HLA-I immunopeptidomes into account (Table S3). Different peptide sequences have binding preferences to different HLA molecules dpending on their sequence motifs and thus if there is a lack of such motifs in the viral sequences, their presentation by for example of HLA-A*02:01 will not be favoured (Figure 23). Such is the case of the Hepatitis delta virus I, which is an outlier in terms of the number of viral-human zwitter HAL-A*02:01 restricted peptides when leveraged against the total number of its theoretical viral-human zwitter peptides.

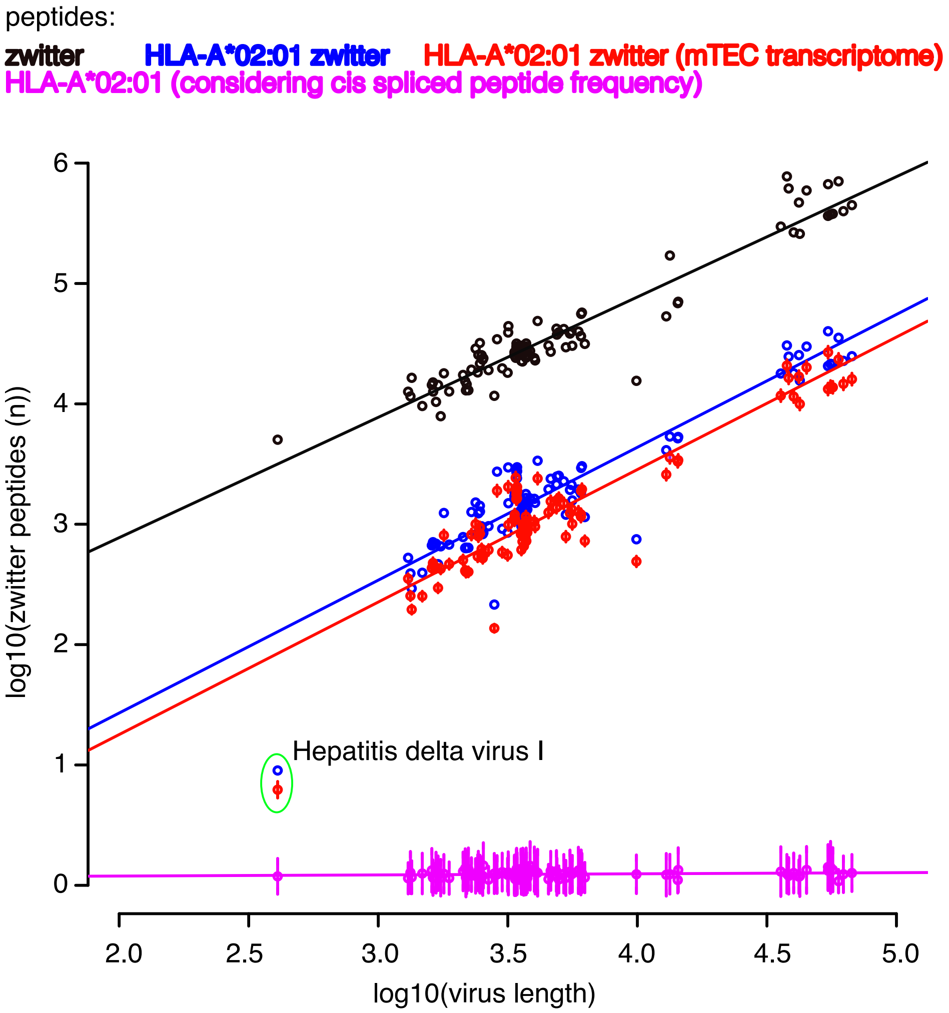


FIGURE 23: Viral-human zwitter peptide frequency depends on virus length and sequence motifs.

Number of viral-human zwitter combined (*i.e.* non-spliced + cis spliced peptides) 9mer peptides per virus, depending on the number of amino acid residues in its proteome. For the groups labeled in pink, we considered a cis-spliced peptide frequency of 15%, as in Figure 4A. Viral-human zwitter 9mer peptides, HLA-A*02:01-restricted viral-human zwitter 9mer peptides are represented with a dot each virus. HLA-A*02:01-restricted viral-human zwitter 9mer peptides either using mTECs RNA-based proteome database (Pinto, 2013) or considering the theoretical cis-spliced peptide frequency in HLA-I immunopeptidomes are represented with a dot (mean) and bars (SD) of sampling iterations. Regression lines are shown. The Hepatitis delta virus I has an underrepresented number of HLA-A*02:01-restricted viral-human zwitter 9mer peptides, which are here labeled.

3.4 Discussion

In light of the discovery that spliced peptides could represent a significant portion of HLAI immunopeptidomes their relevance for the adaptive immune response became a distinct possibility. One important question associated with this potentially huge diversity of epitopes, is the role of spliced peptide in the potential immune evasion by pathogens as a result of molecular mimicry and a removal of CD8+ T-cell self-reactive to self but potentially also reactive to non-self epitopes. Previously, Calis J. *et al.*, showed that 9mer peptides contain sufficient information for the immune system to distinguish self from non-self [24]. This however quickly changed when they considered TCR cross-reactivity. Even according to their conservative estimate of cross-reactivity there was a substantial increase in the potential number of zwitter peptides. Here, we extended this study by investigating the impact of proteasome catalyzed peptide splicing on the number of zwitter peptides in a set of human viruses. We show that even when we consider peptides as zwitter when there is an exact match in all positions of human and viral 9mers, the theoretical number of zwitter peptides increases substantially based on both all 9mers and HLA-A*02:01 binders only.

Zwitter peptides are more likely to be HLA binders than non-zwitter peptides suggesting that zwitter peptides could contain sequence motifs and amino acids that make them more likely HLA-A*02:01 binders compared to non-zwitter peptides. It will be interesting to investigate whether similarly high degree of theoretical zwitter peptides would be observed for other HLA molecules and how their sequence characteristics would impact the number of zwitter peptides and to assess amino-acid distributions in zwitter and non-zwitter peptides.

It will also be interesting to determine the amino acid distributions among viral and human proteomes and to compare amino-acid distributions of zwitter viral non-spliced and spliced peptides to investigate if spliced peptides are enriched in amino acid residues such as Leucine and Isoleucine that are known to be preferred anchor residues for HLA-A*02:01 molecules. This could in part explain the higher prevalence of zwitter viral 9mers being HLA-A*02:01 binders.

For this reason, similar share of viral zwitter peptides based on MHCI binders may not be observed for other HLAI haplotypes in case sequence characteristics of zwitter viral 9mers do not lend themselves to being good HLAI binders. In addition, the more specific HLAI molecules is, the less self-peptides would be presented and thus the chance of encountering zwitter peptide would be smaller. In addition, it is known that some HLAI molecules primarily present pathogen derived peptides.This warrants further investigation.

It would be interesting to investigate if longer than average peptides could aid in viral immune evasion. Peptides longer than 8-14 amino acids can be presented by MHCI molecules [116]. For example, it is known that N-terminally extended precursors are usually processed by ER resident N-aminopeptidases but can sometimes be presented by MHCI. We however don't expect that this would affect the number of zwitter peptides because the main information for the interaction with TCR is present in the middle six positions and TCR don't always use the entirety of those residues.

One interesting avenue of research are the rules that determine which epitopes are more likely to be more immunogenic than others. These are thought to be determined by both stability of pMHC complexes and efficient recognition of pMHC by TCR and their strong interaction. It is likely driven by both physical and chemical properties of presented epitopes and TCRs. It is known that there are stretches in TCR sequences that are shared between multiple individuals suggesting that there is a selection process for certain features of TCR that enhance their recognition of microbial and tumor epitopes. In one, Calis J., *et al* set out to determine what characteristics of peptides presented by MHCI molecules are predictive of their immunogenicity [52]. The authors used peptide-immunization studies in mice and humans from a variety of sources to determine which complexes of 9mers and MHCI (HLA restricted in humans and H-2 restricted in mice) were recognized. They discovered that large, aromatic and to an extent, acidic residues in the middle positions of the 9mers had the highest association with epitopes immunogenicity. Interestingly, different HLA haplotypes were enriched for similar amino-acids in immunogenic peptides. Next, they calculated position depended differences between immunogenic and non-immunogenic peptides. Based on these observations the model for prediction of immunogenicity was proposed. According to the model, human epitopes had substantially higher immunogenicity score than non-epitopes. In terms of enrichment of amino-acid residues, human immunogenic peptides were similar to mice immunogenic peptides. Additionally, viral epitopes tended to be more immunogenic compared to human epitopes.

Recently, Ogishi M. and Yotsuyanagi H. have proposed an alternative framework for analysis of immunogenicity based on thermodynamic properties of interactions of pMHCs with TCR in combination with epitope sequence characteristics that reportedly achieved higher accuracy in immunogenicity prediction [87].

Understanding the rules for immunogenicity could improve the discrimination between immunogenic and non-immunogenic epitopes as well as between zwitter and non-zwitter epitopes. The zwitter peptides would be expected to be less immunogenic. Utilizing existing immunogenicity models, it will be interesting to see if there are differences in expected immunogenicity between zwitter and non-zwitter viral peptides and if zwitter peptides tend to be less immunogenic. Additionally, this assessment could show if on the contrary zwitter epitopes derived from viruses implicated in autoimmunity have a strong predicted immunogenicity.

To further extend our estimates, we considered how the number of zwitter peptide would be affected by rates of presentation in the thymus and actual shares of generated peptides by proteasome. Different antigens have unequal abundances in mTECs which is linked with unequal presentation on the cell surface. This was evident in both of the datasets we considered. This led to drop in theoretical number of zwitter for both datasets suggesting that different levels of production could have an impact on the share of zwitter peptides that would be expected to be presented. A subset of peptides that would otherwise be considered zwitter, would than never be presented or presented in sufficient quantities to induce tolerance. Our estimates here are imperfect as we were only relying in transcriptomics data. It's known that level of expression of a given gene doesn't necessarily mean that the product of this gene would have high protein level [114, 115]. Second, the only data for mature human mTECs that was available to us, was obtained with microarray which is an outdated and an imperfect technique for an assessment of expression levels due to limited dynamic range, potential abundance of non-specific signals and low sensitivity for genes with low level of expression. RNA-seq doesn't have those disadvantages. However, The RNA-seq data that we used was derived from progenitor TECs which can't be used as a direct estimate of the set of antigens presented in mature mTECs.

The second crucial factor, was the fraction of produced peptides which would subsequently be presented on the cell surface.

At the frequencies of non-spliced and spliced peptides that can be theoretically expected to be produced, it seems that the number of zwitter peptides decreases drastically, both spliced and non-spliced. Despite the substantial degree of theoretical sequence overlap of cis-spliced peptides, they don't appear to play a significant role in the formation of T-cell holes according to the mass spectrometry measurements of non-spliced and cis-spliced peptides produced *in vitro* by proteasomes and detected in immuno-peptidomes of human cells. Cis-spliced peptides would play a significant role in shaping the CD8+ T-cell repertoire only if there are large frequencies of cis-spliced peptides in HLAI immunopeptidomes.This has to do with the fact that out of all theoretically possible cis-spliced peptides only a minor fraction will be generated iby proteasome in sufficient quantity and even less will pass the steps in antigen presentation pathway. Spliced peptides are also known to be produced in significantly smaller abundance than non-spliced peptides.

This suggests that while the zwitter peptides could potentially be utilized by viruses provided high frequencies in HLAI immunopeptidomes, its actual role in the immune evasion remains unclear and likely varies from virus to virus. It's however important to keep in mind that in the estimates of fractions of produced non-spliced peptides in the *in vitro* digestions, fractions of spliced peptides and frequencies of spliced peptides in the immunopeptidome, we were not taking the abundances of individual peptides into account and focused strictly on qualitative aspect (diversity of the peptides). Quantity of a generated peptide is one of the deciding factors in antigen presentation. This limitation at large applies to our theoretical estimate of zwitter peptides based on all 9mers. It is notable that we restricted our computations only to cis-spliced peptides with the intervening sequence length of 25 AAs or less and could thus underestimate the frequencies of cis-spliced peptides in the immunopeptidomes.

So far, we focused on the immune evasion of viruses due to the theoretically high number of zwitter peptides. Despite the low expected number of zwitter peptides present in the immunopeptidome owing to the observed frequencies of spliced peptides, the risks of autoimmune response can't be completely disregarded. Presence of even small numbers of viral-human zwitter peptides, which could potentially be cross-reactive thus poses risk for autoimmunity. This could be an issue if a given epitope derived from self is not presented on the cell surface, not presented in sufficient amount to induce tolerance or if there are defects in tolerance induction. In this case, autoreactive T cells would survive negative selection and persist in the pool of circulating T cells. Subsequently, upon viral infection, if similar epitopes end up being displayed on the cell surface this could lead to a strong immune response to hosts own tissues. Some viruses are known to be the possible cause of a range of autoimmune diseases such as Epstein-Barr Virus (EPV) that can cause multiple sclerosis (MS), cytomegalovirus (CMV) that can induce systemic lupus erythematosus (SLE) and enteroviruses that are associated with Type 1 Diabetes (T1D) which we investigated in detail [26].

The TCR degeneracy could on the other hand have a significant impact on the number of zwitter viral-human cis-spliced peptides. Previously, it was shown that a substantial fraction of non-spliced peptides could be indistinguishable from self from TCRs point of view [24]. We investigated this matter in detail (see Chapter 4).

Moreover, driving forces of PCPS such as peptide sequence motifs and amino-acid preferences will have to be taken into the account when assessing the frequencies of the possible viral-human cis-spliced peptides. Peptides sequence motifs were shown to affect both the proteasomal dynamics and the composition and quantities of non-spliced and spliced peptides [27, 63]. Toes *et al.* performed quantitative analysis of cleavage motifs utilized by standard and immuno-proteasomes in the digestion of enolase-1 and discovered that amino acids in both flanking and distant positions relative to the cleavage sites were distributed non-randomly and were instrumental in determining hydrolysis specificity and thus the quantity of generated peptide [117]. In 2012, Mishto *et al.* demonstrated that the quality and quantity of spliced peptides produced by the proteasome were determined by the sequence preferences distinct from those exhibited in non-spliced peptides generation [27]. Liepe *et al.* studied the differences in cleavage site usage between standard and immuno-proteasomes by performing *in vitro* digestions of four synthetic polypeptides and further corroborated that cleavage-site usage was not only determined by the amino acid forming the peptide bond being cleaved but also by their surrounding residues [65]. They also combined experiments and mathematical modelling to simulate peptide-bond hydrolysis and among other things showed that proteasomal dynamics, *i.e.* the frequency of peptide-bond cleavage was dependent on the particular sequence motifs in the time-series of the *in vitro* digestions of a variety of substrates by purified proteasomes [66]. Examination of the peptide products produced in a variety of *in vitro* digestions presented in our recently published data-base shows that different polypeptide sequences result in drastically different relative frequencies of non-spliced and cis-spliced peptides [17]. Moreover length distributions of cleavage and spliced peptides vary drastically between different substrates. Finally, even single amino acid substitutions could impact the proteasomal dynamics. Liepe *et al.* previously showed that single amino acid substitution from Threonine to Methionine in polypeptide derived from gp100 significantly altered the composition of cleavage and spliced peptides produced in the *in vitro* digestions by 20S proteasomes [28]. All of those uncovered factors can reduce the variety of both non-spliced and cis-spliced peptides produced by the proteasome.

Finally, the type of proteasome could affect the frequency of zwitter cis-spliced peptides since different proteasome types (standard-, immuno- and thymo-) have varying amino acid and peptide sequence preferences. Ebstein *et al*. characterised the generation of a spliced eptiope candidate derived from melanoma gp100 and found that the immuno-proteasome derived from spleen catalized the condensation reaction more efficiently than standard proteasome from erythrocytes [2]. Guillaume *et al*. isolated intermediate proteasomes from the lymphoid tissues [118]. They contained either beta5i or beta5i and beta1i subunits of immuno-proteasomes. They found that the presence of one or two beta subunits of the immuno-proteasome had a considerable impact on the quantities of two tumor antigenic peptides recognised by CD8+T-cells. In the follow-up study, Guillaume *et al*. analysed the production of seven tumor epitopes by intermediate proteasomes and standard proteasomes and showed as in their previous study that the efficiency of generation of different epitopes changed drastically depending on the beta subunit composition of the intermediate proteasomes further suggesting the presence of distinct sequence specificieites by different proteasome types [107]. Dalet *et al*. focused on the investigation of the generation of three antigenic spliced peptides known at the time by the standard and the immuno-proteasomes. They showed that two of the spliced peptides were preferentially generated by standard proteasome and one by immuno-proteasome [110]. They hypothesised that the efficiency of spliced peptides' generation was dependent on the efficiency of the production of the N- and C-terminal spliced reactants. Mishto *et al*. performed an in depth analysis of the cleavage site usage by performing *in vitro* digestions of synthetic polypeptides by different purified proteasome isoforms, determining precise quantities of the generated peptides and calculating SCS. They demonstrated that unlike prior observations with less sensitive instruments every single cleavage-site was utilised by all proteasome isoforms [65]. There were, however, substantial quantitative differences in peptides generated by different proteasomes. Next, they combined mathematical modelling and experiments to recapitulate the dynamics of proteasomal cleavage [66]. The key finding was that the rate of the transport of substrate into proteasome's catalytic chamber serves as a rate-limiting step for the rate of degradation and that the differences in the transport efficiency are in large part responsible for the differences in kinetic parameters and the quantities of the products generated by standard and immuno-proteasomes [66]. Kuckelkorn U. *et al*. compared proteasomal dynamics of 20S standard, thymo- and immuno-proteasomes using a combined biochemical and bioinformatics approach. Similarly to standard- and immuno-prtoeasomes there were marked differences in

SCS as well as substrate transport which resulted in quantitative differences in peptide products produced by the different proteasome isoforms [119]. Dianzani *et al.* investigated the dynamics of the proteasomal degradation of Osteopontin (OPN) and its N and C-terminal portions and combined with cell chemotaxis and clinical data of Multiple Sclerosis (MS) patients [120]. There were considerable differences in the processing of OPN by 20S standard and immuno-proteasomes. Mainly, immuno-proteasome was deminstrated to degrade all three versions of OPN significantly faster than the standard proteasome. Fabre *et al.* combined affinity purification with mass spectrometry to profile different proteasome isoforms. They demonstrated that different proteasome isoforms exhibited preferences for the associations with different regulatory subunits which undoubtedly impact the dynamics of substrate's transport and subsequent peptides' generation [121]. All of these uncovered factors about the specifics of the degradation by different proteasomes could have profound effects on the composition of the immunopeptidomes of cells. Faridi *et al.* utilized their bionoformatic pipeline for the identification of MHC-I presented peptides combined with mass spectrometry to characterise the changes in the immunopeptidome of melanoma cells in presence or absence of interferon gamma [23]. They found that IFN-gamma had a substantial impact on the antigenic landscape of the melanoma cells. Only about half of the identified epitopes were shared between treated and untreated cells. This change was in large part attributed to the preferential expression of the immuno-proteasome in the interferon gamma treated cells. Apavaloaei A., *et al.* investigated the role of thymoproteasome in shaping the T-cell repertoire in cortical Thymic Epithelial Cells (cTECs) [122]. Specifically, they examined thymoporteasome specific subunit PSMB11 in mice. Likely thymoproteasome changes the composition of peptides present on the cell surface. However, in addition it was discovered that thymo-proteasome changes the rate of proteolytic destruction or activation of various transcription factors thus changing the patterns of gene expression mainly by repression of the expression of a number of genes. This in turn regulates the maturation and regulation of T-lymphocytes in cTECs and their subsequent localisation to medulla. All of the aforementioned factors could lead to quantitative differences in peptides generation which can impact the repertoire and tolerance of CD8+ T-cells and by extension both the risk of immune evasion and auto-immunity.

It will be prudent to recapitulate the observations we made with the estimates of the numbers of viral-human zwitter peptides of immunogenic and non-immunogenic peptides described above. To fully understand the impact of the zwitter peptides, the immunopeptidome of a variety of cell types infected with different viruses will have to be extensively experimentally investigated and the contribution of zwitter peptides into the immunopeptidome composition will have to be determined.

So far, we can only infer some general trends regarding immune evasion by the viruses. Despite the theoretically large number of viral-human zwitter epitopes the immune system still manages to clear many pathogens and prevents the autoimmunity. This discrepancy requires us to study the connection between the PCPS and the antigen presentation. It is important to remember that the number of spliced zwitter peptides is calculated based on theoretically possible sequences of spliced 9mers which were computed with a number of restrictions. Quantity of a generated peptide is important in antigen presentation. Only a portion of viral spliced and non-spliced peptides revealed to be identical to self-peptides would be relevant for antigen presentation. The determination of specific epitopes responsible for gaps in T cell immunity will require the experimental verification.

Overall, our study provides an additional insight into the potential mechanism of immune evasion by viruses and estimates the impact of peptide splicing on the theoretical number of zwitter peptides. We believe that further research in this direction will be important in vaccine development, identification of uniquely immunogenic targets for immunotherapy which would not be capable of inducing autoimmune response due to being zwitter peptides, and a general understanding of mechanisms of immune evasion.

Chapter 4

The potential impact of T-cell cross-reactivity on the immune evasion through PCPS

4.1 Introduction

CD8+ T-cells have to be able to discriminate between self and non-self antigens and activate only in response to non-self derived epitopes. Ideally and in the most simple case, TCRs would be 100% specific to a particular epitope. However, this is known to not be the case as there exists a certain degree of flexibility of epitopes recognition by TCRs [24]. This is known as T-cell cross-reactivity or T-cell receptor degeneracy. It is thought to be necessary to provide flexibility of recognition and to cover as much sequence space of invading pathogens and emerging tumors as possible with a limited repertoire of T-cells [16, 123, 124, 125, 126, 127, 128]. It was estimated that around 4x10**11 T-cells are present in the peripheral blood in human [129, 130]. The true number of unique TCRs is estimated to be around 10**6-10**8 or at at most 10**12 which is substantially lower than a theoretical 10**15-10**20 breadth of unique TCRs that could be generated by the somatic recombination in human and over 10**15 of unique TCRs in mice [129, 130, 131, 132]. The estimates for the number of unique non-spliced 8mer foreign peptides is in a range of 2x10**10, 9mers - 5.1x10**11, 10mers - 1.2x10**13, 11mers - 2x10**14, 12mers - 4.1x10**15, 13mers - 8.2x10**16, 14mers - 1.6x10**18 in the human proteome [132, 130]. On the other hand, the number of the unique non-spliced 9mers that could be obtained from human proteome is in the range of 10**7 while the number of unique non-spliced 14mers is 10**16 [93, 129]. Thus, in contrast to these large numbers of peptides, a significantly smaller number of TCRs is available compared to the number that would be necessary if only one T-cell receptor would recognize only one epitope - most likely 10**6-10**8 [129, 130, 131, 132]. T-cell cross reactivity is thought to depend on the individual properties of a given TCR as the extent of cross-reactivity varies from T cell to T cell. Not all residues of the displayed peptides are available to TCR. It was shown that the alterations in the epitope that occur outside of the so-called sequence hot spots, most crucial, for the interaction are not sufficient to disrupt the immunogenic response[123, 124, 127].

The extent of cross-reactivity of different TCRs for epitopes is still an ongoing debate. In one study, cross-reactivity of 15 mouse and human CD8+ T-cells was probed against a large set of synthetic peptides from unrelated bacterial and viral antigens considered to be good binders to H-2Kb and H-2Db molecules of mouse and HLA-A*02:01 molecule of human [133]. Out of all tested interactions, only one instance of cross-reactivity was determined for H-2Db for two 9mer peptides sharing only 5th residue in common bringing the estimate of the frequency of TCR cross reactivity for unrelated epitopes to 1/30000. The theoretical estimate for cross-reactivity of T-cells to the two peptides with sequence homology was on the other hand shown to be 1/65000 [133].

Previously, it has been shown that for HLA-I molecules the anchor residues of epitopes are typically at positions 2 and 9 and thus, positions 1,2,9 don't normally participate in the interactions with TCR [134]. In fact, only residues which are oriented outwards and solvent exposed relative to the MHC molecule are utilised for the interactions with TCR. On the other hand, at least a subset of middle six positions are frequently involved in interactions with TCR. It was shown that if amino acid substitutions possess similar chemical and physical properties or if the interacting surface of the epitope on the MHC remains the same, the recognition of such epitope is possible by the same TCR [123, 124, 134].

TCR cross-reactivity would be particularly relevant in context of immune evasion or conversely breaking of tolerance barriers and causing autoimmune diseases. An *in silico* analysis of fungal and bacterial antigens revealed epitopes potentially cross-reactive with CD8+ T-cell clone recognising a naturally processed beta cell epitope HLVEALYLV presented by HLA-A*02:01 allele and implicated in T1D diabetes [135]. All of the top-ranking peptides possessed the same amino-acid in position 4 as the original self-peptide suggesting its key role for the recognition by the isolated T-cell clone. Previously, two distinct epitopes RQFGPDWIVA and MVWGPDPLYV from two bacterial pathogens were shown to be recognised by CD8+ T-cells originally reactive to self insulin derived sequence ALWGPDPAAA [136]. It was suggested that the ability of TCRs to cross-react with different epitopes has to do with the preservations of docking properties TCRs interacting with peptide-MHCIs complexes [137].

In addition to the cross-reactivity to non-spliced epitopes, there were discoveries of several instances of TCR cross-reactivity between non-spliced and cis-spliced peptides with the similar sequences. Several years ago, Mishto *et al.* performed *in vitro* digestions of a polypeptide derived from Listeriolysin O of *Listeria monocytogenes* with purified proteasome in search of epitope candidates presented by H-2Kb molecules of mice and recognized by CD8+ T-cells [71]. Along with the well established linear peptide VAYGRQVYL, a cis-spliced peptide which differed by just one amino-acid was identified - SAYGRQVYL. Both epitopes triggered the cross-reactive response by same population of CD8+ T-cells. Paes *et al.* developed a work-flow for the identification of spliced peptides in HLA-I immunopeptidome and used it to identify a number of spliced peptides presented by HLA-I molecules. Some of those epitopes were partially overlapping with non-spliced peptides known to trigger CD8+ T-cell responses [21]. Those CD8+ T-cells were able to cross-recognise those newly identified epitopes along with non-spliced peptides, likely due to sharing of residues which constitute the primary contact site for the TCRs.

Calis *et al*. hypothesised that their modest assessment of the extent of sequence overlap of 9mer non-self peptides with self when considering complete sequence identity, could be severely underestimated owing to the degeneracy of TCRs of CD8+ T-cell receptors that could lead to a substantially higher immunological overlap between self and non-self. By examining crystal structures of pMHC-TCR complexes, they devised a simple model of TCR cross-reactivity which they then used to re-assess the sequence overlap. They showed that a large fraction of non-spliced viral and bacterial peptides would be expected to indistinguishable from self by TCR of CD8+ T-cells, when TCR cross-reactivity was taken into account [24].

Their degenerate model was in part based on a study by Franklid *et al*l who investigated the extent of TCR cross reactivity [134]. They analyzed the sequence of HIV epitope LFNTVATL and discovered that substitutions in its sequence with similar amino acids had insignificant effect on recognition by TCR. They showed that middle positions appeared to be the most important for TCR recognition. Based on this data, Franklid *et al.* developed a general model for TCR cross-reactivity and demonstrated that non-immunogenic HIV1 peptides appeared to be more similar to self epitopes in terms of amino acid composition [134].

The reason TCR cross-reactivity could be relevant is that during the process of negative selection, if self-epitopes presented on the surface of the antigen presenting cells are recognized by TCRs with high affinity, such TCRs would be removed from the overall pull of T cells to avoid autoimmunity. This is because strong interaction with an epitope could potentially lead to a strong immunogenic response. Due to this cross-reactivity, there would then no longer be TCRs capable or recognising variants of such 9mer. It was hypothesised that if sequences similar to the ones against which T cells were originally neutralised, would originate from the invading pathogens and are displayed by MHCI molecules, they would be no immune response. This would be because there would not be any T cells, capable of recognizing those epitopes. Alternatively, if self reactive T-cells weren't successfully pruned during the negative selection, this could pose a serious risk of autoimmunity.

We aimed to answer the following research questions:

- How does the frequency of viral-human HLA-A*02:01 restricted zwitter non-spliced and cis-spliced peptides change when accounting for TCR cross-reactivity
- How does the combination of TCR degeneracy and the antigen expression in the thymus impact the frequency of viral-human HLA-A*02:01 restricted zwitter peptides

- How would the frequency of viral-human HLA-A*02:01 restricted degenerate zwitter peptides be affected when taking the frequencies of HLA-A*02:01 presented cis-spliced peptides into account

4.2 Materials and methods

4.2.1 Analysis of TCR-pMHC structures

Structures of 9mer pMHCI-TCR complexes were downloaded from PDB database [138] (Table S4). We only considered structures of 9mer-HLA-A*02 complexes with TCR. We counted number of peptide-MHC contact per each position (number of TCR amino-acids within 5 angstrom distance of a given residue of a 9mer). For each structure we calculated the fraction of TCR contacts per position relative to all peptide-TCR contacts.

4.2.2 Estimation of viral-human degenerate zwitter peptides

We computed the degenerate zwitter peptides based on the cross-reactivity of the TCR. It was shown that 9mers presented by MHCI, may tolerate alterations in the middle portion of their sequence with similar amino acids, such that the same TCR could still recognize altered epitope (*i.e.* to be cross-reactive) [123, 124, 125].

Using an approach, similar to the one proposed by Calis *et al.*, we allowed up to two amino acid mismatches in the middle six positions of the 9mer excluding positions 4 and 5, which were shown to have the largest number of contacts with TCRs [24]. Substitutions are allowed for amino acids with similar physical and chemical properties. Substitutions were permitted to occur on either side of residues 4 and 5. We reasoned that if substitutions are sufficiently similar to residues of an original 9mer, the interaction with TCR would not be disrupted. Thus, substitutions are allowed for amino acids with similar physical and chemical properties. Additionally, we allowed any substitutions to occur in position 1 independently from all other substitutions. Position 1 is not typically known to be involved in either MHCI binding or TCR interaction. We did not modify anchor residues 2 and 9, since it is possible that even minor substitutions in those positions could alter or disrupt the binding of the 9mers to MHCI.

We made use of PMBEC matrix which is based on measured binding affinities between various peptides and MHC-I molecules and can be utilized to assess the effect of a given amino acid substitution on peptide protein interaction properties [139]. Covariance of two given amino acids in PMBEC matrix was used as a measure of their similarity. If the absolute covariance of two amino acids was greater than 0.05, such substitutions were allowed to occur in the 9mer in the permitted positions [24].

Prior to computation of sequence variants, we predicted binding affinities of all computed non-spliced and cis-spliced peptides to HLA-A*02:01 molecule. We reasoned that if the original human 9mer which gave rise to such zwitter peptides is itself not an HLA-A*02:01 binder, it would not be presented on the cell surface of antigen presenting cells during negative selection, and thus T cells would not be tolerized against such peptide. As a result, even if there were viral 9mers similar to such peptides, which would be presented on HLA-A*02:01 molecules they would still be recognized by the immune system. Subsequently, such 9mers would not be considered as degenerate zwitter HLA-A*02:01 binders. We subsequently obtained all possible sequence variants that fit the criteria for each human 9mer predicted to be HLA-A*02:01 binders. Following that, we performed sequence matching of all computed variants of human 9mers derived from original HLA-A*02:01 binders to all viral HLA-A*02:01 binders as for the exact overlap (*i.e* complete sequence match of viral and human 9mers). These matched peptides were then used to compute frequency of zwitter peptides in the same manner as for the zwitter peptides (*i.e.* complete sequence match of two given 9mers (see 3.2.1)).

$$dB_v = (dz_b/b_v) * 100$$

Where dz_b - number of degenerate viral-human zwitter 9mers, predicted to be HLA-A*02:01 binders; b_v - number of all possible HLA-A*02:01 binders for a given virus

Similarly to the analysis of zwitter peptides considering the exact match of all 9 positions of viral and human peptides, we analysed the potential impact of gene expression in mTECs on the number of viral human-zwitter peptides utilizing the model described in Chapter 3.

When we considered both predicted peptide-HLA-A*02:01 binding affinity and potential antigen repertoire of mTECs, the viral-human degenerate zwitter peptide frequency was computed as:

$$dMB_v = (dz_{m,b,v}/b_v) * 100$$

Where $dz_{m,b,v}$ - number of sampled degenerate viral-human zwitter 9mers based on mTECs expression values and predicted to be HLA-A*02:01 binders; b_v - number of all possible HLA-A*02:01 binders for a given virus.

Statistical test values for the comparisons between different groups of viral-human zwitter peptides are provided in Table S5 and Table S6.

4.2.3 Estimation of the frequency of viral-human degenerate zwitter peptides weighing PCPS frequency

In order to assess the frequency of viral-human degenerate zwitter peptides based on frequencies of cis-spliced peptides in the immuno-peptidome, we applied the same sampling strategy that we used for viral-human zwitter peptides considering complete sequence homology (see Chapter 3).

For the estimation of sampled degenerate zwitter peptides from the list of sampled human peptides, we predicted which ones were HLA-A*02:01 binders and computed all versions of the binders with one or two substitutions according to the degenerate recognition model. We then computed the all viral-human degenerate zwitter 9mers among the sampled human peptides.

4.3 Results

4.3.1 Estimation of viral-human zwitter peptides

Our prior estimates of the number of viral-human zwitter peptides are based on an assumption that TCR recognition of pMHC is 100% specific, meaning that we considered a viral peptide as zwitter only if it was exactly matching in all 9 positions to the corresponding human 9mer. To get a better idea of a true extent of the sequence overlap of viral and human peptides, we had to determine the viral-human degenerate zwitter peptides due to the cross-reactivity of TCRs.

To account for TCR degeneracy, we opted for a streamlined approach based on the one proposed by Calis *et al.*, which nevertheless doesnt account for all the ways TCR cross-reactivity could be achieved [24]. The model is based on the investigation of the number of contacts between given amino acid residue in a 9mer and T-cell receptor in crystal structures of TCRs with pMHC complexes of a variety of HLAs. The largest number of contacts was observed in positions 4-8. On the other hand, positions 2 and 9 typically serve as anchor residues in HLAI bound peptides and thus, were not considered relevant for the direct interaction with TCR. Position 3 is rarely involved in the interaction with TCRs. Nevertheless, it was demonstrated that it too can be important for the interaction with TCR [24]. Thus, the authors considered positions 3-8 as the primary interaction interface of the presented peptide with TCRs. They reasoned that since position 3 is still within the interacting surface of the epitope, it could be be important for interaction with TCR [24]. According to their model of degeneracy of T-cell recognition, the interaction with a given TCR is still possible if one or two substitutions in middle positions (3-8) of the 9mer occurs. Substitutions were allowed in any of the middle six positions (3-8) with amino acids with similar physical and chemical properties except for position 5, that was shown to have the largest number of contacts with TCR [24].

On our end, we performed similar type of analysis but only for the complexes of peptide-HLA-A*02:01 molecules with TCRs. This was done to check if the rules for interactions proposed by Calis J. *et al* would apply [24]. We observed similar pattern with middle positions 4-8 having the largest number of contacts, on average. In particular, the number of molecular contacts was over-represented for positions 4 and 5, compared to all other positions(Figure 24). Based on our examination of crystal structures we restricted amino acid substitutions to positions 3,6,7 and 8, and additionally allowed any substitutions in position 1 independent from substitutions in the middle positions due to the lack of involvement of this position in TCR interactions (Figure 24).

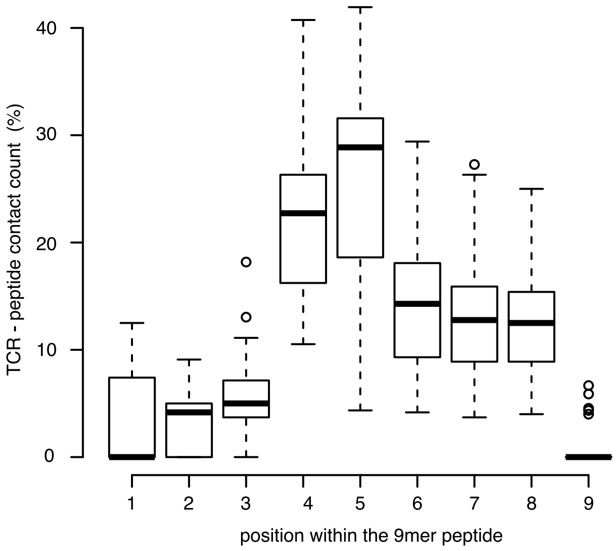
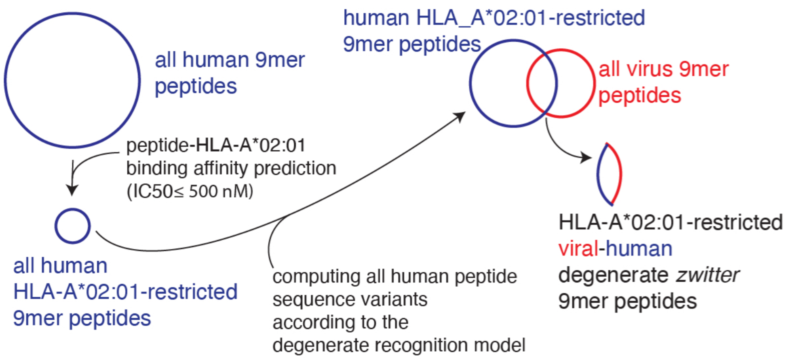


FIGURE 24: Residues of HLA-A*02:01-bound peptides in contact with CD8+ TCR

Degenerate 9mer sequences were computed for Human proteome as during thymic negative selection only epitopes derived from self are presented and cross-reactivity of self-reactive TCRs that are deleted is relevant (Figure 25). Subsequently, non-self peptides that are similar to self wont induce the immune response for further immune evasion.



all human 9mer peptides
peptide-HLA-A*02:01 binding affinity prediction (IC50≤ 500 nM)
all human HLA-A*02:01-restricted 9mer peptides
computing all human peptide sequence variants according to the degenerate recognition model
human HLA_A*02:01-restricted 9mer peptides
all virus 9mer peptides
HLA-A*02:01-restricted viral-human degenerate *zwitter* 9mer peptides

FIGURE 25: An overview of the general approach used to estimate viral-human degenerate zwitter peptides.

Schematic representation of *in silico* pipeline to estimate the frequency of degenerate zwitter peptides predicted to bind HLA-A*02:01 complexes. It is assumed that if a given human peptide is presented, due to the cross-reactivity, TCR specific to such peptide will also be able recognize its sequence variants and thus upon removal during negative selection no TCRs specific to such epitopes will be available which would facilitate immune evasion. Thus, to estimate the degenerate zwitter peptides we only consider human HLA-A*02:01 and their sequence variants within the given parameters. The matching of sequence variants of human HLA-A*02:01 binders with viral 9mers constitutes the degenerate zwitter peptides

The degenerate zwitter peptides were computed based on the binding affinities of the peptides to HLA-A*02:01 molecules. That means that we only only considered the degenerate recognition of the 9mers by TCRs if they were predicted to bind HLA-A*02:01 molecules with IC50 no more than 500 nM. The number of degenerate zwitter HLA-A*02:01 binders relative to all predicted HLA-A*02:01 binders increased considerably for both spliced and non-spliced peptides compared to the numbers of zwitter peptides that we computed based on the complete sequence matches of self and non-self peptides (see Chapter 3). Specifically, when we considered T-cell degeneracy, we identified 8965 and 12150323 theoretical degenerate viral-human zwitter non-spliced and cis-spliced 9mer peptides, respectively. The frequency of degenerate non-spliced zwitter HLA-A*02:01 binders increased from 0.052% to 21.1% on average. The frequency of degenerate zwitter spliced HLA-A*02:01 binders increased to 95.06% on average compared to regular zwitter peptides, suggesting significant contribution of the zwitter peptides into the immune evasion of viruses based on T-cell cross-reactivity (Figure 26).

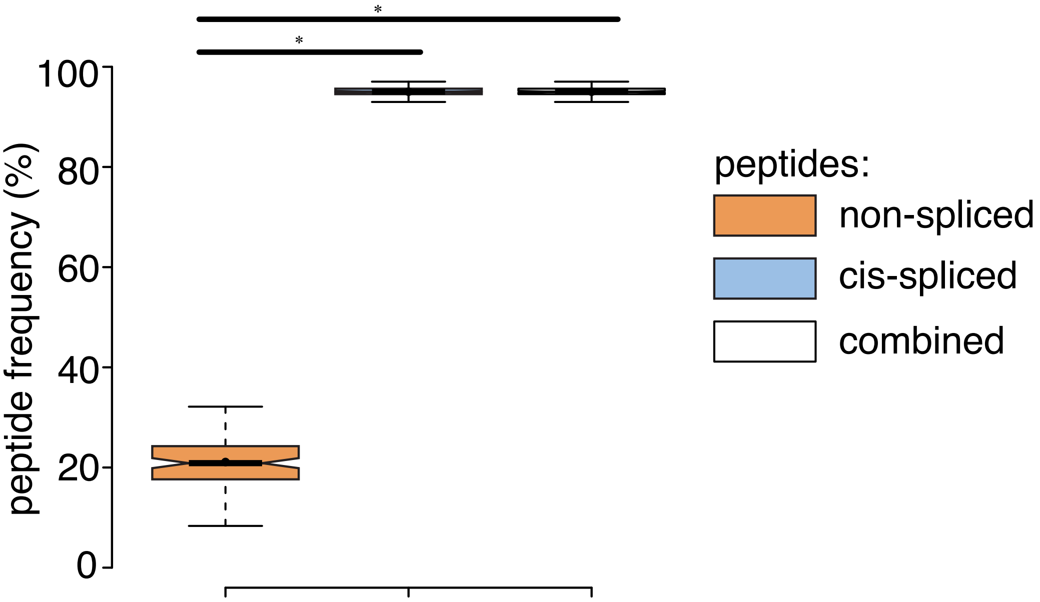


FIGURE 26: Viral-human degenerate zwitter peptides.
Frequency of degenerate viral-human 9mer (non-spliced, cis-spliced and combined) HLA-A*02:01-restricted degenerate zwitter peptides, compared to their cognate viral peptide database and considering the whole human proteome database. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Dots represent the mean. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides. Significant difference between groups are labelled with * (p-value < 0.05).

Similarly, to the analysis of viral-human zwitter peptides assuming exact matches, weighing the number of zwitter peptides based on mTEC transcriptome resulted in a decreased average number of both zwitter non-spliced and cis-spliced peptides, albeit it was relatively modest. On average, 14.83% and 92.6% of the pool of HLA-A*02:01-restricted virus non-spliced and cis-spliced 9mer peptides, respectively, were zwitter peptides using Pintos RNA sequencing database (Figure 27A). Slightly higher decrease was obtained using Zengs RNA sequencing database - 12.35% and 91.07%, respectively (Figure 27B).

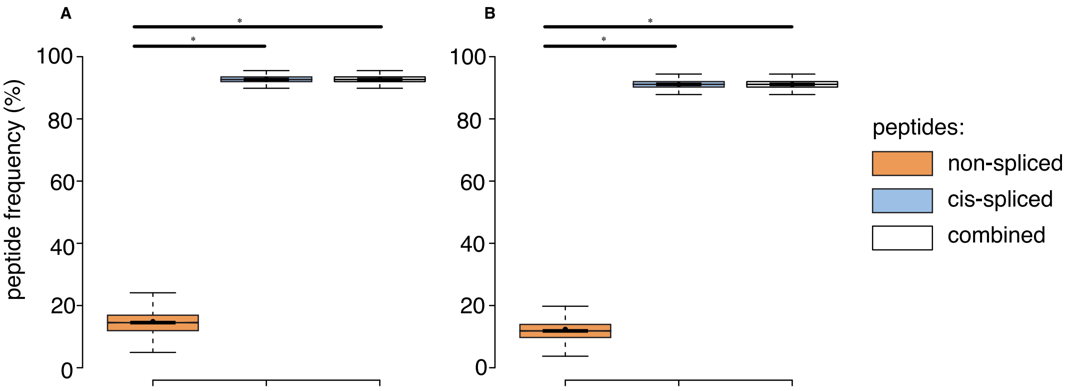


FIGURE 27: Viral-human degenerate zwitter peptides based on mTEC gene expression.

(A, B) Frequency of degenerate zwitter viral-human 9mers (non-spliced, cis-spliced and combined) HLA-A*02:01-restricted zwitter peptides compared to their cognate viral peptide databases considering the human mTEC transcriptome computed either (A) from Pinto *et al.* [101] or (B) from Zheng *et al.* [102]. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Dots represent the mean. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptide. Significant difference between groups are labelled with * (p-value < 0.05).

4.3.2 Estimation of viral-human zwitter peptides weighing cis-PCPS frequency

We reasoned that based on fact that the number of theoretical zwitter peptides increased dramatically when we considered TCR cross-reactivity, we could expect a similar increase in the number of sampled zwitter peptides. Thus, to weigh up the impact of the frequencies of spliced peptides in the immunopeptidome, we utilized the same sampling approach based on the fractions of non-spliced and spliced peptides produced by proteasome and presented on the cell surface as for the estimation of frequencies of zwitter peptides assuming complete sequence identity (Figure 28). We used previously calculated fractions of non-spliced and spliced peptides produced by the proteasome. Using those fractions, we randomly selected degenerate non-spliced and cis-spliced peptides from the human proteome database and the peptides from the viral proteomes, repeated sampling 600 times to reach statistical power and then repeated our entire analysis for each sample.

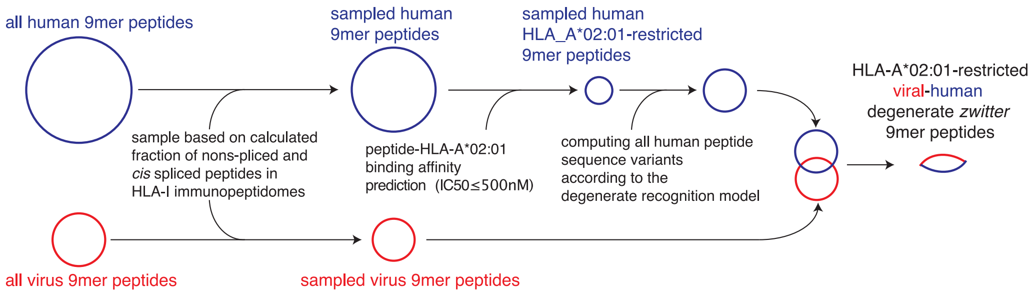

all human 9mer peptides
sampled human 9mer peptides
sampled human HLA_A*02:01-restricted 9mer peptides
HLA-A*02:01-restricted viral-human degenerate *zwitter* 9mer peptides
sample based on calculated fraction of nons-pliced and *cis* spliced peptides in HLA-I immunopeptidomes
peptide-HLA-A*02:01 binding affinity prediction (IC50≤500nM)
computing all human peptide sequence variants according to the degenerate recognition model
all virus 9mer peptides
sampled virus 9mer peptides

FIGURE 28: An overview of the general approach used to estimate viral-human degenerate sampled zwitter peptides.

Schematic representation of *in silico* pipelines to estimate the frequency of degenerate zwitter peptides predicted to bind HLA-A*02:01 complexes accounting for non-spliced and cis-spliced peptide frequency in HLA-I immunopeptidomes.

The number of viruses at each sampling iteration that had non-zero number of degenerate viral-human zwitter peptides, increased dramatically compared to the zwitter HLA-A*02:01 binders for both non-spliced and spliced peptides, particularly at higher frequencies of spliced peptides in the immunopeptidome (Figure 29A). If we assumed a 15% cis-spliced peptide frequency in HLA-I immunopeptidomes, over all randomly sampled peptide pools, we identified, on average, a total of 1312.522 HLA-A*02:01-restricted viral-human zwitter non-spliced 9mer epitope candidates. They correspond to 2.81% of the pool of HLA-A*02:01-restricted virus non-spliced 9mer peptides. This figure strongly varied from virus to virus. On average of sampling, 97 viruses had at least one HLA-A*02:01-restricted viral-human zwitter non-spliced 9mer peptide. In the same analysis, we identified, on average, a total of 308.52 HLA-A*02:01-restricted viral-human zwitter cis-spliced 9mer 1102 epitope candidates. They correspond to 0.00229% of the pool of HLA-A*02:01-restricted virus cis-spliced 9mer peptides, which is a frequency dramatically smaller than the 95.06% computed without accounting for cis-spliced peptide frequency in HLA-I immunopeptidomes. On average of sampling, 68 viruses had an HLA-A*02:01-restricted viral-human zwitter cis-spliced 9mer epitope candidate.

Next, we repeated the non-spliced and cis-spliced peptides sampling and downstream analysis considering a broad range of frequencies of cis-spliced peptides in HLA-I immunopeptidomes. As shown in Figure 29B, the overall picture shifted considerably from lowest to highest frequencies. The average number of HLA-A*02:01-restricted viral-human zwitter non-spliced epitope candidates was estimated to be higher than cis-spliced epitope candidates in the majority of cases, barring the highest frequency of cis-spliced peptides in our analysis. Few outliers of cis-spliced epitope candidates were identified when we assumed very large frequencies of cis-spliced peptide in HLA-I immunopeptidomes. Interestingly, for the non-spliced peptides and for the two highest frequencies of cis-spliced peptides in the majority of sampling iterations at least one virus had one or more zwitter epitope candidate, which is in contrast to the outcome of the same analysis based on the complete sequence identity.

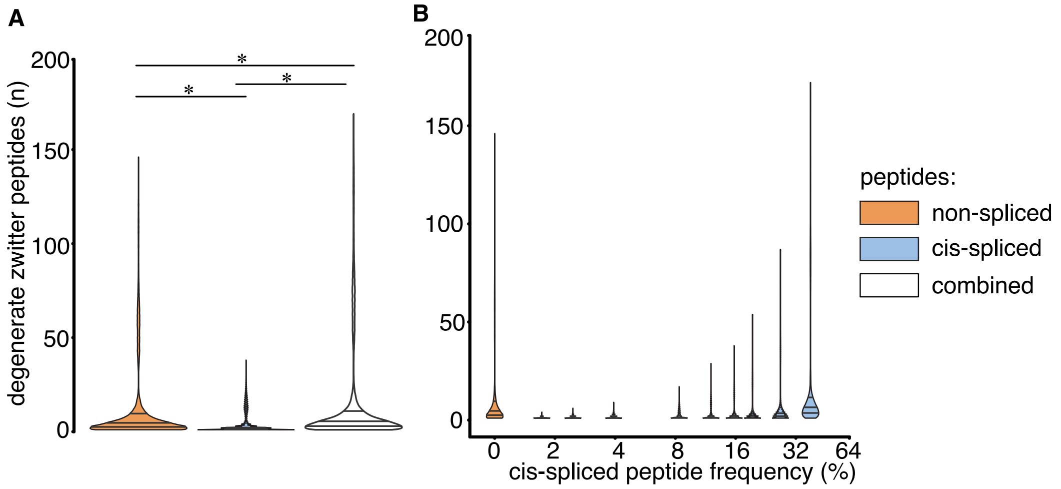


FIGURE 29: Viral-human zwitter peptides considering cis-spliced peptide frequency in HLA-I immunopeptidomes.

Charts are presented as violin plots to show probability densities and to highlight outlier values. (A) Distribution of the number of degenerate viral-human 9mer HLA-A*02:01-restricted (non-spliced, cis-spliced and combined) zwitter peptides per virus across all 600 random samples presented as violin plots (rotated densities). Significant differences between groups are labelled with * (p-value < 0.05). This analysis was carried out by hypothesizing that cis-spliced peptides represent 15% of peptides in HLA-I immunopeptidomes and by using the whole human proteome as database. The distribution of the number of viral-human zwitter cis-spliced peptides has been displayed among viruses that had at least one zwitter peptide. (B) Number of HLA-A*02:01-restricted degenerate viral-human zwitter non-spliced and cis-spliced 9mer peptides per virus per sampling iteration, depending on a broad range of theoretical cis-spliced peptide frequencies in HL A-I immunopeptidomes. Here, viral proteomes are compared to the whole human proteome database. The number of viral-human zwitter non-spliced and cis-spliced 9mer peptides per virus per iteration has been computed among viruses that had at least one zwitter peptide. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides. Significant difference between groups are labelled with * (p-value < 0.05). Horizonal lines represent 1st (25th percentile), 2nd (50th percentile) and 3rd (75th percentile) quartiles.

This phenomenon was reflected also in terms of number of viruses that, on average of sampling, had one or more HLA-A*02:01-restricted viral-human zwitter epitope candidates. In case of the degenerate zwitter peptides, the average number of viruses with one or more HLA-A*02:01-restricted viral-human zwitter epitope candidates was increased by including cis-spliced epitope candidates if we assumed a frequency of cis-spliced peptides in HLA-I immunopeptidomes larger than 8% (Figure 30).

Finally, we considered the impact of the number of amino acid residues in a given viral proteomes and the number of viral-human zwitter peptides. As for the viral-human zwitter peptides assuming complete sequence identity, we observed

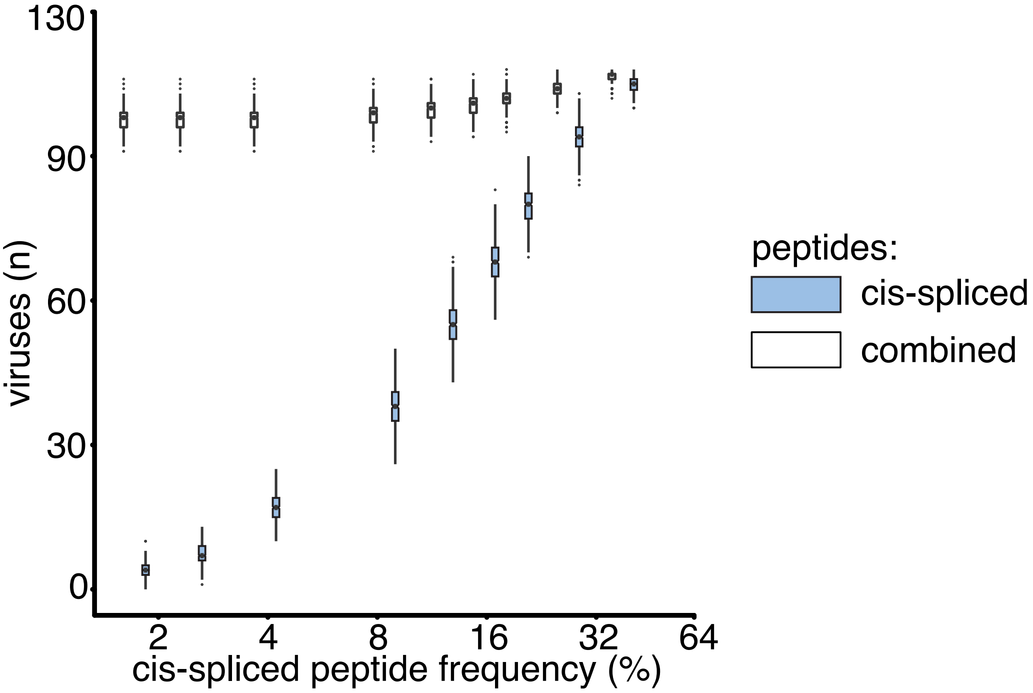


FIGURE 30: Viruses that have degenerate zwittter peptides depending on cis-spliced peptide frequency in HLA-I immunopeptidomes.

Average number of viruses that contain at least one HLA-A*02:01-restricted degenerate viral-human zwitter 9mer peptide per iteration, depending on a broad range of theoretical cis-spliced peptide frequencies in HLA-I immunopeptidomes. Viral proteomes are compared to the whole human proteome database. The boxplots of the combined peptides have been slightly shifted on the x axis for representation purpose. Orange color denotes zwitter non-spliced peptides. Light blue color denotes zwitter cis spliced peptides. White color denotes zwitter combined peptides.

an excelled correlation between viral-human zwitter epitope candidates and the size of virus proteome databases (Figure 31). In contrast to the viral-human zwitter peptides assuming complete sequence identity, this correlation remained similarly strong for the degenerate zwitter peptides when taking frequencies of cis-spliced peptides in HLA-I immunopeptidomes into account (Table S8). The number of viral-human zwitter peptides was still under-represented for Hepatitis delta virus regardless of the degeneracy.

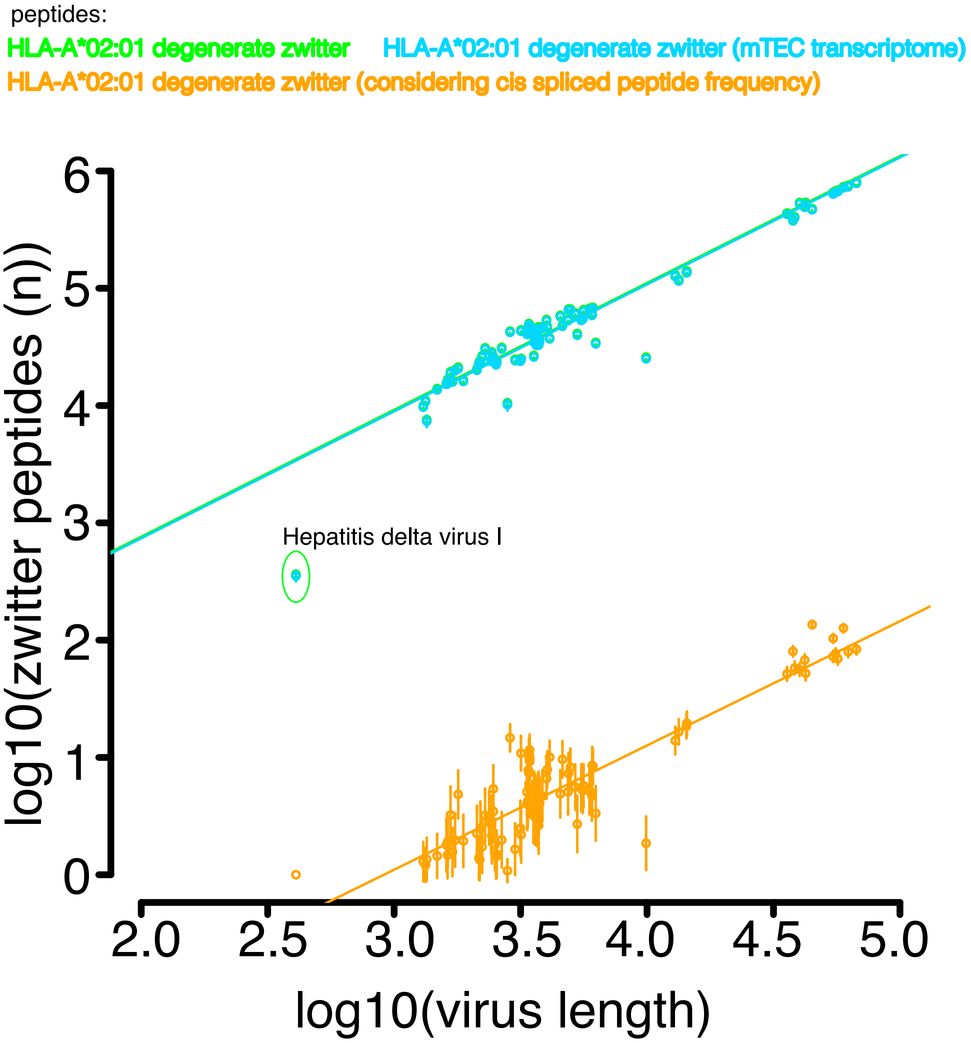


FIGURE 31: Viral-human degenerate zwitter peptide frequency depends on virus length and sequence motifs.

Number of viral-human zwitter combined (*i.e.* non-spliced + cis spliced peptides) 9mer peptides per virus, depending on the number of amino acid residues in its proteome assuming TCR degeneracy. For the groups labeled in yellow, we considered a cis-spliced peptide frequency of 15%, as in Figure 4A. Viral-human HLA-A*02:01-restricted degenerate viral-human zwitter 9mer peptides are represented with a dot each virus. HLA-A*02:01-restricted degenerate viral-human zwitter 9mer peptides either using mTECs RNA-based proteome database [101] or considering the theoretical cis- spliced peptide frequency in HLA-I immunopeptidomes are represented with a dot (mean) and bars (SD) of sampling iterations. Regression lines are shown. The Hepatitis delta virus I has an underrepresented number of HLA-A*02:01-restricted viral-human zwitter 9mer peptides, which are here labeled.

4.4 Discussion

Previously, it was shown that a substantial fraction of non-spliced peptides could be indistinguishable from self from TCRs point of view [24]. We aimed to determine the impact of cis-spliced peptides on the frequency viral-human zwitter peptides considering flexibility of TCR recognition. With the degenerate recognition of pMHCs taken into account, the theoretical frequency of viral-human zwitter peptides increases dramatically to over 95% based on all cis-spliced 9mers that could in principle be generated by the proteasome, and over 95% when considering the rates of antigen presentation by mTECs. These frequency estimates could theoretically explain how viruses escape the immune response - by significantly limiting number of epitopes available to the human immune system. It is important to emphasise that these estimates were obtained without accounting for the expected frequencies of production of cis-spliced peptides by the proteasome and the presentation on the cell surface by MHCI molecules. In fact, only a minor fraction of all theoretically possible peptides would be produced and presented and have an impact on the immune evasion.

To reflect this, when we took the frequencies of cis-spliced peptides that would in fact be produced by the proteasome and presented by MHCI molecules, this originally very high estimate of the frequency of zwitter HLA-A*02:01 restricted cis-spliced peptide dropped from 95.06% to 0.00229%, when assuming the a 15% cis-spliced peptide frequency in HLA-I immunopeptidomes. Thus, similarly to the estimates based on the complete sequence match of self and non-self peptides, we would expect that no or very few zwitter peptides (single digits range) would be presented on the cell surface. This shows that even when taking TCR degeneracy into account, cis-spliced peptides don't increase the size of T-cell holes nor meaningfully impinge upon the variety of circulating T-cells. On the other hand, the cross-reactivity would still increase the probability of zwitter peptides being presented compared to the estimates based on the complete sequence match of self and non-self peptides. Due to this increase, there would be a higher likelihood that such presented zwitter peptide would be recognised by a T-cell, that wasn't removed during the process of negative selection in the thymus. This would in turn increase the probability of an autoimmune response to self in the aftermath of the viral infection.

In our estimates of frequency of zwitter peptides, we didn't consider alterations of the anchor residues of the 9mers. Allowing alterations in the anchor residues would increase the estimate of the frequency of zwitter cis-spliced peptides even further to almost 100% of all of the possible non-self peptides. However, depending on the different anchor residues, an epitope would be expected to have a different binding affinity to HLA-I molecule, which would mean that epitopes with different anchor residues but same middle positions would have different probabilities of being presented and would have different strengths of interactions with TCRs. Subsequently, epitopes with different anchor residues would have different impacts on the immune evasion by the viruses. We reasoned that if for a given peptide the substitutions would occur in one or both anchor residues ,this would be more likely to alter the properties of interaction of the epitope with MHC molecule and possibly prevent its binding to MHCI molecule or decrease the likelihood of presentation.

The level of specificity that we consider here is much larger than what has been shown experimentally. We only considered mismatches in the middle six positions with only up to 2 substitutions and allowed any mismatch in the first position of the 9mer. It is however known that more than two substitutions can still maintain TCR reactivity [16, 123, 124, 125, 126, 127, 128, 133]. Additionally, depending on an epitope, HLAI molecule and TCR, different parts (hot spots) of epitope can be the most important for pMHC-TCR interactions and thus, more mismatches with dissimilar residues could be allowed outside of that area. It was estimated in mice that one TCR would recognize between roughly 1/30000 and 1/100000 of peptides and conversely 1/30000 or 1/100000 of pMHC would be expected to be recognized by a single TCR [133]. According to the degenerate recognition model proposed by Calis *et al.*, one TCR would be expected to recognize 1 in 2.7 million of pMHCs [24]. Similarly, our estimates could be considered rather conservative despite high frequencies of viral-human degenerate zwitter cis-spliced peptides, and could underestimate the frequency of zwitter peptides.

It is also worth pointing out that the true extent of cross-reactivity is still a matter of intensive debate. It can vary dramatically from TCR from TCR. Moreover, peptides with significant sequence overlap were shown to elicit non-cross-reactive T cell responses. In one study, two epitopes derived from neuraminidase protein of Influenza A (SGPDNGAVAV and SGPDNGAVAVL) share complete sequence identity except for the L residue in the longer peptide [140]. Despite this, these two peptides stimulated two different populations of CD8+ T-cells with distinct TCRs. This suggests that residues not in the direct contact with TCR could still have a big impact on the epitope's recognition and that the length of the peptides also plays a role in the properties of the interaction by changing peptide landscape displayed to the TCRs. Therefore, it's not unreasonable to suggest, that with how we implemented the model for TCR degeneracy the number of zwitter peptides that could be recognized by the same TCR could be overestimated as not all TCRs would have the same extent of cross-reactivity. Moreover, the immunological relevance of CD8+ TCR cross-reactivity is still a matter of debate, and even largely overlapping viral epitopes can induce an independent and non-cross-reactive T cell response [133, 135, 137, 140].

Chapter 5

Viral molecular mimicry and Type 1 Diabetes

The results of this chapter were published in [26]

5.1 introduction

It was suggested that the molecular mimicry of the pathogens to human self could lead to an escape from the immune response due to the removal of potentially auto-reactive T-cells during thymic negative selection that could also recognize viral epitopes [24, 82]. An alternative scenario however is possible if there is a failure to neutralize self-reactive T-cells during central or peripheral tolerance. In that case, theoretically even if small number of cross-reactive viral-human zwitter epitopes is present on the cell surface this could prime such auto-reactive T-cells and lead to autoimmune diseases.

One such condition is Type 1 Diabetes (T1D). T1D is a disease during which the body's immune system attacks beta-cells of the pancreas leading to ablation of the production of insulin and cells starvation due to the lack of glucose [141]. In the previous studies, islet-infiltrating CD4+ and CD8+ T-cells were demonstrated to be the hallmark of T1D [51, 141, 142]. Generally, auto-reactive T-cells primarily interact with epitopes derived from a variety of beta islet associated antigens - chiefly insulin and its precursors - pre-proinsulin and pro-insulin, and presented by HLAI and HLAII molecules. The presence and proliferation of self-reactive T-cells is hypothesised to be caused by the incomplete/defective negative selection of T-cells and/or pro-inflammatory islet micro-environment promoting T-cell infiltration into pancreas [51]. The failure of tolerance mechanisms was suggested to be linked to the under-representation of various post-translational modifications (PTMs) in the thymus which change HLA binding and TCR interaction properties as well as usage of alternative mRNA splice variants [141]. Considerable efforts have been attempted to discover novel epitopes that could explain the emergence of T1D. Owing to the technological improvements in the mass spectrometry and bioinformatics, unconventional antigenic peptides (*e.g.* derived from putative non-coding regions, alternative open reading frames (ORFs) and post-translational peptide splicing) emerged as a significant part of the HLA presented peptides. Cis-spliced peptides generated by proteasome, in particular, emerged as a potentially reach source of epitopes that could be targeted by CD8+ T-cells

[6, 58].

However, in context of T1D it's the trans-spliced peptides that are currently under spotlight [143]. A prominent example of such peptides are so-called Hybrid Insulin Peptides (HIPs) which emerged as major auto-antigens in Type 1 Diabetes priming CD4+ T-cell responses [143, 144, 145, 146, 147]. HIPs consist of fragment of insulin as the N-terminal peptide and a fragment of natural cleavage product on as the C-terminal peptide such as chromogranin A (CHGA), secretograninses I-V, Glucose regulatory protein 78 (GRP78) and islet amyloid polypeptide (IAPP) and are produced in secretory granules of beta-islets [143, 144, 145, 146, 147]. HIPs are presented by major histocompatibility complexes class II molecules (MHC-II) in nonobese diabetic (NOD) mice and by HLA-II (HLA-DP, HLA-DM, HLA-DOA, HLA-DOB, HLA-DQ, and HLA-DR) in humans [143, 144, 145, 146, 147]. If this process only occurs in pancreatic beta islets but not in the medulla of the thymus where negative selection occurs, this could explain how auto-reactive T-cells escape tolerance [147]. Delong *et al.* were the first to report the formation of HIPs and their recognition by pathogenic auto-reactive CD4+ T-cells from diabetic mice and PBMCs of T1D patients [143]. They synthesized a synthetic library of hypothetical fusion peptides of insulin and granule proteins detected via mass spectrometry and found CD4+ T-cells recognising one such hybrid peptide. Moreover, those HIP reactive CD4+ T-cells were present in the beta islets. The identity of the peptide in beta islet cells was later confirmed by mass spectrometry. Likewise, Baker *et al.* showed that HIPs are presented in human beta islets and are recognised by Islet-infiltrating T cells [144]. They synthesised various HIPs consisting of covalent linkages of C-peptide fragment on the N-term and b cell secretory granule peptide on the C-term and presented by HLA-DQ2 and HLA-DQ8 molecules and tested for the reactivity of PMBCs to these peptides via INF-gamma assays. There were strong and persistent IFN-gamma responses to those peptides in at least a half of T1D patient derived PBMCs. In addition, they demonstrated that HIP reactive T-cells were present not just in beta-islets but also in the periphery. Layton-Arribas *et al.* constructed a library of theoretical HIP sequences that could be formed by the ligations of insulin fragments or insulin fragment at the N-terminus to peptides derived from secretory granule and islet associated proteins at the C-terminus and presented by DRB1*04:01 MHCI molecules [145]. They then probed those hybrid peptides for the recognition by CD4+ T-cells by stimulating PBMCs of T1D patients with synthesised HIP peptides. They showed that several of such HIPs peptides were indeed recognized by CD4+ T-cells and elicited a strong immunogenic response in the form of high production of IFN-gamma, TNF-alpha and Interleukin-4. Importantly, HIP reactive T-cells were present in significantly higher frequencies in T1D patients compared to healthy subjects. Wang *et al.* resolved crystal structures of TCRs in complex with HIP bound to MHCII and suggested that the transpeptidation occurs in the lysosomes of beta-islet cells or in APCs residing in beta-islets [147]. Wan *et al.* examined MHCII immunopeptidomes of diabetic mice and were able to identify not just conventional peptides but also a HIP formed by fusion of Insulin C-peptide and IAPP. In addition, they found that multiple HIPs were formed in crinosomes suggesting that they indeed are the source of these hybrid peptides for MHCII molecules [142].

Babon J. *et al.* analysed self-antigen specificity of islet-infiltrating T-cells in T1D patients [146]. It was shown that both CD4+ and CD8+ islet-infiltrating T-cells were present in T1D patients in considerably higher frequencies that in healthy subjects. In addition, islet-infiltrating CD4+ T-cells were demonstrated to be autoreactive to a variety of peptides presented by HLA-DR3, HLA-DR4 and HLA-DQ8 and derived from known T1D associated antigens. Notably, IFN-gamma secretion was triggered by various post-translationally modified peptides as well as hybrid trans-spliced insulin peptide fusions [146].

Spliced peptides are a subject of interest for T1D not just in the narrow context of HIPs and CD4+ T-cells. Recently, Gonzales-Duque S. *et al.* applied a systematic discovery approach to identify epitopes presented by HLA-1 molecules of pancreatic islet beta-cells and recognized by islet-reactive CD8+ T cells [141]. They have demonstrated that diverse epitopes associated with pancreatic beta-cells such as secretory granule proteins are generated by a variety of mechanisms, including post-translational peptide splicing and presented by MHC class I molecules. They are selectively recognised by CD8+ T-cells which are enriched in the pancreas in T1D patients. The presence of islet-infiltrating T-cells is the necessary condition for the development of T1D as demonstrated by Culina *et al.* who showed that the human leukocyte antigen (HLA)-A2-restricted zinc transporter 8 (ZnT8) reactive CD8+ T-cells were circulating in both T1D and healthy individuals but the islet-reactive CD8+ T-cells were only found in T1D patients [51].

Previously, we observed that a significant portion of viral cis-spliced peptides was indistinguishable from self peptides based on the entire viral proteomes which however decreased drastically due to the fact that only a small fraction of possible cis-spliced peptides is produced by proteasome and presented on the cell surface. Nevertheless, even a few stray viral-human zwitter peptides could still be sufficient to trigger an autoimmune response provided a self-reactive CD8+ T-cell wasn't removed during negative selection. Thus, we hypothesised that a failure to tolerize self-reactive CD8+ T-cells, followed by viral infection and presentation of cognate viral epitope by MHCI molecules that is similar or identical to self could be one of the mechanisms of triggering of CD8+ T-cell mediated autoimmunity against pancreatic beta-islets. For example, Coxsackie B4 enetrovirus (EV) (CVB4) was isolated from beta islet cells of three out of six T1D patients and that beta cells infected with the virus were dysfunctional. Moreover, virus infected cells exhibited inflammatory phenotype even though auto-reactive T-cell responses weren't detected [148]. There are several examples of sequence similarity of viral proteins to humans antigens implicated in T1D. A non-spliced 9mer epitope derived from P2C non-structural protein of CVB4 with sequence similarity to glutamate decarboxylase 2 (GAD2/GAD65) was identified and predicted to bind with moderate affinity to HLA-A*02:01 molecule. CD8+ T-cells recognising this epitope were also identified and were shown to display cytotoxic activity against target cells presenting this peptide by HLA-A*02:01 molecules. These

T-cells however weren't cross-reactive with cells displaying homologous peptides derived from GAD65 in non-diabetic donors [149]. An immune dominant epitope derived from VP7 protein of human rotavirus and with a high sequence similarity and partial sequence identity to human tyrosine phosphatase IA-2 derived epitopes and presented by HLA-DR4 MHCII molecule was also described. It was shown to elicit CD4+ T-cell responses [150]. Likewise, sequence homology was described between VP7 protein of rotavirues, P2C protein of coxsackieviruses and glycoprotein of adenoviruses and GAD65, although any immunogenic responses were also limited to CD4+ T-cells [151]. Other viruses were also associated with T1D such Human parechovirus (HPeV) (*i.e.*in boys in 6 months before the secretion of T1D auto-antibodies) [152].

Moreover, a link was suggested between persistent viral infections such as Human cytomegaloviruses (HCMV) and T1D [153, 154]. HCMV was found in 22% of T1D patients wheres only 2-6% were found in healthy subjects [153]. In another study HCMV was found to infiltrate pancreatic islets of T1D patient which was also accompanied by the infiltration of CD8+ and CD4+ T-cells into the islets [154]. Moreover, Rodriguez-Calvo and colleagues found CD8+ T-cells in the pancreas of T1D patients reactive to HCMV [155]. Puzzlingly, other groups reported no association between HCMV and T1D progression. No statistically significant differences in frequency of HCMV specific antibodies were found in T1D susceptible and healthy groups of infants [156]. In fact, one study found an inverse correlation between HCMV infection and the progression of T1D [157].

In addition to the aforementioned viruses, Epstein Barr Virus (EBV) and Human Herpesvirus-6 (HHV-6) were also linked to the development of T1D. Bian *et al.* examined antibody responses of T1D patients to an array of viral proteins. The antibody responses of T1D patients to a variety of EBV proteins were significantly more frequent than in healthy subjects [158]. Sabouri *et al.* discovered that glycoprotein B (gB) of HHV-6 was detected more frequently in the beta islets and exocrine pancreas of T1D patients compared to healthy individuals although there was no observable correlation between that and MHCI expression or the infiltration of CD8+ T-cells into the pancreas [159].

Due to the emerging evidence that peptide splicing can play a role in the immune response to pathogens and since enteroviruses and other viral infections have historically been associated with T1D, we aimed to determine the potential effect of mimicry of cis-spliced epitopes derived from viruses associated with type 1 diabetes (T1D) and human pancreatic beta cells in the context of CD8+ T cell autoimmune response on T1D.

5.2 Materials and Methods

5.2.1 Estimation of viral-human zwitter peptides

Viral proteomes were obtained via ViralZone and the human proteome via from Swiss-Prot [98, 99]. The human proteome database contained 20191 protein entries with a total of 11323862 amino acid residues. Per viral species, one strain was selected. Only viruses with human trophism and association to T1D were included in any downstream analysis here presented (n = 8) (Table S7).

Sequences of all possible non-spliced and cis-spliced peptides were computed and were then used to compute the frequency of viral-human zwitter peptides as described (*i.e.* complete sequence match of two given 9mers (see 3.2.1)). In this study we didn't consider T-cell cross-reactivity and only deemed a pair of viral and human 9mer peptides as zwitter if they were exactly matching in all of their 9 positions.

5.2.2 Peptide-HLA-I binding affinity prediction and Immune epitope database (IEDB)

The study focused on non-spliced and cis-spliced 9 amino acid long (9mer) peptides and HLA-A*01:01, -A*02:01, -A*03:01, -A*11:01, -A23:01, -A*24:02, -B*07:02, -B*08:01, -B*15:01, -B*35:01, -B*39:06, -B*40:01, -B*44:02, -B44:03 complexes. This pool of HLA-I alleles covers over 90% of the Caucasian population. For each HLA-I allele, we computed a cut-off comparable to the threshold of a predicted inhibitory constant (IC50) 500 nM of peptide-HLA-A*02:01 complex as follows: we downloaded all 9mer peptides detected through peptide elution from HLA-I complexes, and reported in the IEDB database [160]. We restricted the analysis to the HLA-I alleles specified above. For each peptide-HLA-I complex, we predicted the IC50 of these 9mer peptide sequences by using NetMHCpan-BA4.0 algorithm [161]. IC50 estimates the binding affinity of HLA-I-peptide complexes. The lower the IC50, the higher the binding affinity between peptide and HLA-I complex. To have a similar IC50 cut-off among HLA-I alleles, we determined the quantile of the HLA-A*02:01 for IC50 = 500 nM, which resulted in 91.4%-ile of peptides present in the HLA-A*02:01-specific HLA-I immunopeptidome database of the IEDB (analysis performed by Dr. Juliane Liepe). We then applied this quantile to the predicted IC50 distributions of all other peptide-HLA-I complexes, thereby identifying the IC50 cut-offs of each HLA-I allele, which corresponded to the peptide-HLA-A*02:01 IC50 = 500 nM. Values are provided in Table S8. These cut-offs corresponded to 91.4%-ile of peptides present in the HLA-I immunopeptidome databases of IEDB database.

For the identification of peptides already determined in HLA-I immunopeptidomics or analyzed (with positive outcome) for T-cell recognition, we consulted the IEDB. We downloaded and selected all HLA-I-restricted peptides for which a positive T cell assay was reported [160]. The latter included experiments, for example, performed through tetramer staining, IFN-gamma assays with co-culture of APCs pulsed with synthetic peptide candidates and either peripheral blood mononuclear cells (PBMCs) or CD8+ T cell clones as well as Cr51 cytotoxicity. For the computation of antigenic hotspot regions see below.

5.2.3 Estimation of viral-human zwitter epitope candidates considering antigenic hotspots and the potential antigen repertoire of human mTECs and pancreatic beta cells

To determine the potential hotspot regions among antigens that might be the origin of zwitter epitope candidates, we collected all peptide sequences present in IEDBs human HLA-I immunopeptidome database and mapped them to the reference proteome database (analysis performed by Dr. Juliane Liepe). For each amino acid in the reference proteome database, we counted how many unique peptides of IEDBs human HLA-I immunopeptidome database contained that residue. For any given zwitter 9mer peptide we computed the average count over the 9 residues on its sequence, which was our hotspot score. Finally, we applied a cut-off score of 1 to define hotspot regions. Therefore, a hotspot score of 1 was computed if each residue of a given 9mer peptide was identified at least once in IEDBs human HLA-I immunopeptidome database.

To determine the potential antigen repertoire of human medullary thymic epithelial cells (mTECs) and pancreatic beta cells, we extracted gene expression values from the RNA sequencing dataset of human mTECs and pancreatic beta cells, published by Gonzalez-Duque *et al.* [141], for each antigen in our study. We filtered all antigens based on their expression values, such that the expression was smaller than 0.1 RPKM in mTECs and larger than 5 RPKM in pancreatic beta cells (analysis performed by Dr. Juliane Liepe).

5.2.4 Predicted protein structures

For visualization purpose, the structure of HCMV DNA primase (UL70) and human IA-2 (a.k.a. PTPRN) antigens was determined using iTasser [162] with default settings without inclusion or exclusion of structural templates.

5.3 Results

5.3.1 Estimation of viral-human zwitter epitope candidates potentially associated to T1D

In order to determine which HLAI epitope candidates are viral-human zwitter peptides that could be derived from T1D associated antigens and trigger an autoreacticve CD8+ T-cell response, we first computed the overall 9mer viral-human zwitter peptides shared between human and viruses ostensibly linked to T1D (table S7). The reason we focused on the peptides of length 9 AAs is because this is a preferred length of the peptide binding to HLAI molecules [24]. As in the [82], we only considered non-spliced and cis-spliced peptides with an intervening sequence length of 25 AAs (Figure 32).

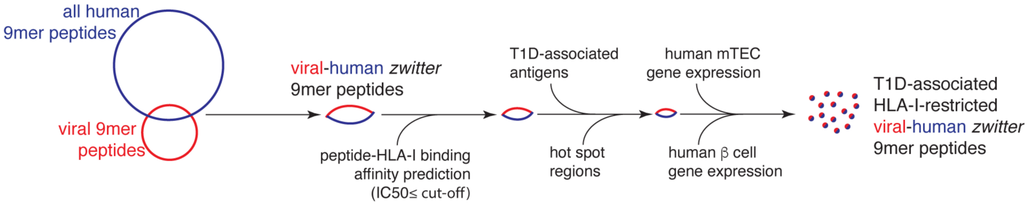


FIGURE 32: *in silico* pipeline for the identification of T1D-associated zwitter epitope candidates

in silico pipeline to identify a pool of zwitter non-spliced or cis-spliced epitope candidates associated to T1D.

In this study, we focused our attention on the HLAI alleles known to be associated with T1D (Table S7) and then sequentially narrowed down our search to arrive to the final set of the relevant 9mers (Figure 32). As we anticipated, there was a drastic discrepancy in terms of the numbers of viral-human zwitter non-spliced and cis-spliced peptides. When we considered non-spliced peptides, only 332 were found to be zwitter (Figrue 33A). Moreover, out of the 8 viruses that we considered, only HHV-6A, HHV-6B, EBV and HCMV had such non-spliced peptides. Out of those 332 peptides, 45 were predicted to bind any of the HLAI molecules (Figure 33B). Out of 45 zwitter epitope candidates, 12 had been previously identified by MS (Figure 33C) but only one peptide had been shown to trigger a positive T-cell assay (Figure 33D). Interestingly, six out of 332 peptides were derived form T1D-associated antigens that had previously been proposed by Gonzalez-Dunque *et al.*, [141] even though none were predicted to be good binders to any of the HLAI molecules we considered (Figure 33E,F). Despite this, two of those peptides had been eluted from the cell surface (Figure 33G). Finally, none of the putative T1D associated non-spliced peptides were demonstrated to elicit CD8+ T-cell response.

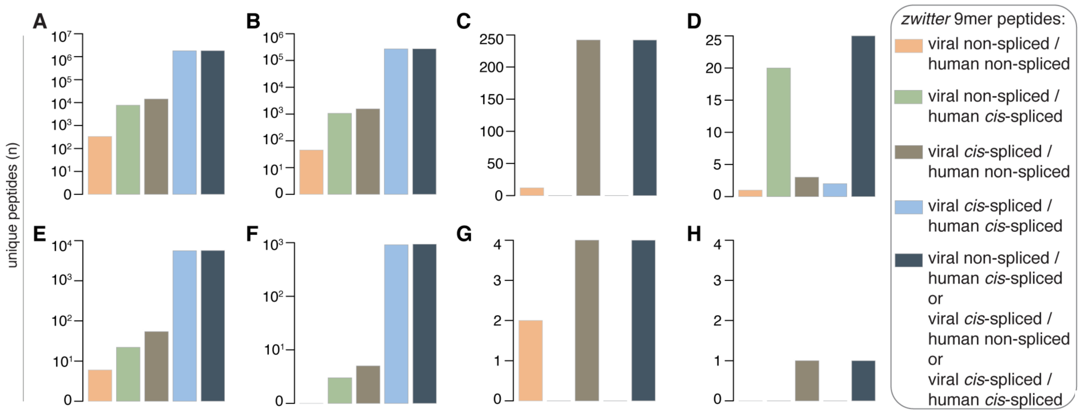


FIGURE 33: Theoretical viral-human zwitter 9mer peptide frequency and potential association with T1D

(A-H) Number of theoretical viral-human 9mer (non-spliced or cis-spliced) (A) zwitter peptides, (B) zwitter epitope candidates predicted to efficiently bind selected HLA-I complexes, (C) zwitter epitope candidates described in published HLA-I immunopeptidomes, (D) zwitter epitope candidates that showed a positive T cell response in published studies, (E) zwitter epitope candidates derived from T1D-associated antigens, (F) zwitter epitope candidates predicted to efficiently bind selected HLA-I complexes and derived from T1D-associated antigens, (G) zwitter epitope candidates described in published HLA-I immunopeptidomes and derived from T1D-associated antigens, (H) zwitter epitope candidates that may be derived from T1D-associated antigens and showed a positive T cell response in published studies. For the identification of epitope candidates already identified in HLA-I immunopeptidomics or analyzed (with positive outcome) for T cell recognition, we consulted the IEDB database.

The theoretical outcome changed when we took cis-spliced peptides into account. In total, almost two million viral-human cis-spliced peptides were identified, while 270000 were also predicted to be HLAI binders (Figure 33A,B). 242 peptides had been identified at the cell surface via MS. It's worth noting that each of those peptides is cis-spliced viral/ non-spliced human and as such they had been considered as traditional human cleavage peptides (Figure 33C). Further, 25 viral-human zwitter cis-spliced peptides had also elicited a CD8+ T-cell response (Figure 33D). Of those 25 peptides, 20 were non-spliced viral / cis-spliced human and were describe as non-spliced peptides in the T-cell assay, accordingly. The other 5 peptides were either cis-spliced viral / cis-spliced human or cis-spliced viral / non-spliced human. For those peptides, T-cell responses were described as either to human non-spliced peptide or viral non-spliced peptide from the other viral strain that we didn't include in our set (Table S7). Coming back to the list of T1D-associated antigens suggested by Gonzalez-Dunques *et al.*, 5000 of the viral-human cis-spliced peptides we computed could be derived from them (Figure 33E). Moreover, close to a 1000 of those peptides were predicted to be HLAI binders (Figure 33F). Four of those peptides had previously been described as non-spliced peptides (Figure 33G), and one viral-human cis-spliced peptide (LLPPLLEHL), found in T1D associated antigens was shown to elicit a weak but detectable T-cell response (Figure 33H). INF-gamma production at the detectable level was detected just in 1 out of 11 T1D patients and no responses were observed from healthy subjects [163]. This peptide can be generated either via regular peptide-bond cleavage from insulinoma-associated antigen 2 (IA-2/ PTPRN) or alternatively from the DND primase (UL70) of HCMV via peptide splicing (Figure 34). Previously, this peptide had been demonstrated to be presented by HLA-A*02:01 with a predicted IC50 of 45 nM and measured IC50 of 444 nM [163, 164]

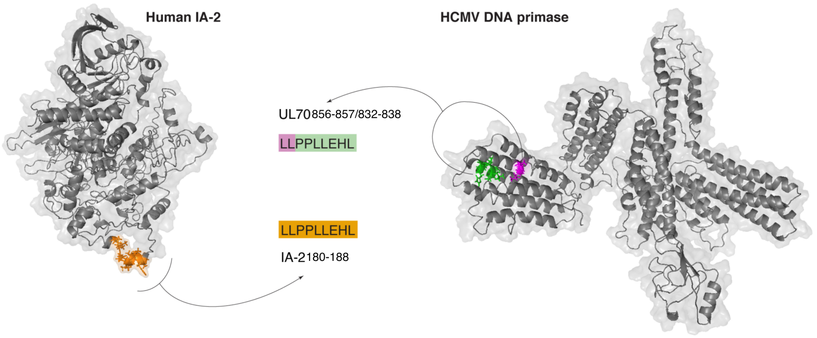


FIGURE 34: Example of potentially immunogenic zwitter viral cis-spliced / human non-spliced peptide.

Predicted crystal structure of the human IA-2 (a.k.a. PTPRN) and HCMV DNA primase (UL70) and theoretical localization of the viral-human zwitter peptide candidate LLPPLLEHL. This peptide may be generated through peptide hydrolysis from human IA-2 and through peptide splicing from HCMV DNA primase. The zwitter non-spliced peptide candidate IA-$2_{180-188}$ [LLPPLLEHL] is depicted in orange. For the zwitter cis-spliced peptide candidate $\text{UL70}_{856-857/832-838}$ [LL][PPLLEHL], the two splice-reactants are depicted, too.

5.3.2 Prioritization of viral-human zwitter peptide candidates potentially associated to T1D

There is a multitude of factors that determine whether or not a given peptide would be able to pass all steps in APP pathway and be presented on a cell surface. Thus, only a small portion of all hypothetically possible peptides would in fact be presented. The amount of of a peptide presented on the cell surface is hypothesised to be determined by the amount of the source antigen, the efficiency of its degradation and crucially on the location of the peptide within so called "hotspot" regions, which seem to give rise to the majority of HLAI presented peptides derived from a given antigen [14, 115]. In addition, despite the fact that gene expression does not mirror the HLA-I immunopeptidomes, it appears, to some extent, to be a predictor of antigen presentation [114, 115]. Thus, it's feasible to assume that if a given antigen has a low expression level in mTECs, peptides from this antigen would be less likely to be presented and less likely to trigger negative selection for CD8+ T-cells that could recognise them. Subsequently those T-cells would go on to circulate in periphery and upon encountering those epitopes could trigger an auto-reactive response. In case of T1D, if a given antigen is highly expressed in pancreatic beta islets, epitopes from this antigen could then trigger immune mediated destruction of beta islets. By applying this logic, we selected viral-human zwitter peptides based on (a) their binding of HLAI molecules; (b) high expression of an antigen in beta islets and low expression in mTECs; (c) association of an antigen with T1D and (d) derivation of a peptide of interest from the hotspot region of the antigen. We collected information on the gene expression and antigen association with T1D from the database generated by Gonzalez-Duque *et al.* [141]. It was predicted whether or not a given peptide was derived from the hotspot region of an antigen based on the available data from HLAI immunopeptidomes while the distribution of gene expression values of the antigens from which viral-human zwitter 9mers could be derived are shown and which expression was detected in mTECs and beta islets is shown in Figure 35.

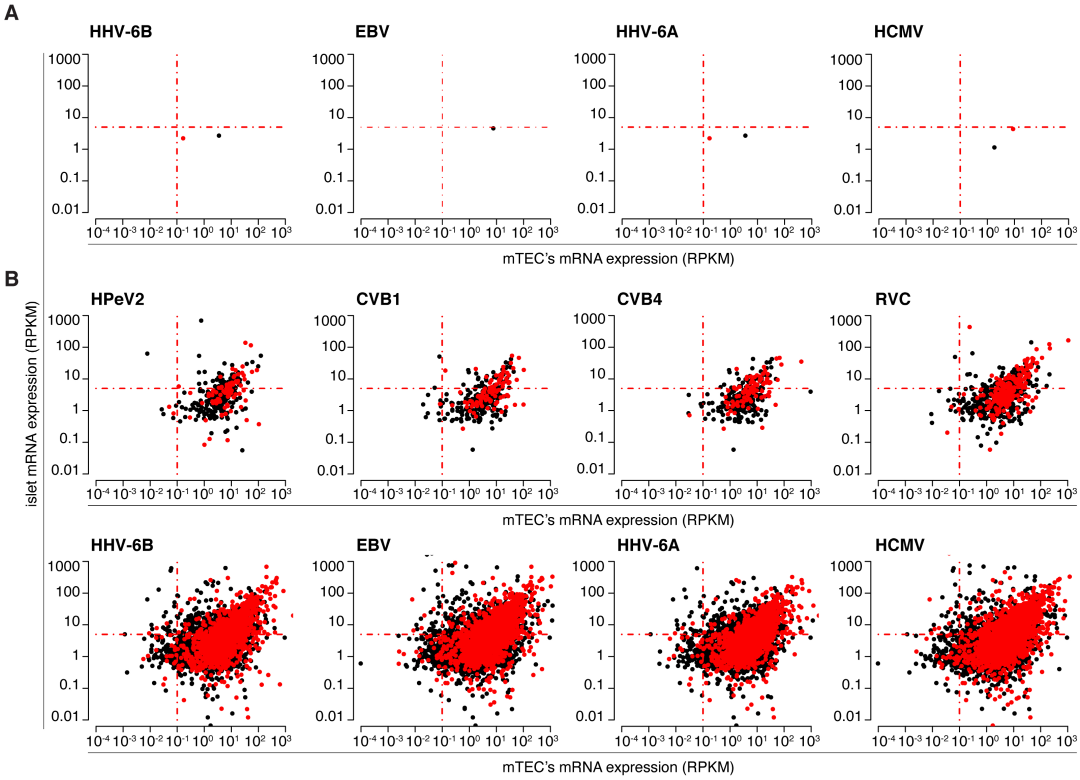


FIGURE 35: Human pancreatic islets and mTECs mRNA expression of antigens potentially carrying HLA-A*02:01-restricted viral-human non-spliced and cis-spliced zwitter peptide candidates.

The scatter plots depict the distribution of RPKM of mRNA of human antigens, as measured by Gonzalez-Duque and colleagues [141] in human pancreatic islets and mTECs, that theoretically can carry viral-human zwitter (A) non-spliced and (B) cis-spliced epitope candidates. Scatter plots are divided based on the corresponding theoretical virus origin. In (A) only four out of eight viruses are shown because for four viruses no viral-human non-spliced peptide candidates with the required characteristics were estimated. Black dots represent antigens carrying epitope candidates predicted to bind the HLA-A*02:01 allele. Red dots represent antigens carrying epitope candidates predicted to bind the HLA-A*02:01 allele and located in hotspots, according to IEDB database.

For the analysis, we selected the gene expression thresholds that had previously been suggested by Gonzalez-Duque *et al.* [141] - 0.1 RPKM in mTECs and 5 in beta islets. With this restrictions we observed no viral-human zwitter non-spliced epitope candidates (Figure 36A). If we disregarded antigens' expression and instead focused on the location of an epitope in the hotspot regions, 16 viral-human non-spliced epitope candidates were identified (Figure 36B). On the other hand, over 900 viral-human cis-spliced epitope candidates were identified if we were to select based on the HLAI binding affinities and preferential expression in beta islets over mTECs (Figure 36A). Based on the derivation from hotspot regions, over 60000 viral-human cis-spliced epitope candidates could be predicted (Figure 36B). Combining hotspot origin and gene expression, over 100 viral-human cis-spliced epitope candidates could be identified, while no zwitter non-spliced peptides remained (Figure 15C). If we were to focus on T1D associated antigens again no non-spliced epitope candidates could be estimated based on HLAI binding and preferential expression in mTECs. On the other hand, over 200 viral-human cis-spliced epitope candidates could be generated (Figure 36D).

One interesting example of such viral-human cis-spliced epitope associated with T1D is a peptide with sequence STATNMFTY which could either be derived from human GAD65 protein as [STA][TNMFTY] or from CVB4 Genome Polyprotein POLG as [STATN][MFTY] (Figure 36E). This peptide is predicted to bind HLA-A*01:01, HLA-A*11:01 and HLA-B*35:01 with IC50 < 100 nM. The other promising peptide has a sequence of IPTQLYHFL (Figure 36F). It can be derived from the Inner Capsid Protein VP2 of RVC as $VP2_{278-284/260-261}$ [IPTQLYH][FL]. In human, this peptide can be derived from Glucose-6-phosphatase 2 as $G6PC2_{312-314/305-310}$ ([IPT][QLYHFL]). A multitude of viral-human zwitter cis-spliced peptides could be derived from HHV-6A and -6B. For example, peptide [IV][LSVALNI] could be produced from ackaging protein UL32 of HHV-6A and HHV-6B. In human, this peptides is derived from from islet amyloid polypeptide (IAPP) as [IVLSVALN][I]. If we continue to narrow down the selection of the epitope candidates and require derivation of a given peptide from T1D associated antigen, we could identify over 100 viral-human cis-spliced epitope candidates and again no non-spliced peptides (Figure 15H). Out of those peptides, 25 could be derived from antigens with high expression level in beta islets and low expression in mTECs (Figure 36I). Such epitope candidates could originate either from EBV or HCMV. One example of T1D associated epitope derived from a hotspot is a peptide LLLAYGGRV, which as binding affinity to HLA-A*02:01 of 93 nM, that could be derived from human KCNK16 as $KCNK16_{17-21/11-14}$ ([LLLAY][GGRV]) or from EBV's Glycoprotein 42 as $BZLF2_{21-25/34-37}$ ([LLLAY][GGRV]). When it comes to IA-2, another major T1D associated antigen, a representative example of an epitope candidate would be GLVNAILKA which is predicted to binding HLA-A*02:01 with high binding affinity. It can be derived from human IA-2 as $IA\text{-}2_{951-953/971-976}$ ([GLV][NAILKA]) and from EBV Major DNA-binding protein DBP as $DBP_{853-855/835-840}$ ([GLV][NAILKA]) (Figure 36M).

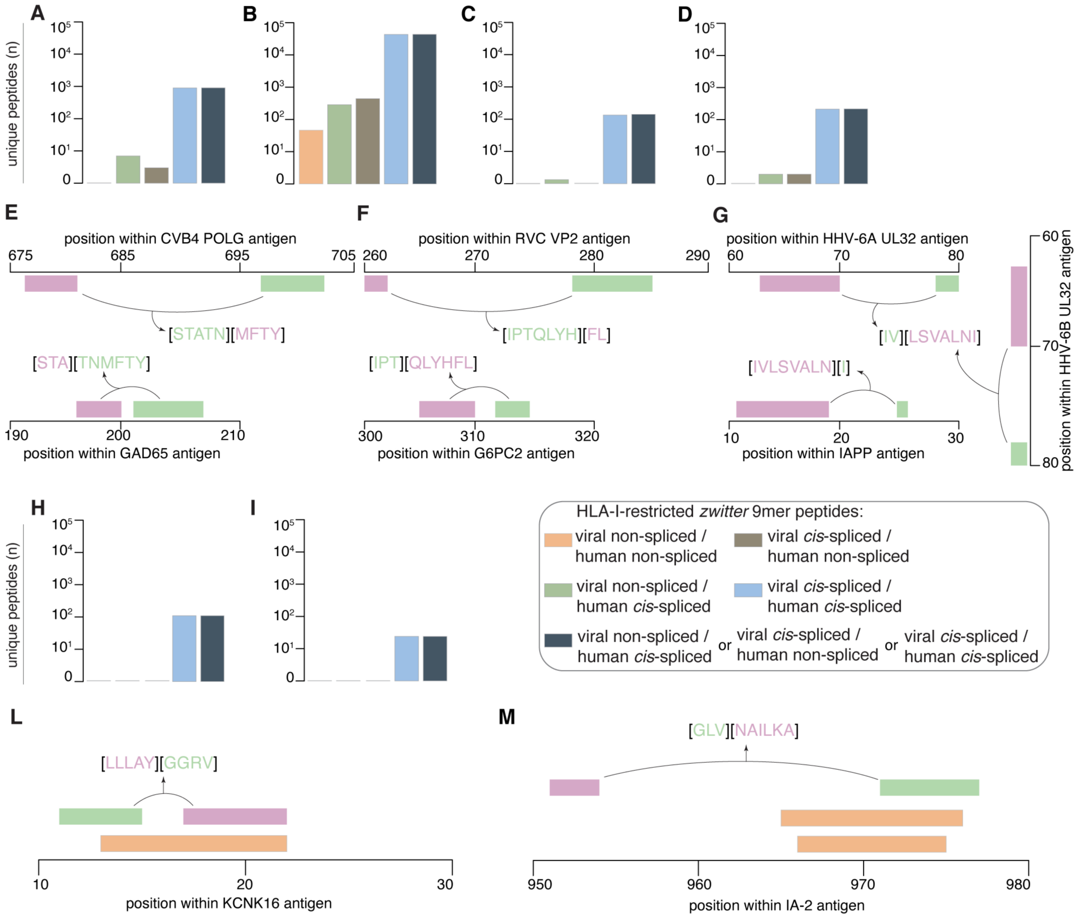


FIGURE 36: Prioritization of viral-human zwitter 9mer peptide candidates and examples.

(A-D) Number of theoretical viral-human zwitter non-spliced or cis-spliced 9mer epitope candidates predicted to efficiently bind selected HLA-I variants and either (A) derived from antigens preferentially expressed in pancreatic islets over mTECs, or (B) located in hotspot regions, or (C) located in hotspot regions of antigens preferentially expressed in pancreatic islets over mTECs, or (D) derived from T1D-associated antigens preferentially expressed in pancreatic islets over mTECs. (E-G) Examples of zwitter viral-human cis-spliced epitope candidates derived from T1D-associated antigens preferentially expressed in pancreatic islets over mTECs: (E) GAD65$_{198-200/202-207}$ [STA][TNMFTY] and CVB4-derived POLG$_{698-702/677-680}$ [STATN][MFTY], (F) G6PC2$_{312-314/305-310}$ [IPT][QLYHFL] and RVC-derived VP2$_{278-284/260-261}$ [IPTQLYH][FL], as well as (G) IAPP$_{11-18/25-25}$ [IVLSVALN][I], HHV-6A-derived UL32$_{79-80/64-70}$ [IV][LSVALNI] and HHV-6B-derived UL32$_{79-80/64-70}$ [IV][LSVALNI] cis-spliced peptides. (H, I) Number of theoretical viral-human zwitter non-spliced or cis-spliced 9mer epitope candidates derived from T1D-associated antigens, predicted to efficiently bind selected HLA-I variants and either (H) located in hotspot regions, or (I) located in hotspot regions of antigens preferentially expressed in pancreatic islets over mTECs. (L,M) Examples of zwitter viral-human cis-spliced peptide candidates potentially associated to T1D and located in hotspots: (L) KCNK16$_{17-21/11-14}$ [LLLAY][GGRV] and (M) IA-2$_{951-953/971-976}$ [GLV][NAILKA] cis-spliced epitope candidates are located in area where non-spliced antigenic peptides (orange bars) have been identified by mass spectrometry in HLA-I immunopeptidomes by others. In (E-G and L-M) bars color code corresponds to that used in Figure 33.

Finally, it's worth pointing out that both of the major antigens IA-2 and KCNK16 in our epitope database are not in fact prevalent in the IEDBs HLA-I immunopeptidome database. It's likely that their domination in the group of optimal viral-human zwitter cis-spliced 9mer epitope candidates is due to a partial sequence homology between their sequence and the viral antigen sequences. It's particulary evident for IA-2 antigen, which among the T1D-associated antigens is one of the largest contributors into the overall pool of viral-human zwitter cis-spliced peptides.

5.4 Discussion

In this study we performed an *in silico* evaluation of zwitter non-spliced and cis-spliced peptides that could trigger an auto-reactive CD8+ T-cell response against pancreatic beta islets and consequently contribute to T1D progression. It is important to treat our results with caution as there are many confounding factors that obscure which possible epitopes would be relevant in the immune response (*i.e.* incomplete understanding of the rules governing peptide-bond cleavage and peptide splicing, difficulties in modelling the entire APP and T-cell degeneracy) [17, 27, 58, 63, 66, 165]. In addition, we only considered the effects of central tolerance on pruning of the CD8+ T-cell repertoire. Due to this we decided to focus on the viral-human zwitter peptides that have complete sequence identity and treated each peptide that satisfied our selection criterion as being capable of being produced by the proteasome and presented on the cell surface. With these restrictions and when ignoring antigen expression in mTECs and beta iselts and binding affinity to HLAIs we estimated that the theoretical pool of viral human zwitter peptides, was in the range of a few hundred for non-spliced and millions for cis-spliced peptides. Upon narrowing down our search to the potentially immunogenic epitopes that could be presented by HLAIs and derived from T1D associated antigens highly expressed in beta islets, these estimates dropped drastically. No non-spliced and only a hundred of cis-spliced peptides passed our selection steps. This figure will likely go down significantly considering that only a portion of theoretically possible non-spliced and cis-spliced peptides is actually produced by the proteasome at the detectable level and not all of these peptides make it through all of the downstream APP steps to be presented on the cell surface [17]. The matter is further complicated by the notion of the T-cell cross-reactivity (degeneracy) wherein one TCR could recognise multiple different epitopes, provided the interaction properties are not significantly altered by the AA substitutions. Predictably, this would lead to an increase of the number of viral-human cis-spliced peptides that could be linked to T1D. On the other hand, It's still neither clear how important T-cell degeneracy for the immune response is, nor what determines the extent of cross-reactivity of a TCR and so the estimates with TCR degeneracy taken into account would have to be approached with caution [131, 133, 137, 140]. There are examples of almost completely identical peptides that elicit non-cross-reactive T-cell responses [140]. Szomolay *et al.* developed a strategy based on combinatorial peptide library (CPL) for determining which peptides are recognized by the individual TCRs the extent of cross-reactivity of CD8+ T-cells and showed broad cross-reactivity of CD8+ T-cell clones recognizing 9mer peptides from HCMV, EBV and HIV-1 to a multitude of similar peptides with the same length [131]. Yet, there were other T-cell clones which were highly specific to particular ligands [131]. Despite the ongoing debate about the relevance of T-cell degeneracy, there is however evidence that TCR degeneracy might play a role in T1D pathogenesis [51, 135, 136]. Culina *et al.* demonstrated that what distinguished healthy and T1D patients were islet-destructing naive CD8+ T-cells self-reactive to ZnT8 antigen enriched in pancreas which interestingly was cross-reactive with mimotope from Bacteroides stercoris [51]. Epitopes derived from fungal and bacterial antigens, showed that they could be recognized by the same CD8+ T-cell clone originally specific to nonamer epitope presented by beta cells and implicated in T1D [135]. In addition, two distinct bacterial epitopes were shown to be recognised by CD8+ T-cells originally reactive to self insulin derived epitope [136]. The other layer of complexity which will have to be addressed in the future is the alteration of the beta islets's microenvironment due to immune activation and the subsequent changes in the antigens expression and the composition of HLA immunopeptidomes.

The viral-human cis-spliced epitope candidates that we identified further suggest a possible contribution of the molecular mimicry into T1D development. The generation of these peptides, their presentation on the cell surface and the ability to trigger CD8+ T-cell responses are still yet to be confirmed however. This will require extensive experimental evaluation. It is prudent to suggest that in some cases viral infections and the presentation of the zwitter peptides in the beta islets of the individuals susceptible to T1D may exacerbate the condition. In this context, the secretion of IFN-gamma and exposure of beta islets to it would likely changed the islet microenvironment and make it more favourable for the antigen presentation and lymphocyte infiltration into the pancreas [166, 167]. Further investigation will be necessary to understand the mechanisms of epitopes generation, presentation and the consequences of the cytotoxic T-cell responses.

Chapter 6

Effect of single point mutations on proteasome catalysed peptide splicing and their consequences for anti-cancer immunotherapies

6.1 Introduction

In light of the discovery that spliced peptides could represent a significant portion of HLA-I immuno-peptidomes their relevance for the adaptive immune response became a distinct possibility [13, 14]. One of the attractive strategies for elicitation of targeted adaptive immune responses against viruses and tumors is the identification of the MHCI presented peptides from relevant antigens (*e.g.* carrying tumor-specific mutations) which could subsequently induce specific CD8+ T-cell responses. For example, Kubuschok *et al.* have focused on the epitopes derived from RAS proto-oncogenes which encode 21 kDa proteins and are GTPases (HRAS, KRAS and NRAS) [44]. These proteins are mutated in over 90% of pancreatic cancer patients. The mutations typically lead to point amino acid substitutions and often occur at codon 12 or RAS proteins. The mutations lead to replacement of wild type (WT) glycine (G) with an aspartic acid (D), valine (V), cysteine (C) or arginine (R) in mutated RAS proteins (MUT). Over half of the patients in the study were carrying G12V substitution in the KRAS protein. In addition more than half of the patients were HLA-A*02:01 positive and thus the authors focused on the elicitation of CD8+ T-cell responses agains KRAS derived epitopes carrying the tumor-specific G12V mutation. First, they identified a 17mer KRAS G12V-derived peptides (KRAS-5-21) that contained binding motifs for both MHCI and MHCII molecules. Subsequently two non-spliced peptides derived from the 17-mer were selected due to their high affinities to HLA-A*02:01 molecules KRAS-5-14 and KRAS-6-14. Both of these peptides elicited a strong CD8+ T-cell responses. Moreover, those T-cells displayed a cytotoxic activity against a variety of cancer cell lines expressing HLA-A*02:01 and loaded with the KRAS-derived peptides. Importantly, those T-cells responses were naturally occurring [44].

A significant limitation of approaches exclusively targeting non-spliced epitopes is that such epitopes carrying for example tumor-specific mutations are either not produced by the proteasome in sufficient quantities to elicit a meaningful immune response or their sequence characteristics are not optimal for the MHCI presentation pathway. One of the potential advantages of spliced peptides is that they could bypass those problems and potentially enlarge the landscape of eptiopes that that could be used in the immunotherapies [13, 14]. For example, Dalet *et al.* performed a screening of antigen-specific CD8+ T-cells capable of responding to melanoma associated antigens and discovered a population of tumor-infiltrating lymphocytes (TILs) recognising HLA-A*24 epitope derived from tyrosinase [22]. Treatment of a melanoma patient with these CD8+ T-cells led to the complete remission. In a follow-up study this same peptide was shown to be produced by reverse cis-splicing of the tyrosinase antigen and naturally processed in the APP [12]. Other studies have documented a generation of cis-spliced peptides that were generated in the proteasomal digestions of melanoma-associated antigens that were presented by MHCI molecules and were capable of eliciting CD8+ T cells responses in the peripheral blood of melanoma patients [2, 23]. We previously utilised a hybrid *in silico* and *in vitro* pipeline to identify a spliced 9-mer epitope derived from KRAS G12V carrying the tumor specific mutation which was produced by the 20S proteasome, was a TAP substrate, could be bound to HLA-A*02:01 with strong affinity and was capable of a strong interaction with the TCR [72].

It has long been known that even relatively minor alterations in antigens sequences could have a profound impact on their proteasomal digestions and the release of the antigenic peptides that could trigger specific CD8+ T-cell responses. Such mutations could have both positive effects such as enhancing generation of MHCI epitopes and their binding affinity and also negative effects via the introduction of strong cleavage sites negating the production of relevant epitopes [28]. Tenzer S. *et al.* analysed the effect of flanking and intraepitope amino acid point mutations on the generation of epitopes from HIV-1 proteins p17 and p24 [168]. They showed that these mutations could have both small and large effects on the cleavages in the HIV-1 p17 and p24 derived polypeptides. These mutations could either completely negate the production of some non-spliced epitopes while others enhanced their abundance and binding to MHCI molecules which had a positive effect on CTL responses [168]. Del Val M. *et al.* showed that not just the epitope sequence itself but also the flanking residues in the antigen strongly impact the generation of such epitope and its presentation [169]. Thus they modified the surrounding sequence of the epitopes of interest derived from Hepatitis B antigens without changing the epitope itself by cloning the epitope's sequence into different positions of the hepatitis B virus core antigen. They found that changing the surroundings of the epitope could have a strong effect on the generation of that epitope and thus its amount. The abundance of the generated epitope was a crucial factor in its presentation by MHCI molecules [169]. Similarly, Eggers M. *et al* cloned a CMV derived peptide into hepatitis B virus derived antigen at several positions and determined that there was a strong effect on the proteasomal processing and the release of the HLA-A*02:01 presented epitopes depending on it's flanking residues [170]. Velders M. *et al* also showed that flanking sequences around the tumor-related epitopes of interest were a deciding factor in the immune response to those epitopes owing to the efficiency of their generation - both upstream and downstream of the epitope [171]. Theobald M. *et al* investigated the impact of a single point R/H mutation located C-terminally relative to the p53 derived non-spliced HLA-A*02:01 presented epitope on its processing by 20S proteasome. The introduction of that single mutation upstream of the epitope sequence inhibited the C-terminal cleavage site necessary to release the epitope preventing the presentation of the epitope by HLA-A*02:01 [172]. Beekman N. *et al.* also investigated an epitope release during 20S proteasomal digestions and its presentation in the presence/absence of a single amino acid substitution flanking the C-terminal proteasome cleavage site of SSWDFITV eptiope from Moloney murine leukemia virus (MuLV) and a homologous epitope from a closely related Friend MuLV. The 26-mer synthetic polypeptide containing the epitope with the original or altered C-terminal flanking residue was digested. The replacement of neutral N with negatively charged D in the flanking C-terminal position resulted in the abolition of C-terminal cleavage after V [173]. Along these lines, Seifert U. investigated the impact of the single-point mutation in the flanking region of HLA-A*02:01 restricted CD8+ T-cell Hepatitis C NS3 derived epitope on it's proteasomal processing and presentation. Y/F substitution in the residue located just terminally of the 9-mer epitope of interest strongly affected the epitope carboxyterminal proteasomal cleavage in the longer precursor peptide containing the epitope sequence. Significantly larger amounts of epitope containing Y in the C-terminal flanking position (WT) was generated in the *in vitro* proteasomal digestion compared to the digestion of the altered peptide [174].

One important area of study with the direct relevance to the epitope selection is the effect of point mutations on the products generation by the proteasome. Little is currently known about the impact of such substitutions on the processing of antigens by the proteasome both quantitatively and qualitatively. The understanding of these effects could have significant implications for epitope discovery. Ossendorp F. *et al.* compared proteasomal processing of two versions of an epitope derived from either AKV/MCF or FMR types murine leukemia virus (MuLV) - KSPWFTTL and RSPWFTTL, respectively by performing the *in vitro* proteasomal digestions of 26-mer polypeptides containing either one or the other epitope sequence [175]. They discovered that K/R substitution introduced a new strong proteasomal cleavage site which resulted in the epitope's destruction. On the other hand, KSPWFTTL was generated by the proteasome, presented by tumor cells and recognised by CD8+ T-cells [175]. Most recently, Fidanza M. *et al.* have investigated the impact of G/Y substitution on the proteasomal processing of LEEKKGNYVVTDH peptide (pepVIII) [29]. This peptide covers the stretch of sequence derived from the junction site formed due to an in-frame deletion in the EGFR gene that results in the removal of exons 2-7 and leads to the creation of a unique glycine at that junction in glioblastoma (GMB) patients. The peptides containing this G could subsequently be used to raise a specific immune response against EGFR in GBM. The authors wanted to know if replacement of G with other amino acid residues could positively impact the pepVIII proteasomal digestion and performed *in vitro* digestions of the peptide variants with the substitutions of pepVIII at position 6, followed by LC-MS/MS, MHCI binding and immonogeneicity assays. Specifically G/Y substitution (Y6pepVIII) had the biggest positive effect on tumor regression and survivial in mice models. This was due to the enhanced proteasomal degradation of the Y6pepVIII compared to pepVIII. The frequency of cleavages after G/Y position as well as amino acids distant to it was increased which resulted in the generation. As a consequence, the digestion resulted in the elevated generation of one non-spliced and nine spliced peptides all of which were closely related to the original sequence. Those peptides exhibited a strong binding to HLA-A*02:01, HLA-B*07:02 and H-2kb MHCI molecules. G/Y substitution didn't affect the peptides recognition with TCRs and the vaccinations of mice with those peptides resulted in significantly improved survival compared to the original pepVIII [29].

In 2015, Textoris-Taube K. *et al* performed *in vitro* digestions of gp100 melanoma derived 30-mer peptide with T/M substitution. This substitution was introduced to improve the immunogenicity of epitopes that could be derived from that 30-mer. They demonstrated that T210M substitution in gp100 melanoma derived peptide had a strong impact on the proteasomal cleavage peptide usage [28]. Moreover, the introduction of mutation resulted in the generation of both non-spliced and spliced peptide products that weren't found in the digestion of WT substrate. The T/M substitution resulted in dramatically altered rates of cleavage after multiple residues in the polypeptide and remarkably the effect of mutation was observed even for positions distant to the substitution. The impact was expressed through the differences in the kinetics of peptide product generation. While the substrate degradation rates weren't impacted the dynamics of the production of both non-spliced and spliced MHCI epitopes of interest was altered. It was hypothesised that the differences in processing could be linked to the binding of the peptides to the substrate binding sites in the proteasome as well as their transport along the proteasome [28]. The main findings of each of the groups highlighted above regarding the impact of the mutations on the proteasomal processing of peptides are listed in Table 2

Table 2: Main findings of the effect of the amino acid substitutions on the generation of eptiopes

Author	reference	year	finding
Del Val M. *et al*.	[169]	1991	introduction of the mutation in the regions surrounding the Hepatitis B derived epitope sequences significantly affect the amount of the generated epitope
Eggers M. *et al*.	[170]	1995	introduction of the mutation in the regions surrounding the CMV derived HLA-A*02:01 restricted epitope sequences significantly affect the epitope generation
Seifert U. *et al*.	[175]	1996	Replacement of the N-terminal K with R in MuIV derived epitope KSPWFTTL resulted in the epitope destruction due to the introduction of the strong cleavage site

Theobald M. *et al.*	[172]	1998	introduction of the R/H substitution upstream of the p53 derived HLA-A*02:01 restricted epitope inhibited it's generation by the 20S proteasome
Beekman N. *et al.*	[173]	2000	introduction of the N/D substitution upstream of the MuLV derived epitope resulted in the abrogation of the C-terminal cleavage after the last residue of the epitope by the 20S proteasome
Velders M. *et al.*	[171]	2001	introduction of the mutation in the upstream and downstream flanking regions surrounding the tumor-derived epitope could abolish the immune response
Seifert U. *et al.*	[174]	2004	introduction of the Y/F substitution upstream of the Hepatitis C NS3 derived HLA-A*02:01 restricted epitope significantly decreased the amount of the epitope produced by the 20S proteasome
Tenzer S. *et al.*	[168]	2009	introduction of flanking and intraepitope mutations either abolish or enhance the production of epitopes derived from HIV-1 proteins
Textoris-Taube K. *et al*	[28]	2015	Introduction of T/M substitution into the sequence of gp100 melanoma derived epitope significantly changed cleavage and splicing preferences and resulted in the increased generation of multiple MHCI peptides
Fidanza M. *et al*	[29]	2021	Introduction of G/Y substitution into the sequence of EGFR derived tumor epitope LEEKKGNYVVTDH enhanced the generation of MHCI restricted non-spliced and spliced epitopes and resulted in tumor clearance in mice

The identification of the spliced epitopes carrying tumor-specific mutations is an attractive strategy for tumor immunotherapies and an elucidation of the impact of amino acid substitutions on the proteasomal dynamics is of great consequence for the determination of how a peptide pool would be altered and if an epitope of interest would be efficiently generated. In light of that and the fact that currently there is a lack of understanding of systematic impact of the single-point substitutions on the proteasomal degradation of substrates, we aimed to answer the following research questions:

- What are the qualitative features of the digestions of WT and MUT substrates and whether the introduction of the mutation affects the expected peptide lengths and the number of products
- How does the mutation affect the generation dynamics of the sequence identical non-spliced and spliced peptides that could be derived from the digestions of both the WT and MUT polypeptides
- Does the proximity to the mutation impact the generation of the peptide products
- Does the mutation alter utilisation of different amino acids for cleavage/splicing
- Are overall cleavage/splicing preferences of the substrates in line with the previously made observations
- Could we use the *in vitro* digestion kinetics to identify and quantify the potential MHCI epitope candidates for use in the immunotherapies

6.2 Materials and Methods

6.2.1 Peptide Synthesis and Proteasome Purification

The peptides that were used in the study were derived from a variety of tumor associated antigens (Table 3) [44, 176, 177, 178, 179]. For each antigen there was an unaltered wild type (WT) version of the peptide and one or more (for KRAS and NRAS) corresponding mutated (MUT) analogues containing one amino acid (AA) substitution. All peptides were synthesized using Fmoc solid phase chemistry. 20S standard proteasome was purified from peripheral blood by Dr. Michele Mishto's group as follows: (i) 10 ml peripheral blood was homogenized, lysed and centrifuged; (ii) the supernatant was fractionated by ammonium sulfate precipitation (35% and then 75%); (iii) the latter pellet was fractioned by chromatography on DEAE-Sephacel; (iv) the selected fractions were separated by 10-40% sucrose gradient and followed by (v) anion exchange chromatography on Mono Q in an Akta-FPLC system; (vi) the selected fractions (2-4 mL) were further purified by DEAE-Affi-gel-blue chromatography. In each of the (ii-vi) steps, the fractions were monitored by degradation assays of standard short fluorogenic substrate Suc-LLVY-AMC. Proteasome concentration was measured by Bradford staining and verified by Coomassie staining of an SDS-Page gel.

Table 3: List of synthetic poly-peptides used for the *in vitro* digestions

Antigen of origin	Substrate sequence	Number of biological replicates
KRAS WT	TEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPT	3
KRAS G12R	TEYKLVVVGARGVGKSALTIQLIQNHFVDEYDPT	3
KRAS G12D	TEYKLVVVGADGVGKSALTIQLIQNHFVDEYDPT	3
KRAS G13D	TEYKLVVVGAGDVGKSALTIQLIQNHFVDEYDPT	3
NRAS WT	VIDGETCLLDILDTAGQEEYSAMRDQYMRTG	3
NRAS Q61R	VIDGETCLLDILDTAGREEYSAMRDQYMRTG	3
NRAS Q61K	VIDGETCLLDILDTAGKEEYSAMRDQYMRTG	3
BRAF WT	LHEDLTVKIGDFGLATVKSRWSGSHQFEQLSG	3
BRAF V600E	LHEDLTVKIGDFGLATEKSRWSGSHQFEQLSG	3
JAK2 WT	KLSHKHLVLNYGVCVCGDENILVQEFVKFGSL	3
JAK2 V617F	KLSHKHLVLNYGVCFCGDENILVQEFVKFGSL	3
MPL WT	LSAVLGLLLLRWQFPAHYRRLRHA	3
MPL W515L	LSAVLGLLLLRLQFPAHYRRLRHA	3
IDH1 WT	RLVSGWVKPIIIGRHAYGDQYRATDFV	3
IDH1 R132H	RLVSGWVKPIIIGHHAYGDQYRATDFV	3

List of synthetic poly-peptides derived from various cancer-associated antigens. WT denotes wild-type version of any given peptide. The numbering of the mutation positions corresponds to the whole protein

6.2.2 *in vitro* Digestions and MS Measurements

Synthetic polypeptides were digested by 20S standard human proteasomes isolated from blood at different time points (0-4h and 20 h) at 37C. For the digestion of BRAF V600E and BRAF WT, 1.5 ug of 20S proteasome from peripheral blood mononuclear cells (PBMCs) in 50 ul was used. For the digestion of JAK2 V617F and JAK2 WT, 1.5 ug of 20S proteasome from PBMCs in 50 ul was used. For the digestion of MPL W515L and MPL WT, 1.2 ug for the first biological replicate and 1.5 ug for the second and third biological replicates of 20S proteasome from PBMCs in 50 ul was used. For the digestion of IDH1 R132H/WT, NRAS Q61R/Q61K/WT and KRAS G12R/G12D/G13D/WT, 1 ug of the 20S proteasome from erythrocytes in 40 ul was used. In all cases, the stock concentration of the synthetic peptides in the digestions was 40 uM. The 0 - 4h (0,1,2,3,4h) digestions were used for the kinetic analysis and 20h digestions were used to identify putative epitope candidates. The concentration of each peptide in the digestion was 40 uM. The buffer used in the digestion has the following composition: 50mM Tris-HCL (pH 7.8), 20 mM KCl, 5 mM MgAc, 1mM DTT. Each digestion was performed 3 times (3 biological replicates). For the kinetics each of the biological replicates was measured either 2 times (2 technical replicates) (for the 0-4 h kinetics) or 1 time (for no proteasome control and 20 h digestions) by mass spectrometry (MS). The *in vitro* digestions were performed by Dr. Michele Mishto's group.

To this end, 20h *in vitro* digestions with 20S proteasomes were measured by Fusion Lumos Mass Spectrometer (Thermo Fisher Scientific) at MPI-BPC. Prior to measurement, the samples were diluted with the loading buffer (2% acetonitrile, 0.05% Trifluoroacetic acid) to a final substrate concentration of 25 uM. Eight ul of those dilutions (corresponding to 200 pmol of substrate initially present in the sample) were injected. Samples were loaded and separated by a nanoflow HPLC (RSLC Ultimate 3000) on an Easy-spray C18 nano column (30 cm length, 75 um internal diameter) coupled on-line to a nano-electrospray ionization Fusion Lumos mass spectrometer (Thermo Fisher Scientific). Peptides were eluted with a linear gradient of 5-55% buffer B (80% ACN, 0.1% formic acid) over 88 min at 50C at a flow rate of 300 nl/min. The instrument was programmed within Xcalibur 3.1.66.10 to acquire MS data in a Data Dependent Acquisition mode using Top 20 precursor ions. We acquired one full-scan MS spectrum at a resolution of 120,000 with an automatic gain control (AGC) target value of 1,000,000 ions and a scan range of 300-1,600 m/z with maximum injection time set to 50 ms and intensity threshold set to 50,000. The MS/MS fragmentation was conducted using HCD collision energy (35%) with an orbitrap resolution of 30,000 at 1.4 m/z isolation window with Fixed First Mass set to 105 m/z. The AGC target value was set up at 100,000 with a maximum injection time of 128 ms. A dynamic exclusion of 30 s and 1-7 included charged states were defined within this method.

in vitro proteasome-mediated digestion kinetics (0-4 h) were measured by LC-MS/MS as follows: Prior to measurement, samples were diluted with the loading buffer and insulin as described above. Eight ul (*i.e.*, 200 pmol substrate) of those dilutions were loaded. For the measurement of the 0-4h kinetics of the digestions of JAK2, MPL and BRAF derived polypeptides (Table 2) human insulin (Sigma-Aldrich) at concentration of 2 uM was added. Insulin was used as a coating polymer to prevent binding of peptides to the glass vials used for measurements and to improve reproducibility between technical replicates. Samples were loaded and separated by a nanoflow HPLC (RSLC Ultimate 3000) on an Easy-spray C18 nano column (30 cm length, 75 um internal diameter; Dr. Maisch) coupled on-line to a nano-electrospray ionization Q Exactive Hybrid-Quadrupol-Orbitrap mass spectrometer (Thermo Fisher Scientific) at MPI-BPC. Peptides were eluted with a linear gradient of 5-55% buffer B (80% ACN, 0.1% formic acid) over 88 min at 50C at a flow rate of 300 nl/min. The instrument was programmed within Xcalibur 3.1.66.10 to acquire MS data in a Data Dependent Acquisition mode using Top 20 precursor ions. We acquired one full-scan MS spectrum at a resolution of 70,000 with an automatic gain control (AGC) target value of 1,000,000 ions and a scan range of 350-1,600 m/z. The MS/MS fragmentation was conducted using HCD collision energy (30%) with an Orbitrap resolution of 35,000 at 2 m/z isolation window with Fixed First Mass set to 110 m/z. The AGC target value was set up at 100,000 with a maximum injection time of 128 ms. For Data Dependent Scans the minimum AGC target value and the Intensity threshold were set to 2,600-20,000 accordingly. A dynamic exclusion of 25 s and 1-7 included charged states were defined within this method.

6.2.3 Spliced and Non-spliced Peptide Identification and Quantification

Peptides were identified using the Mascot version 2.6.1 (Matrix Science) search engine. Mass spectra were searched against a customized database that includes all theoretically possible cis and trans spliced and non-spliced peptides that could be derived from any given substrate. No restrictions on the intervening sequence length between splice reactant (SR) 1 and 2 of cis-spliced peptides were imposed. M oxidation, N-terminal acetylation and NQ deamidation were set as variable post-translational modifications (PTMs). The enzyme was set to NoCleave, meaning that each MS/MS spectra was matched to entries in the database as such without additional *in silico* cutting. For the peptide identification in the Orbitrap Q Exactive measurements, we set as mass tolerances for MS and MS/MS 6 ppm and 20 ppm, respectively. For the peptide identification in the Fusion Lumos measurements, we set as mass tolerances for MS and MS/MS 5 ppm and 0.02 Da, respectively.

Peptide hits were filtered using an ion score cut-off of 20 and a q-value cut off of 0.05. Following DB search and quantification, the data was subjected to post-processing. In order to distinguish between two top splice peptide hits or a top scoring spliced peptide and a lower scoring non-spliced peptides suggested by the database search engine we used the difference between the two top-scoring MS2 spectra suggested by Mascot - delta score. the two top-scoring matches were spliced peptides, the top scan was included into the final data-set provided that it's delta-score relative to the second best scoring match was at least 30%. If the two top-scoring matches were a spliced peptide and a non-spliced peptide, the spliced peptide was selected only if the delta-score was no more than 30%. Otherwise, a non-spliced peptide was picked as the most likely explanation for the given MS2 spectra. When assigning peptides the issue of multimappers had to be considered and Isoleucine/leucine redundancies as a specialised case of mutli-mappers. Multimappers are product peptides such that the sequence could be recapitulated from the substrate in several different ways (for example, as normal-cis spliced, reverse cis-spliced and trans-spliced peptide). In those cases, we had to determine the most likely route for such peptides' generation. When dealing with such ambiguous assignments, we applied a hierarchical order of peptides. That is if a given sequence could be non-spliced, all other alternative splice peptides were discarded. When given a choice between trans-spliced and cis-spliced peptide, we applied a hierarchy in which cis spliced peptide took a priority. Trans-spliced peptides were only included in the subsequent downstream steps if they could be the only possible explanation for a sequence. Finally, if several normal or reverse cis-spliced peptides were possible, such sequences were included in the subsequent downstream analysis as multi-mapper cis. The other issue that we had to consider was Isoleucine (I) and Leucine (L) redundancy. Mass Spectrometers can't distinguish between the two and thus all variants of sequences containing Is and Ls had to be considered in our analysis. We replaced all Is with Ls thus compressing peptides containing Is or Ls but otherwise indistinguishable, into one hit. If non-spliced peptide could be possible with I/L redundancy, we discarded all alternative spliced sequences

Mascot Distiller's label-free quantification (LFQ) add-on (2.8.0) was used to automatically extract MS ion peak areas (extracted ion chromatograms (XICs)) of precursor ions of all identified peptides for all five time points (0-4 h). In LFQ the peak area for each MS1 level scan is calculated based on their signal intensities and used as an estimate of the quantity of each given peptide and charge state. All three biological and technical replicates were processed simultaneously. The resulting peptide kinetics were filtered for peptide synthesis artifacts and non-reproducible peptide kinetics between technical replicates. If MS/MS for a given non-spliced peptide was detected at 0h, such peptide was kept and was removed based on it's kinetic behaviour over time. On the other hand, no spliced peptides were expected to be present in the digestion mixture at 0h and thus if MS/MS of spliced peptide or it's N or C-terminally extended precursors were detected at 0h, such peptides were removed. Furthermore, peptides that showed unrealistic generation kinetic behaviour (such as alternating MS ion peak areas between consecutive time points or monotonously decreasing signal intensity from 0h to 4h) were removed. Moreover, the kinetic was filtered based on the standard deviation of numeric values between biological replicates. If more than one charge or PTM variant of a given sequence was available, the numeric intensity values were summed up. In the final analysis, only peptides that were detected and quantified in at least two biological replicates out of three were considered.

The quantification of the substrate degradation was performed by manually extracting MS ion peak areas based on the m/z ratios of the precursor ions corresponding to the substrate sequences.

6.2.4 Qualitative analysis of the *in vitro* digestions

We were interested in the general qualitative features of all of the digestions that we generated, for both WT and MUT polypeptides. To characterise qualitative properties of the *in vitro* digestions of WT and MUT polypeptides, we extracted the numbers of unique peptides of each type from WT and MUT polypeptides (*i.e.* non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced peptides) and compared distributions of lengths of peptide products and splice reactants (SRs) both between peptides of different types as well as within each category of peptides between wild-type and mutated substrates. This was done to determine if there were clear preferences for particular peptide and splice reactant length ranges of different peptide types. In addition, we were interested in whether there was an impact of the mutations on the lengths of the generated peptide products and splice reactants.

If peptide hydrolysis and splicing were randomly performed by the proteasome without particular specificities, we would expect to observe peptide and splice reactant length distributions identical to those of the random control dataset. This is not what we and other observed in the past. We wanted to verify that the proteasomal cleavage and splicing indeed follows certain rules. We therefore compared the observed length distributions of identified products and splice-reactants to those of the random control dataset. In order to evaluate the sequence characteristics of the identified peptides in the database, we generated a random control dataset. The data-set was generated by combining the generated data-bases for all the substrates used in the study into one file. Next, we removed all theoretically possible sequences, which have a length shorter than 6 amino acids and do not exceed a molecular weight of 7kDa. Following that, we generated a random peptide sample from each product type (*i.e.* non-spliced, normal cis spliced, reverse cis spliced and trans spliced peptides). The sample size of each of the random samples was set to be such that the ratios of numbers of peptide products of different types would be the same as in the actual *in vitro* digestions. If peptide hydrolysis and splicing were randomly performed by the proteasome without particular specificities, we would expect to observe peptide and splice reactant length distributions identical to those of the random control dataset. We therefore compared the observed length distributions of identified products and splice-reactants to those of the random control dataset.

We also wanted to compare the qualitative features of the peptides products to the previously published datasets to compare the methodologies and to reveal unique properties of the digestions of particular groups of polypeptides. In addition, we further characterised our data-set by comparing the numbers of different peptide categories, peptide length distributions and SR length distributions with those generated from the large database of *in vitro* digestions of synthetic polypeptides previously generated by our group [17]. For each polypeptide in our WT-MUT dataset we also counted the number of peptides of each type and computed the frequency of peptides of each type relative to all of the peptides quantified in the kinetics for each WT and MUT polypeptides. Similarly, such frequencies were computed for the polypeptides included in the DB by Specht *et* *al*. This was done to check if the digestions of the polypeptide sequences in our WT-MUT data-set were distinct from the polypeptides that were included in our previously published database.

6.2.5 Quantitative analysis of the *in vitro* digestions

To perform the general quantitative analysis of the data-set, we extracted numeric values for all quantified peptides that passed the kinetic filtering and compared the overall distributions of signal intensities of non-spliced and different types of spliced peptides over time between all poly-peptides. In addition, we computed the mean and total signal intensities of the non-spliced and spliced peptides over time. The primary goal of this analysis was to check if the intensity distributions of non-spliced and different types of spliced peptides followed the previously observed patterns, wherein non-spliced peptides were shown to be considerably more abundant that spliced peptides.

We were interested in determining the potential impact of the mutations in the polypeptides on the quantitative dynamics of the generation of the peptide products by the proteasome. There were several ways we could determine the impact and the first, most straightforward comparison was based on comparison of the signal intensities. In absence of knowledge of the precise concentrations of each of the product peptides, signal intensities could serve as approximation of the quantity of the peptide as it's known that the signal intensity of any given peptide is correlated with it's chemical amount in the *in vitro* digestion setting. To this end, in order to perform the comparison of signal intensities of the product peptides derived from *in vitro* digestions of WT and corresponding MUT peptides, we first determined which of the sequence identical peptides were identified and quantified in the digestions of both WT and MUT peptides. The comparison was focused only on sequence identical peptides because we reasoned that they would have identical ionisation behaviour during MS measurement which would allow us to more accurately judge the impact of the single point mutation on the processing of a given poly-peptide. By contrast, the comparisons of the peptides with the same coordinates but differing in one amino-acid (*i.e.* peptides carrying position with the mutation) based solely on raw signal intensities wouldn't be conclusive due to the unpredictable impact of the substitutions of the peptide behaviour during MS measurement.

To further account for the different rates of substrate degradation in WT and MUT substrates over time and mitigate the impact of the single point substitution on the comparison we normalised the signal intensity values of product peptide to the substrate degradation. The fraction of degraded substrate was calculated as:

$$X_i = (S_0 - S_i/S_0) * 100$$

Where X_i - The relative amount of the substrate degraded at timepoint i; S_0 - Signal intensity of the substrate at timepoint 0; S_i - Signal intensity of the substrate at timepoint i.

Next, the signal intensity of a product peptide P at time-point i was normalised to the amount of substrate degraded as:

$$Pn_i = (P_i/X_i) * 100$$

Where Pn_i - The normalized signal intensity of the peptide P at timepoint i; P_i - Signal intensity of the peptide P at timepoint i; X_i - The relative amount of the substrate degraded at timepoint i.

Thus, the downstream comparisons were conducted based on the normalised signal intensity values.

Due to the vastly different amino acid composition of the different polypeptides and the different properties of the substituted amino acids, in particular, we reasoned that assessing the impact of the mutations on the quantitative dynamics wouldn't produce informative results if we were to compare all of the polypeptides together without separating them based on their unique features. Thus, to assess the effect of the mutation on peptides generation, all of the substrates were split into groups based on the transition of properties between the original amino acid and the substituted amino acid. We reasoned that despite the difference in sequence composition of substrates in each given group we could expect a similar impact of a substitution on the overall processing of the polypeptide by the proteasome. As such, the properties that we considered included amino-acid charge and hydrophobicity (Table 4). For the charge, the substrates were split into those where there was no change in charge (IDH1 R132H, MPL W515L and JAK2 V617G), those in which the neutral AA was substituted with positively charged AA (KRAS G12R, NRAS Q61R and NRAS Q61K) and those in which the neutral AA was substituted with negatively charged AA (KRAS G12D, KRAS G13D and BRAF V600E). For the hydrophobicity, the substrates were split into those where there was no change (IDH1 (R132H), NRAS (Q61R), NRAS (Q61K), JAK2 (V617F), MPL (W515L)) and those where hydrophobic residue was substituted with hydrophilic residue (BRAF (V600E), KRAS (G12D), KRAS (G13D), KRAS (G12R)). For these groups the numeric values for sequence identical peptides were combined and these aggregated distributions were analysed.

Table 4: Substrates group based on the property transition of the mutated amino acid

Substrate and mutation	charge	hydrophobicity
BRAF V600E	neutral to negative	hydrophobic to hydrophilic
IDH1 R132H	no change	no change
JAK2 V617F	no change	no change
KRAS G12D	neutral to negative	hydrophobic to hydrophilic
KRAS G12R	neutral to positive	hydrophobic to hydrophilic
KRAS G13D	neutral to negative	hydrophobic to hydrophilic
MPL W515L	no change	no change
NRAS Q61K	neutral to positive	no change
NRAS Q61R	neutral to positive	no change

First, we counted the numbers of peptides of each type (non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced) detected in the digestions of each given group of polypeptides based on the charge and hydrophobicity and then computed the Pearson correlation coefficients between the summed numbers of product peptides found in the digestions of WT polypeptides and MUT polypeptides. This was done to understand if presence/absence of the mutation could on the qualitative level significantly impact the overall number of the peptide products that would be derived from the digestion of any given WT-MUT polypeptide pairs.

To directly estimate the impact of the mutation on the production rate of sequence identical peptides, we reasoned that we could compute the ratio of signal intensities for each sequence identical peptide between the digestions of MUT and WT polypeptide. If the ratio was larger than one, the mutation would have a positive effect on the generation of a given peptide and vice versa. To do this, for each given sequence identical peptides identified in each given pair of WT and MUT polypeptide, we computed the log10 ratios of signal intensity values of the peptide in the digestion of the mutated polypeptide to the same peptide in analogous wild-type polypeptide at each timepoint. The distributions of the signal intensity ratios were analysed for each peptide type separately (*i.e.* non-spliced, normal cis spliced, reverse cis spliced and trans spliced peptides and all combined peptides) for the timepoints 1-4h. It would be expected that if the mutation had no effect on the generation of peptide the mean of the log10 ratios would be centred around 0. Thus, we determined if the mean of log10 of ratios of signal intensities was significantly different from 0.

Next, we were interested to see if the introduction of mutation had a positional effect on the upregulation/downregulation of the peptide products. That is, we aimed to find out if the upregulation or downregulation of the peptides tended to occur distantly or in close proximity to the mutation. To determine the relationship between the extent of intensity change between MUT and WT polypeptides (*i.e.* log10 ratios, introduced above) we computed the distance of the position with the mutation to the P1 residue for spliced peptides (C-terminal residue of N-terminal spliced reactant) and C-terminal (P1) residue for the non-spliced peptides. If the p1 residue was located N-terminally relative to the position with mutation the distance was expressed as the negative value. Next, the Pearson correlation coefficients were computed between log10 ratios and the absolute distance of p1 residues to the position with the mutation.

Finally, we wanted to know if the presence of the mutation had the strongest effect on the peptides containing specific amino acid residues at P1 positions and located at certain distance ranges to the position with the mutation. This would allow us to determine whether introduction of the residues with particular characteristics resulted in the increased or decreased utilisation of the amino acids from a given group of physical and chemical properties (*e.g.* the mutation resulted in a frequent use of hydrophobic residues). To compare the most highly upregulated/downregulated peptide products in each of the polypeptide digestion groups we computed the mean log10 ratios of MUT to WT peptide products between the 1-4h time-points and selected those that were located below 25th percentile or above 75th percentile in the distribution of log10 ratios signal intensities. After extracting the highly upregulated/downregulated product peptides, for each position in the substrate we counted the number of highly upregulated/downregulated spliced and non-spliced peptides that had that residue used in cleavage or splicing reaction to form the corresponding product peptide. Then we computed the frequency of such upregulated/downregulated peptides relative to the total number of sequence identical peptides sharing the same position. We considered a given position in the substrate to be of significance if at least 50% or more of the peptides having that position as p1/pC were upregulated or downregulated upon introduction of the mutation.

6.2.6 Calculation of site specific cleavage strength (SCS) and frequency of cleavage after P1 residue (PSP-P1) calculation

SCS-P1 (site specific cleavage strength after amino acid residue P1/pC) and PSP-P1 (frequency of peptide splicing catalyzed using the C-terminus of the N-terminal splice-reactant (P1) as splicing site) were calculated based on the absolute signal intensities of each product identified in the proteasome-catalyzed digestions (developed by Dr. Juliane Liepe and Hanna Roetschke). Briefly, for each time point and each amino acid in the substrate, the sum over all mean intensity values between the three biological replicates of product non-spliced amount that have the corresponding substrate amino acid at their C-terminus has been computed and normalized, so that they add up to 100%, resulting in SCS-P1. For each time point and each amino acid in the substrate, the sum over all mean intensities of the three biological replicates of the spliced peptide amount that have the corresponding substrate amino acid at their C-terminus of the N-terminal splice-reactant was computed and normalized, so that they add up to 100%, resulting in PSP-P1.

We then computed the Pearson correlation coefficients between SCS-P1 frequency values of the WT and MUT digestions for each WT-MUT substrate pair and similarly for PSP-P1 frequency values. Next, we extracted all SCS-P1 and PSP-P1 frequency values for each amino acid residue in all polypeptides combined, both WT and MUT compared their SCS-P1 and PSP-P1 frequency distributions. The comparisons were performed for residues belonging to particular broad property group - polar positively charged (R, H, K), polar negatively charged (D, E), polar uncharged (S, T, N, Q), special case residues (C, G, P) and hydrophobic residues (A, V, I, L, M, F, Y, W). In addition the distributions computed for each separate residue and their SCS-P1 and PSP-P1 frequencies were compared.

Additionally, the extracted SCS-P1 and PSP-P1 frequency values for each amino acid residue were grouped for substrates based on their charge and hydrophobicity and their distributions of SCS-P1 values were compared for each broad group of residues between the digestions of WT and MUT polypeptides. The same was done for PSP-P1 frequency values.

6.2.7 HLA-I Peptide Binding Affinity Prediction and Epitope Selection

We predicted binding affinity of all of the theoretically possible non-spliced and cis-spliced peptides carrying the mutations to the following HLA molecules: -A*01:01; - A*02:01; -A*03:01; -A*11:01; -A*23:01; -A*24:02; -A*31:01; -A*68:01; -B*07:02; -B*08:01; -B*14:01; -B*15:01; -B*27:05; -B*35:01; -B*40:01; -B*44:02; -B*44:03; -B*51:01; -C*08:02; -C*15:06 . The binding affinities were predicted using the standalone NetMHCPan algorithm (IEDB). The prediction was performed for 8-14 mer peptides (min and max possible lengths of peptides for which IC50 could be predicted, respectively). The IC50 cut-off was set to 5000 nM. All trans-spliced peptides were removed. If a given sequence of the MUT substrate could be re-capitulated in the corresponding WT substrate, such sequences were removed from the final list of candidates.

The identification of the putative epitope candidates or their N/C-terminally extended precursors with the same splice-site was performed in the 20h and no-proteasome control samples measured on Fusion Lumos Mass Spectrometer (Thermo Fisher Scientific) at MPI-BPC as described. The identification of peptides was performed according to scoring criteria described above.

The epitope candidates for further experimental verification were selected if they were identified in 20h digestions and not found in the no-proteasome control samples as such or as N/C-terminally extended precursors which length doesn't exceed the length of epitope candidates as such by more than 5 AAs (for cis-spliced peptides), and based on whether they could be quantified in the 0-4h digestion kinetics. The kinetics was filtered based on whether the signal intensity values of the candidate quantified binders were within at least the 90% percentile of the distribution of signal intensity values of all quantified peptides for at least 2 out of 3 latter time-points (2-4h). If a given set of peptides could be a binder regardless of any I/L redundancies in their sequences, such peptides were counted as identified and quantified. We considered a given epitope candidate identified and quantified either if the minimal epitope was detected or if at least an N/C-terminal precursor of a given epitope no larger than by 5 amino acids compared to the minimal epitope was found.

It was checked if the quantified candidates or its N/C-terminally extended precursors were found in the human proteome considering all isoforms as non-spliced, as cis-spliced with intervening sequence length not exceeding 25 AAs or with no restriction on the intervening sequence length. When searching for matches, all Isoleucines were replaced with Leucines. In addition, the epitope candidates were matched to human proteome as described without I/L redundancies.

6.2.8 Statistical Analysis

All statistical tests have been done in R and differences in distributions have been tested using the Kolmogorov-Smirnov test (ks.test) with cut-off for significance set to p-value < 0.05. The linear correlations were computed using Pearson coefficient with cut-off for significance set to p-value < 0.05. To determine if the distributions of log10 ratios of MUT to WT intensities of peptide products were significantly different from 0, Wilcoxon rank sum test (wilcox.test) with cut-off for significance set to p-value < 0.05 was used. The statistical test values are reported in Tables S9-S26

6.3 Results

In order to evaluate the impact of point mutation on the proteasomal processing of substrates, our collaborators chemically synthesised a number of polypeptides derived from tumour associated antigens. These polypeptides were then digested by human 20S standard proteasomes derived from blood *in vitro* over the course of 4h. Per polypeptide, the experiment was performed 3 times. After the database search of the generated peak lists and LFQ, we performed additional post-processing to deal with the sequence multi-mappers and I/L redundancies. Finally, after obtaining the numeric intensity values, we performed filtering of the quantified peptides based on their behaviour over time to exclude synthesis errors and and unreliably produced peptides. These steps resulted in a total of 2243 non-spliced, 2465 normal-cis spliced, 1593 reverse cis-spliced and 3426 trans-spliced peptides when considering all WT and MUT polypeptides

6.3.1 Qualitative characteristics of the WT-MUT data-set

Firstly, we aimed to perform a general quantitative and qualitative assessment of our WT-MUT dataset and perform a qualitative comparisons of length distributions of peptides and splice reactants (SRs) (Figure 37 and Table S9 and S10) In terms of the overall length distributions there are significant differences in lengths of different peptide types across all of the peptide types (Figure 37A). Despite this, the median lengths of non-spliced, reverse-cis and normal-cis spliced peptides is comparable and the length of these peptide types rarely exceeds 20 amino acids (AAs). On the other hand, the trans-spliced peptides are significantly longer than all other peptide types. This can be attributed to the fact that trans-spliced peptides are formed by the ligation of SR1 and SR2 derived from two separate substrate molecules unlike non-spliced and cis-spliced peptides which lengths are limited by the lengths of the substrate. The length distributions of peptides derived from WT and MUT substrates per type were significantly different despite similar median lengths except for trans-spliced peptides (Figure 37B). Likewise, when comparing the length distributions for different peptide types for WT or MUT substrates separately, we observed the same pattern as in Fig. 37 A (Figure 37C,D).

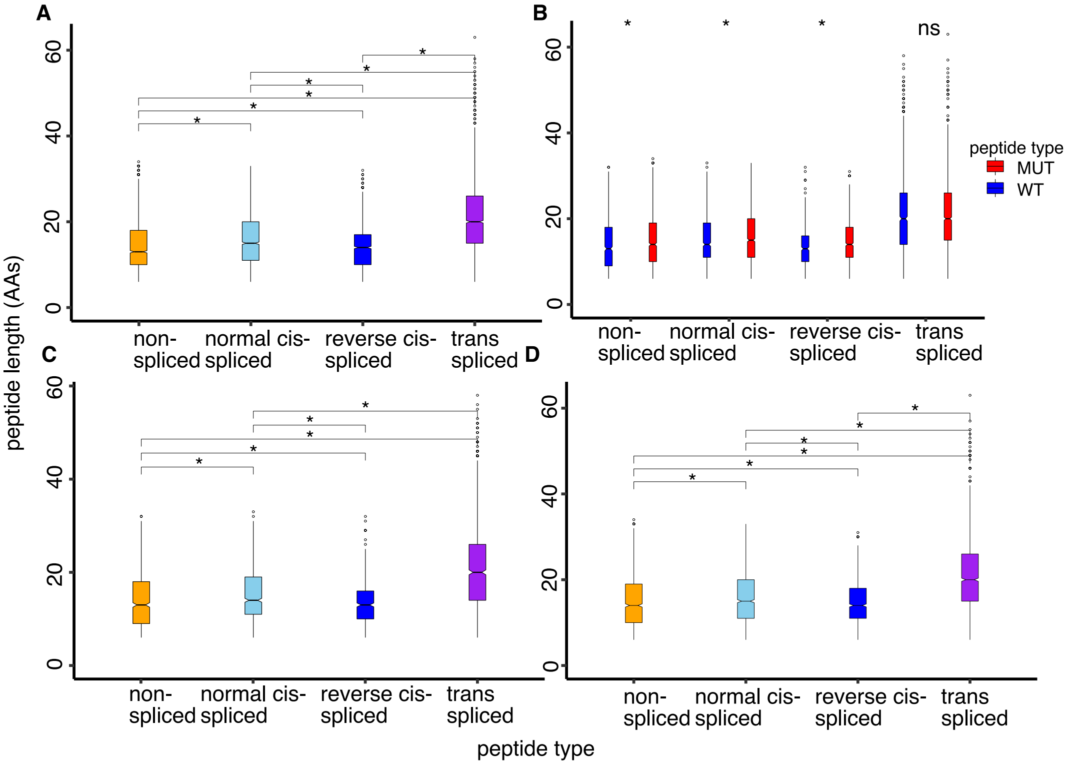


FIGURE 37: Peptide length comparison

(A) Length distributions of all non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides in all of the digested polypeptides (B) Comparison of length distributions of all non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides between digested WT and MUT polypeptides (C) Length distributions of all non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides in all of the digested WT polypeptides (D) Length distributions of all non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides in all of the digested MUT polypeptides. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Black horizontal line represents the median. Orange color denotes non-spliced peptides. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides. Significant difference between groups are labelled with * (ks.test; p-value < 0.05).

We observed a strong correlation between the number of non-spliced and normal cis spliced peptides and the length of substrates from which they were derived. This correlation was not observed from reverse cis- and trans-spliced peptides (Figure 38 and Table S11). This could be attributed to the relatively small number of polypeptides included in our data-set - 15 in total. The correlation between polypeptide length and number of reverse cis- and trans-spliced peptides would likely become stronger and significant upon the increase in the number of analysed substrates.

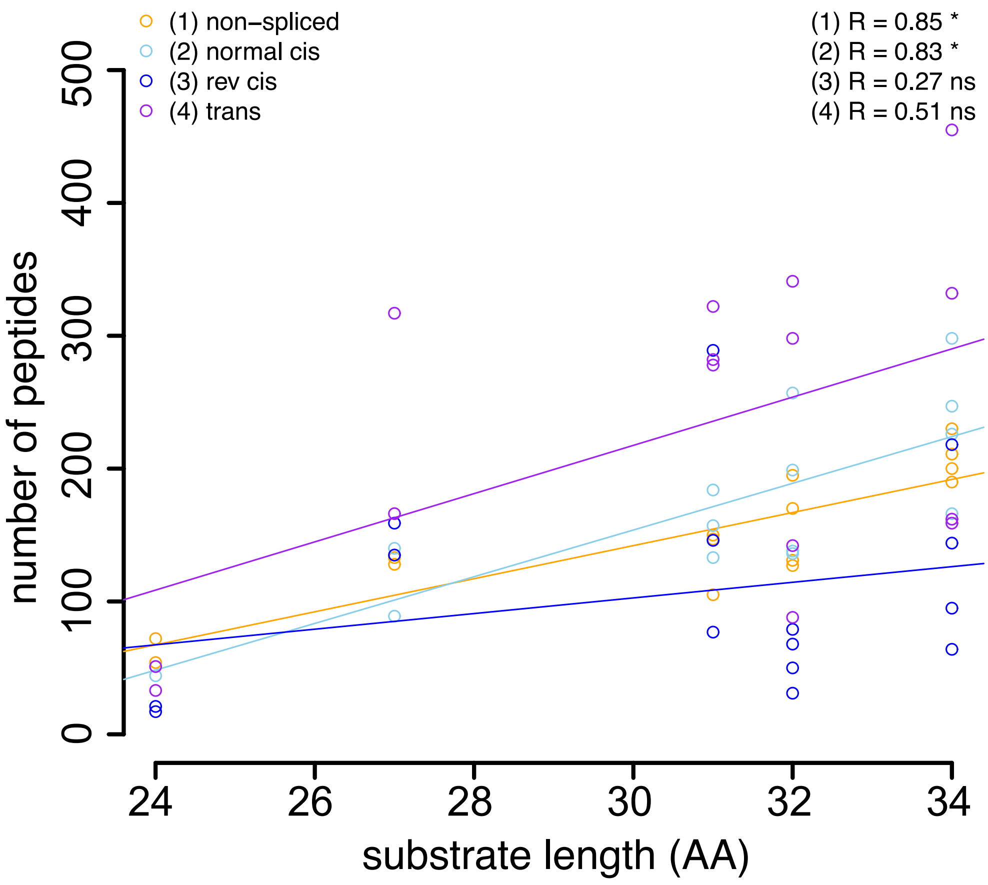


FIGURE 38: Correlation of substrates length and number of peptides

The scatter plot depicts the lengths of polypeptides included into WT-MUT digestion set and the number of products of each type. Orange color denotes non-spliced peptides. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides. Significant Pearson correlation coefficients between lengths and number of product peptides are labelled with * (p-value < 0.05).

Next, we examined the length distributions of SR1 and SR2. Overall, there were significant differences in the length distributions of normal cis- and trans-spliced peptides (Figure 39 and Table S12). Notably, SR2 was significantly longer than SR1 for trans-spliced peptides (Figure 39A). When splitting data-set into WT and MUT derived peptides, the length of SR2 was significantly longer than the length of SR1 for MUT derived trans-spliced peptides, which wasn't the case for WT derived trans spliced peptides, despite the higher median length of SR2 (Figure 39B,C). Similarly to combined peptide products, there were significant differences in length distributions of SR1 and SR2 lengths for norma cis-spliced and reverse cis-spliced peptides, for both WT and MUT derived product peptides. Interestingly, SR1 of normal-cis spliced peptides derived from WT polypeptides was much longer than SR2, compared to comparable median lengths of SR1 and SR2 of MUT derived normal cis-spliced peptides. On the other hand, SR2 was significantly longer than SR1 of reverse cis-spliced peptides of WT derived peptides. Yet, the opposite was true for reverse cis-spliced peptides derived from

MUT polypeptides, suggesting the overall impact of single point mutations on the length preferences of splice reactants for cis-spliced peptides, in particular.

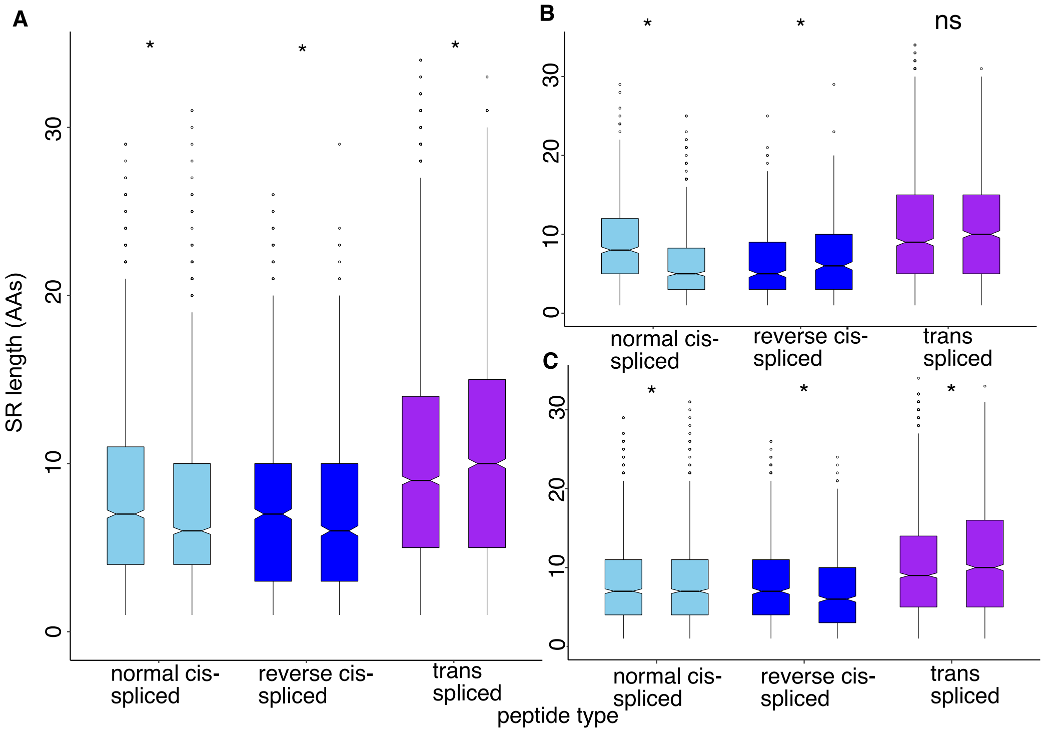


FIGURE 39: SR1 and SR2 length comparison

(A) Length distributions of SR1 and SR2 reactants of all normal cis-spliced, reverse cis-spliced and trans-spliced product peptides in all of the digested polypeptides (B) Length distributions of SR1 and SR2 reactants of all normal cis-spliced, reverse cis-spliced and trans-spliced product peptides in all of the digested WT polypeptides (D) Length distributions of SR1 and SR2 reactants of all normal cis-spliced, reverse cis-spliced and trans-spliced product peptides in all of the digested MUT polypeptides. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Black horizontal line represents the median. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides. Significant difference between groups are labelled with * (ks.test; p-value < 0.05).

Previously, we generated a large dataset of *in vitro* digestions of synthetic polypeptides by the standard- and immuno-proteasomes for the determination of rules of canonical peptide bond cleavage and splicing [17]. The data-set includes a variety of polypeptides digested over the course of 1-20h. We wanted to compare the qualitative characteristics of our WT-MUT data-set with that previously generated data-set - *i.e.* peptide and SRs length distributions. There were noticeable differences in peptide and SR length distributions between peptides in the WT-MUT dataset and peptides included in the published data-base (Figure 40 and Table S13 and S14). On average all non-spliced and spliced peptides were significantly longer than analogous peptides in the published database (Figure 40A). This discrepancy could likely be explained by the fact that we only considered peptides generated in the 4h digestions. By contrast the published data-set includes a multitude of peptides digested for 20 and 24h hours leading to shorter peptides overall. SR1s were significantly longer for normal cis-, reverse cis-spliced and trans-spliced peptides included into WT-MUT dataset compared to the published database for the same reason as for the full peptide length distributions (Figure 40B). This was true for SR2 for reverse cis-spliced and trans-spliced peptides but not for the normal cis-spliced peptides. This could be potentially explained by a stricter restriction on the length of SR2 reactants imposed by the proteasome on normal cis-spliced peptides.

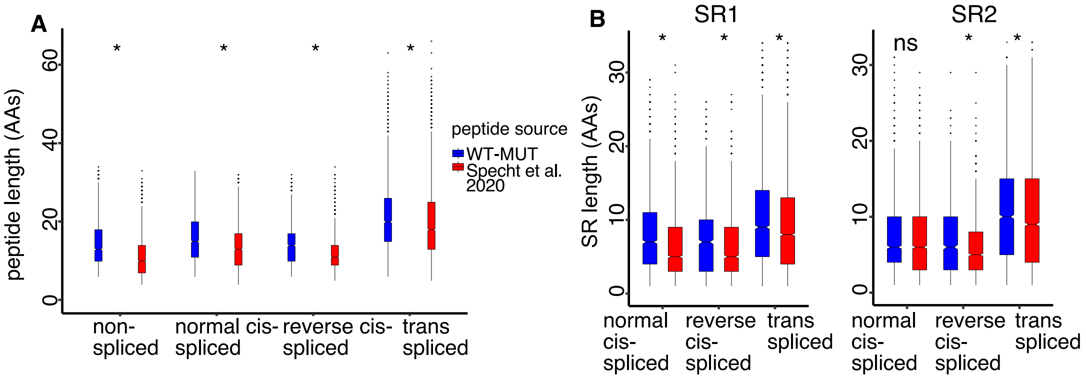


FIGURE 40: Peptide and SR length comparison between WT-MUT and DB2020

(A) Comparison of length distributions of all non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides between digested WT-MUT polypeptides and the DB published in Specht *et al.* 2020. (B) Comparison of length distributions of SR1 and SR2 reactants of all normal cis-spliced, reverse cis-spliced and trans-spliced product peptides between digested WT-MUT polypeptides and the DB published in Specht *et al.* 2020. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Black horizontal line represents the median. Significant difference between groups are labelled with * (ks.test; p-value < 0.05).

In addition, we compared the length distributions of peptide products generated in the digestions of WT and MUT polypeptides with a random control DB. This was done to demonstrate that the proteasome cleavages follow a set of rules and are not performed randomly. If that was the case, we would expect a length distribution that wouldn't be statistically significant from the peptides generated randomly. To assess if that was the case, we compressed all of the DBs for each individual WT and MUT polypeptide into one and randomly selected non-spliced and spliced peptides such that their numbers would observe the same ratios as the numbers of peptides that were included into the final WT-MUT dataset. There were significant differences in both peptide and SR1/SR2 lengths distributions for all of the peptide types between peptides generated in WT-MUT digestions and the randomly selected peptides, except for SR2s of normal cis-spliced peptides, overall confirming that peptide generation by the proteasome wasn't random (Figure 41 and Table S15 and S16).

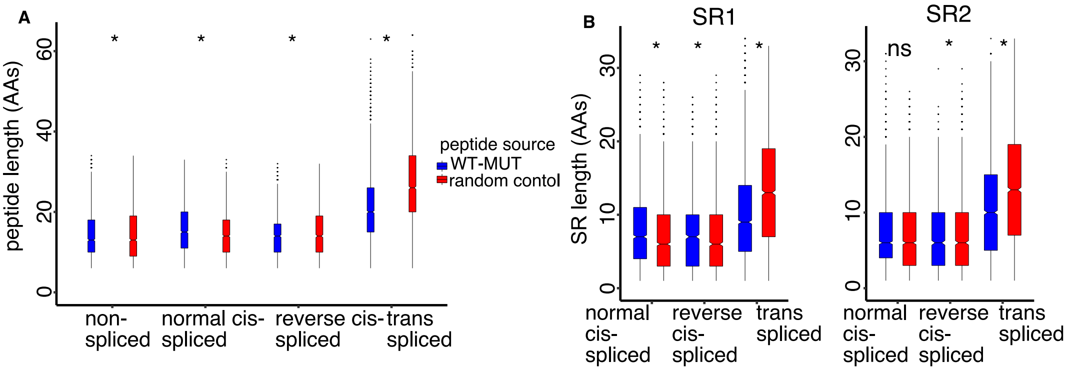


FIGURE 41: Peptide and SR length comparison between WT-MUT and Random DB

(A) Comparison of length distributions of all non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides between digested WT-MUT polypeptides and a random control DB (B) Comparison of length distributions of SR1 and SR2 reactants of all normal cis-spliced, reverse cis-spliced and trans-spliced product peptides between digested WT-MUT polypeptides and a random control DB. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Black horizontal line represents the median. Significant difference between groups are labelled with * (ks.test; p-value < 0.05).

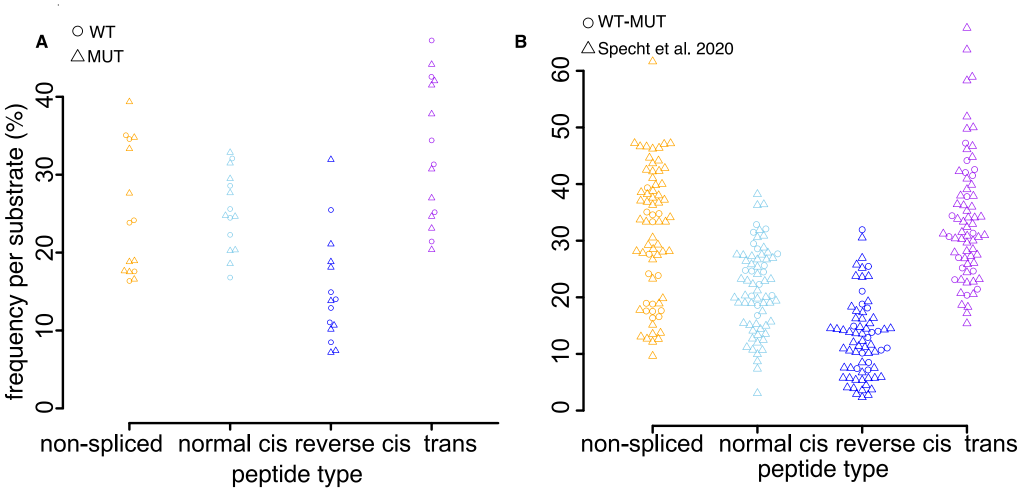


FIGURE 42: Peptide frequency per type

(A) Beeswarm plot depicting distribution of frequencies of non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides relative to the total number peptides detected in the digestions of each individual WT and MUT substrate (B) Beeswarm plot depicting distribution of frequencies of non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced product peptides relative to the total number peptides detected in the digestions of each individual WT-MUT substrate and the substrates included in the database published in Specht *et al.* 2020. Orange color denotes non-spliced peptides. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides.

Finally we compared the frequencies of peptides of different type per substrate for WT and MUT polypeptides and between WT-MUT polypeptides and the substrates included into the published DB. For the WT and MUT polypeptides we didn't observe any clear preference for higher or lower frequency of the peptides of different type depending on the mutation status (Figure 42A). The frequency of product peptides is dependent on both the length and the overall amino acid composition of each given polypeptide. When comparing the frequencies of product peptides between WT-MUT substrates and the substrates included into the published DB we observed very similar distributions of frequencies of different peptide types suggesting that the frequency of the product peptides scales with the digestion time and is again dependent on the sequence composition and length of the substrates (Figure 42B).

6.3.2 Comparison of peptide products generation dynamics

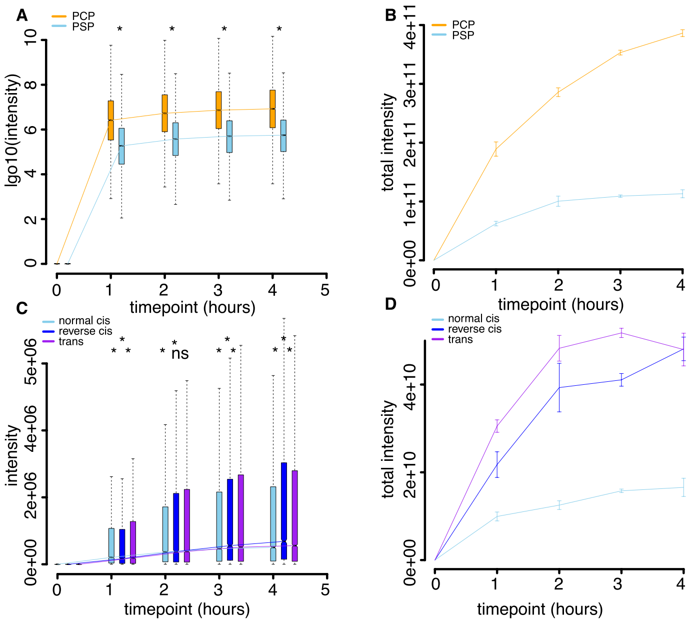


FIGURE 43: Overall intensity comparison

(A) Signal intensity distributions of all non-spliced and spliced peptides in all of the digested polypeptides at each time-point of the digestion. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Black horizontal line represents the median. Orange color denotes non-spliced peptides. Light blue color denotes spliced peptides. Significant difference between groups are labelled with * (ks.test; p-value < 0.05). (B) Comparison of total signal intensities of all non-spliced and spliced product peptides at each time-point of the digestions. Error bars represent standard deviation (SD) between the three biological replicates. Orange color denotes non-spliced peptides. Light blue color denotes spliced peptides. (C) Signal intensity distributions of all normal cis-spliced, reverse cis-spliced and trans-spliced peptides in all of the digested polypeptides at each time-point of the digestion. Box plots depict the median and 25-75 percentiles. Bars represent 5-95 percentiles. Black horizontal line represents the median. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides. Significant difference between groups are labelled with * (ks.test; p-value < 0.05). (D) Comparison of total signal intensities of all normal cis-spliced, reverse cis-spliced and trans-spliced peptides product peptides at each time-point of the digestions. Error bars represent standard deviation (SD) between the three biological replicates. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides.

Before performing the quantitative comparisons of the peptide products generated in the digestions of WT and MUT polypeptides, we wanted to assess the overall quantitative features of the entire dataset (Figure 43 and Table S17 and S18). In terms of overall intensity, the non-spliced peptides are on average much more abundant than spliced peptides at all time-points which is in line with what we observed previously (Figure 43A) [72]. The same held true for the total signal intensity summed between all of the peptides at all timepoints (Figure 43B). On average the trans-spliced peptides are more abundant that normal cis-spliced peptides and reverse cis-spliced peptides are more abundant than trans-spliced peptides (Figure 26 C). The total intensity of trans-spliced peptides is higher than that of the reverse cis-spliced peptides (Figure 43C). Interestingly, normal cis-spliced peptides are the least abundant spliced peptides despite being more numerous than reverse-cis spliced peptides.

In order to perform quantitative comparison the digestions of either WT or analogous MUT polypeptides we had to use the numeric intensity values in absence of knowledge on the precise chemical amounts of each peptide. This meant that for our analysis we only had to compare sequence identical peptides detected and quantified in the digestions of both WT and corresponding MUT substrates. This is because that when it comes position identical but not sequence identical peptides (*i.e.* those that contained the position with the altered residue) we couldn't reliably compare the quantities of such peptides based on signal intensities due to the fact that even a single point mutation could significantly affect peptide ionisation and overall behaviour during MS measurement thus impacting its measured intensity. To further mitigate the impact of mutations on further analysis, when comparing the peptide intensities we elected to normalise them to the fraction of the degraded substrate at each given time-point by dividing the signal intensity of each product peptide by the share of substrate degraded.

First, we looked at the relative amount of each WT and MUT substrate degraded over time. We elected not to directly compare signal intensities of the substrates for the reason discussed above - the unpredictable impact of the mutation on the detected ion current for the polypeptides. Instead, to mitigate the impact of the amino acid alterations, we compared the relative substrate degradation rate. We didn't observe a consistent effect of the mutation on the substrate degradation that would be consistent with the properties of amino acids that were altered (Figure 44). To this end, in some cases the inclusion of mutation seemed to significantly decrease the rate of substrate degradation (BRAF V600E, MPL W515L, KRAS G12D and KRAS G12R) while in others the rate of substrate degradation was comparable (IDH1 R132H, JAK2 V617F, KRAS G13D, NRAS Q61R and NRAS Q61K). The impact of the mutation on the substrate degradation seems to have less to do with the amino acid chemical properties and more with the overall amino acid composition of the polypeptides and the effect of the mutation on the accommodation of the substrate within the proteasome's catalytic chamber.

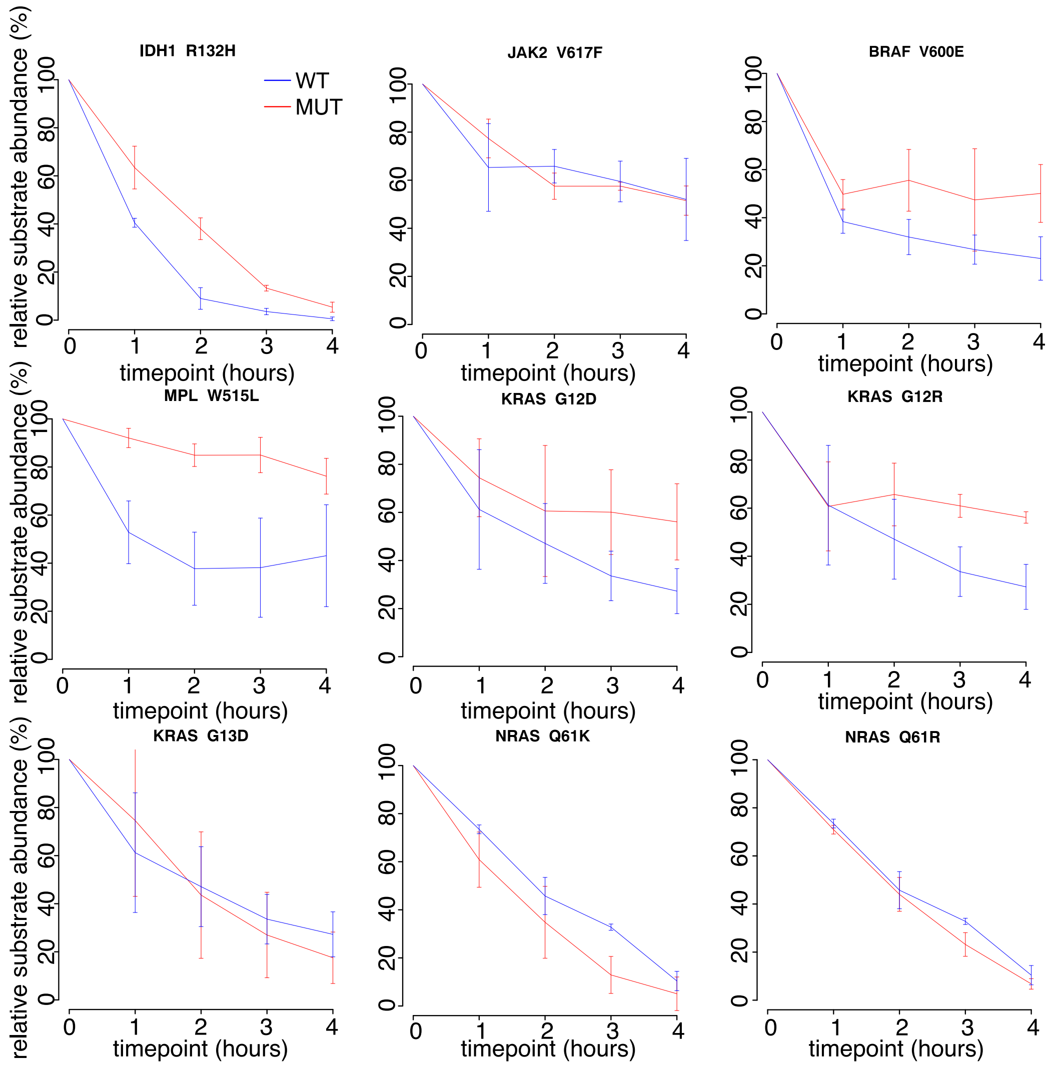


FIGURE 44: relative substrate degradation

Comparison of relative substrate degradation of each given WT and MUT polypeptide pairs. The fraction of peptide degraded was calculated by dividing the mean intensity of the substrate at each given time-point by the mean substrate intensity at timepoint 0h. Red color denotes MUT polypeptides. Blue color denotes WT polypeptides. Error bars represent standard deviation (SD) between the three biological replicates.

In order to discern the impact of different mutation of the proteasomal processing of substrates, we split all of the MUT polypeptides into two large groups based on the general transitions of two key chemical properties between the original amino acid and the altered amino acid - charge and hydrophobicity of the residue. Those groups were further divided into subgroups. Given the polypeptides that were included in our dataset, for the charge group those were mutations where there was no change in charge, those where there was transition from neutral to positively charged residues and those where there was transition from neutral to negatively charged residues. When considering transitions in hydrophobicity the polypeptides were split into those where there was no change in the overall hydrophobicity and those in which hydrophobic residues were replaced with hydrophilic.

Initially, we wanted to evaluate the potential impact of the mutations on the number of generated and quantified products in the proteasomal digestions. To do this we counted the number of identified peptides of each type (non-spliced, normal cis-spliced, reverse cis-spliced and trans-spliced) for each pair of WT-MUT polypeptides and for each of the groups of properties outlined in the previous paragraph and then computed the correlation coefficients between the total numbers of products in the digestions of WT and MUT polypeptides. Regardless of the group of polypeptides that was considered, overall there was a strong and statistically significant correlation between the number of peptide products generated in the digestions of WT and analogous MUT polypeptides (Figure 45 and Table S19). This suggested that the mutations didn't directly impact the number of the produced product peptides. It was yet to be determined, however, if the amino acid alterations impacted the composition of peptide digests and the quantitative dynamics of the products generation. Thus, we aimed to perform a quantitative comparison of the generated sequence identical peptide products found in both the digestions of WT and MUT polypeptides and to discern the potential impact of the mutations on the proteasome's polypeptide sequence preferences via the estimates of frequencies of canonical peptide bond cleavage and splicing after each amino acid position for all of the polypeptides.

Accordingly, we compared signal intensities of sequence identical peptides quantified in digestions of both WT and MUT versions of the same polypeptide. For each sequence identical peptide found in both related WT and MUT polypeptides, we computed a log10 of the ratio of the intensity of the peptide found in the digestions of a MUT polypeptide by the intensity of the corresponding peptide found in WT polypeptide and transformed those ratios on log10 scale. If there was no measurable impact of the mutation on the generation of sequence identical peptides one would expect the mean of the distributions of log10 transformed ratios to not be significantly different than 0 (Figure 46 and 47, Table S20).

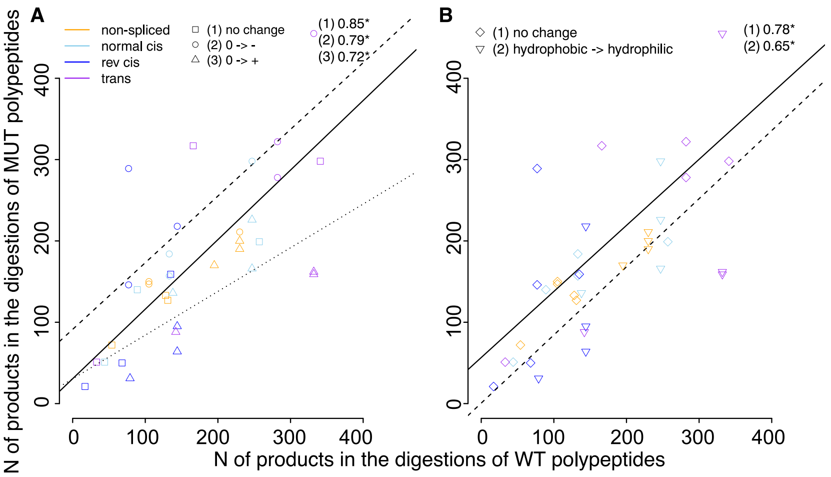


FIGURE 45: Comparison of number of product peptides in different groups of substrates

(A) The scatter plot depicts the number of products included into WT digestion plotted against the number of products included into MUT digestion for each group of polypeptides based on the transition of charge of the altered amino acids. Black solid line correspond to group of polypeptides where there was no change in charge, black dashed line - group of polypeptides where a neutral amino acid was replaced with positively charged, black dotted line - group of polypeptides where a neutral amino acid was replaced with negatively charged (B) The scatter plot depicts the number of products included into WT digestion plotted against the number of products included into MUT digestion for each group of polypeptides based on the transition of hydrophobicity of the altered amino acids. Black solid line correspond to group of polypeptides where there was no change in hydrophobicity, black dashed line - group of polypeptides where a hydrophobic amino acid was replaced with hydrophilic amino acid. Orange color denotes non-spliced peptides. Light blue color denotes normal cis-spliced peptides. Dark blue color denotes reverse cis-spliced peptides. Purple color denotes trans-spliced peptides. Significant Pearson correlation coefficients between the total summed numbers of product peptides in the digestions of WT and MUT polypeptides are labelled with * (p-value < 0.05).

When we considered changes in charge, for the first group of polypeptides - *i.e.* those where the mutation didn't result in change in charge, we didn't observe on average a statistically significant impact of the mutation on the overall generation dynamics of the different types of peptide products over time (Figure 46A). There was however, small but statistically significant negative impact of the mutations on the generation of trans-spliced peptides which nevertheless was only observed during the first 2 hours of the digestions. On the other hand, when a neutral amino acid was replaced with positively charged one, there was a statistically significant increase in the abundances of non-spliced peptides at 1 hour and importantly 4 hour time-point. Remarkably, there was a statistically significant increase in the efficiency of generation of reverse cis-spliced peptides at all time-points but not the other types of spliced peptides (Figure 46B). On the opposite end of the spectrum, when a neutral amino acid was replaced with a negatively charged residue we saw a dramatic decrease in the generation of both non-spliced and spliced peptides - specifically reverse cis-spliced and trans-spliced peptides (Figure 46C).

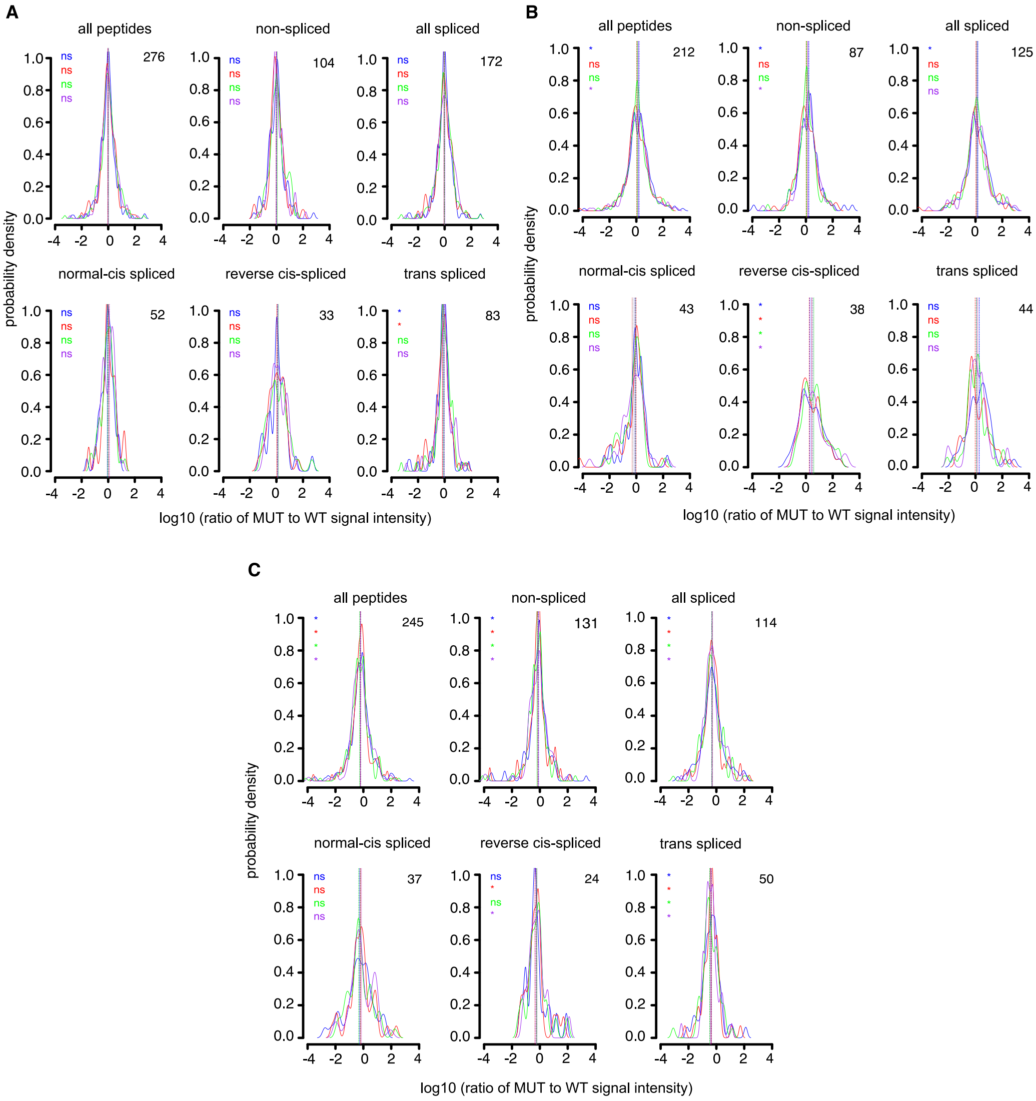


FIGURE 46: Probability density of the signal intensity ratios of the digestions of MUT to WT polypeptides based on charge

Probability density of log10 of mean signal intensity ratios of sequence identical peptides found in the digestions of MUT and WT polypeptides for polypeptides (A) in which there was no change in charge, (B) in which a neutral amino acid was replaced with positively charged, (C) in which a neutral amino acid was replaced with negatively charged. The ratios were computed by dividing the mean signal intensity value of the product peptide from the digestion of MUT polypeptide by the mean signal intensity value of the same product peptide from the digestion of the analogous WT polypeptide at each time-point. The probability densities are provided for each broad group of product peptides - all peptides, all non-spliced peptides, all spliced peptides, normal cis-spliced peptides, reverse cis-spliced and trans-spliced peptides. Blue color corresponds to ratios computed at 1h of the digestions, red - at 2h of the digestions, green - at 3h of the digestions, purple - at 4h of the digestions. The dotted line represents the mean log10 signal intensity ratio, dot dashed line - median. Significant differences of the mean of the distribution from 0 are labelled with * (wilcox.test; p-value < 0.05). The value on the top right corner of each plot is the number of unique peptide products of each type

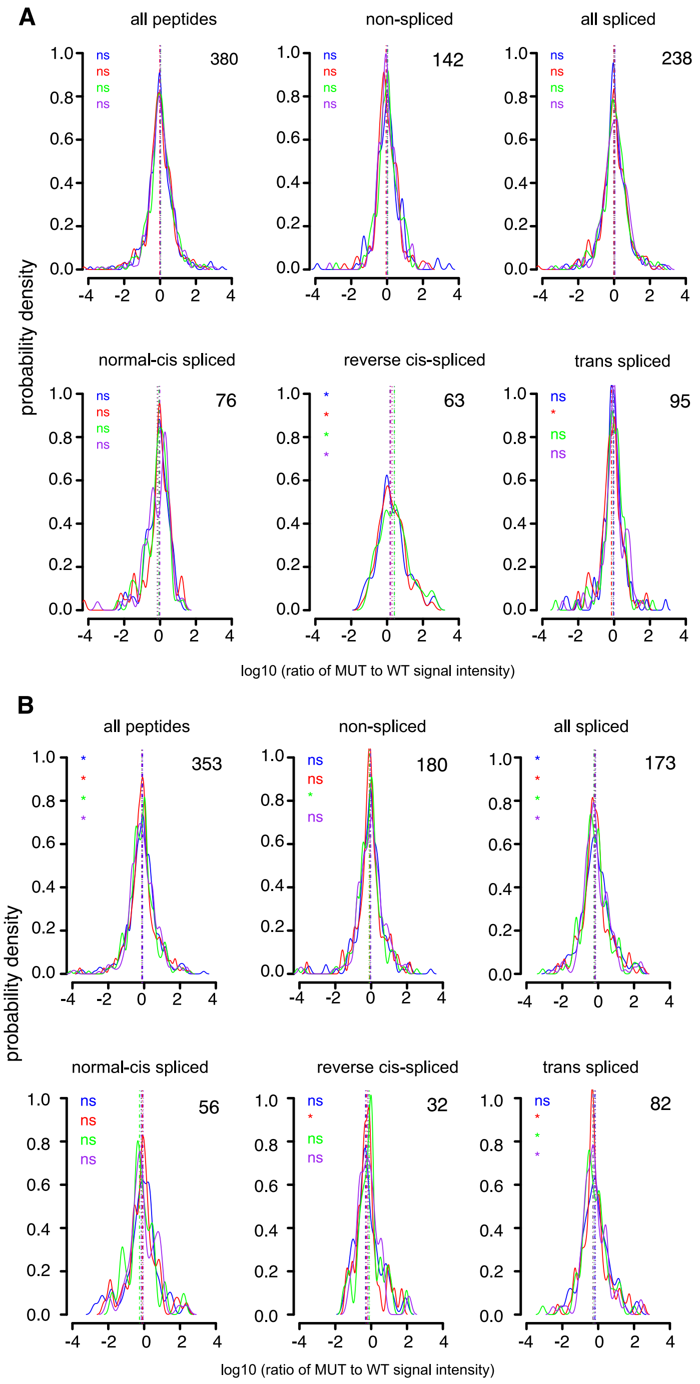


FIGURE 47: Probability density of the signal intensity ratios of the digestions of MUT to WT polypeptides based on hydrophobicity

Probability density of log10 of mean signal intensity ratios of sequence identical peptides found in the digestions of MUT and WT polypeptides for polypeptides (A) in which there was no change in hydrophobicity (B) in which a hydrophobic amino acid was replaced with hydrophilic amino acid. The ratios were computed by dividing the mean signal intensity value of the product peptide from the digestion of MUT polypeptide by the mean signal intensity value of the same product peptide from the digestion of the analogous WT polypeptide at each time-point. The probability densities are provided for each broad group of product peptides - all peptides, all non-spliced peptides, all spliced peptides, normal cis-spliced peptides, reverse cis-spliced and trans-spliced peptides. Blue color corresponds to ratios computed at 1h of the digestions, red - at 2h of the digestions, green - at 3h of the digestions, purple - at 4h of the digestions. The dotted line represents the mean log10 signal intensity ratio, dot dashed line - median. Significant differences of the mean of the distribution from 0 are labelled with * (wilcox.test; p-value < 0.05). The value on the top right corner of each plot is the number of unique peptide products of each type

When we considered polypeptides with no change in hydrophobicity, the significant increase in the peptide generation in the MUT polypeptides were only observed for reverse cis-spliced peptides but not the other peptide types (Figure 47A). On the contrary, when hydrophobic amino acids were replaced with hydrophilic, there was a significant decrease in the abundance of trans-spliced peptides (Figure 47B). Given the smaller degree of change in peptides generation in the digestions based on hydrophobicity compared to those based on charge it appears that at least when it comes to signal intensities of sequence identical peptides, the charge of the permuted amino acid has a more pronounced effect compared to hydrophobicity.

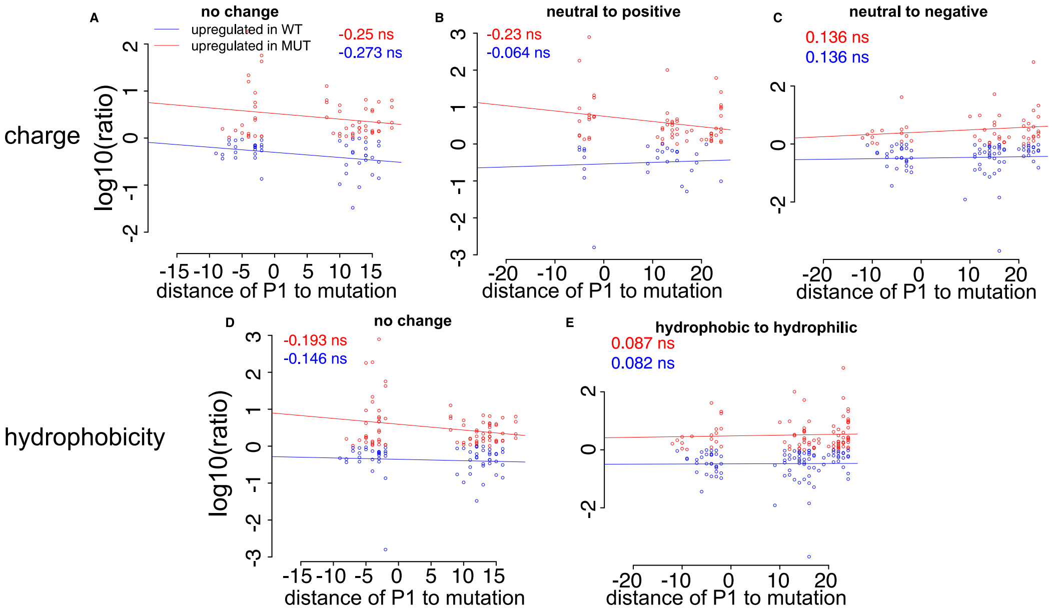


FIGURE 48: Distance of pC residue to the position with the mutation vs log10 signal intensity ratios of non-spliced peptides

The scatter plot depicts the distance of pC residue of non-spliced sequence identical products found in both the digestions of WT and MUT polypeptides plotted against the log10 ratio of the intensity of those product peptides for polypeptides (A) in which there was no change in charge, (B) in which a neutral amino acid was replaced with positively charged, (C) in which a neutral amino acid was replaced with negatively charged, (D) in which there was no change in hydrophobicity (E) in which a hydrophobic amino acid was replaced with hydrophilic amino acid. The ratios were computed by dividing the mean signal intensity value of the product peptide from the digestion of MUT polypeptide by the mean signal intensity value of the same product peptide from the digestion of the analogous WT polypeptide. The mean value of those ratios was then computed between all 4 time-points. The product peptides up-regulated upon alteration of the amino acid are marked in red. The product peptides up-regulated in the digestions of WT polypeptides compared to MUT polypeptides are marked in blue. Significant Pearson correlation coefficients between the distances of pC residue to the position of the mutation and the log10 signal intensity ratios of peptide products found in the digestions of MUT and WT polypeptides are labelled with * (p-value < 0.05).

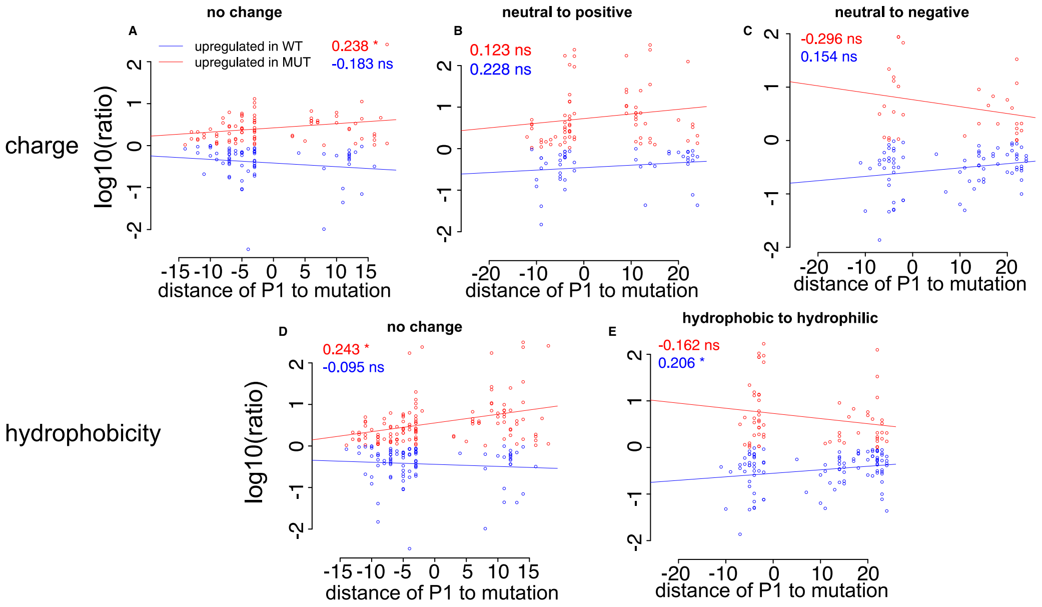


FIGURE 49: Distance of p1 residue to the position with the mutation vs log10 signal intensity ratios of all spliced peptides

The scatter plot depicts the distance of p1 residue of spliced sequence identical products found in both the digestions of WT and MUT polypeptides plotted against the log10 ratio of the intensity of those product peptides for polypeptides (A) in which there was no change in charge, (B) in which a neutral amino acid was replaced with positively charged, (C) in which a neutral amino acid was replaced with negatively charged, (D) in which there was no change in hydrophobicity (E) in which a hydrophobic amino acid was replaced with hydrophilic amino acid. The ratios were computed by dividing the mean signal intensity value of the product peptide from the digestion of MUT polypeptide by the mean signal intensity value of the same product peptide from the digestion of the analogous WT polypeptide. The mean value of those ratios was then computed between all 4 time-points. The product peptides up-regulated upon alteration of the amino acid are marked in red. The product peptides up-regulated in the digestions of WT polypeptides compared to MUT polypeptides are marked in blue. Significant Pearson correlation coefficients between the distances of p1 residue to the position of the mutation and the log10 signal intensity ratios of peptide products found in the digestions of MUT and WT polypeptides are labelled with * (p-value < 0.05).

In addition to the direct comparison of signal intensities, we wanted to determine if there was a connection between signal intensities of peptides upregulated or downregulated in the MUT polypeptides and the proximity to the position with the mutation. To this end, for each given sequence identical peptide, we computed the distance of the position with the mutation to the p1 residue for spliced peptides and C-terminal residue for the non-spliced peptides. Overall, regardless of the peptide type (spliced or non-spliced) and the substrate property group, we didn't observe any strong correlations between proximity to mutation and the signal intensity (Figure 48, 49, Table S21). In some cases, there was a statistically significant correlation but the it wasn't strong enough to suggest any direct connection between distance of p1/pC residues to mutation and the signal intensity. Interestingly p1 of spliced peptides and pC of non-spliced peptides tended to be situated further from the position with the mutation if they were located upstream of it. On the contrary, the peptides for which p1/pC residues were located N-terminally relative to the position with mutation were positioned closer.

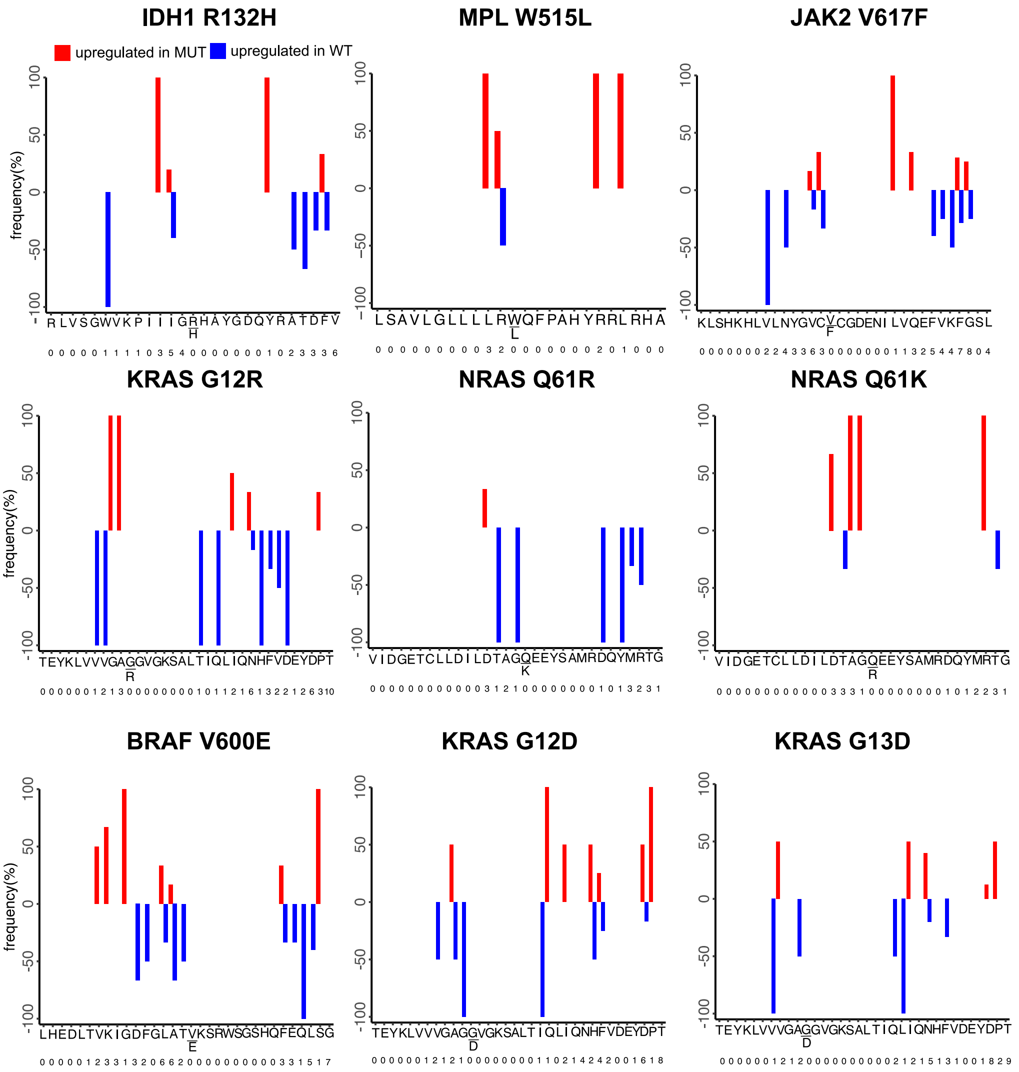


FIGURE 50: P1 frequency of upregulated non-spliced peptides

The barplots depict the frequencies of the sequence identical non-spliced peptides highly upregulated/downregulated in the digestions of WT and MUT polypeptides for each position in the polypeptides that is a pC residue for the corresponding product peptides. The peptides were selected based on whether the log10 signal intensity ratios were above 75% percentile or below 25% percentile of the distribution of log10 signal intensity ratios. Each bar corresponds to the peptides with that pC position in the polypeptide sequence that are upregulated in the digestions of MUT polypeptides (red) or upregulated in the digestions of WT polypeptides (blue). The numbers along the X axis denote the total number of product peptides that were found to have that position in the substrate as pC.

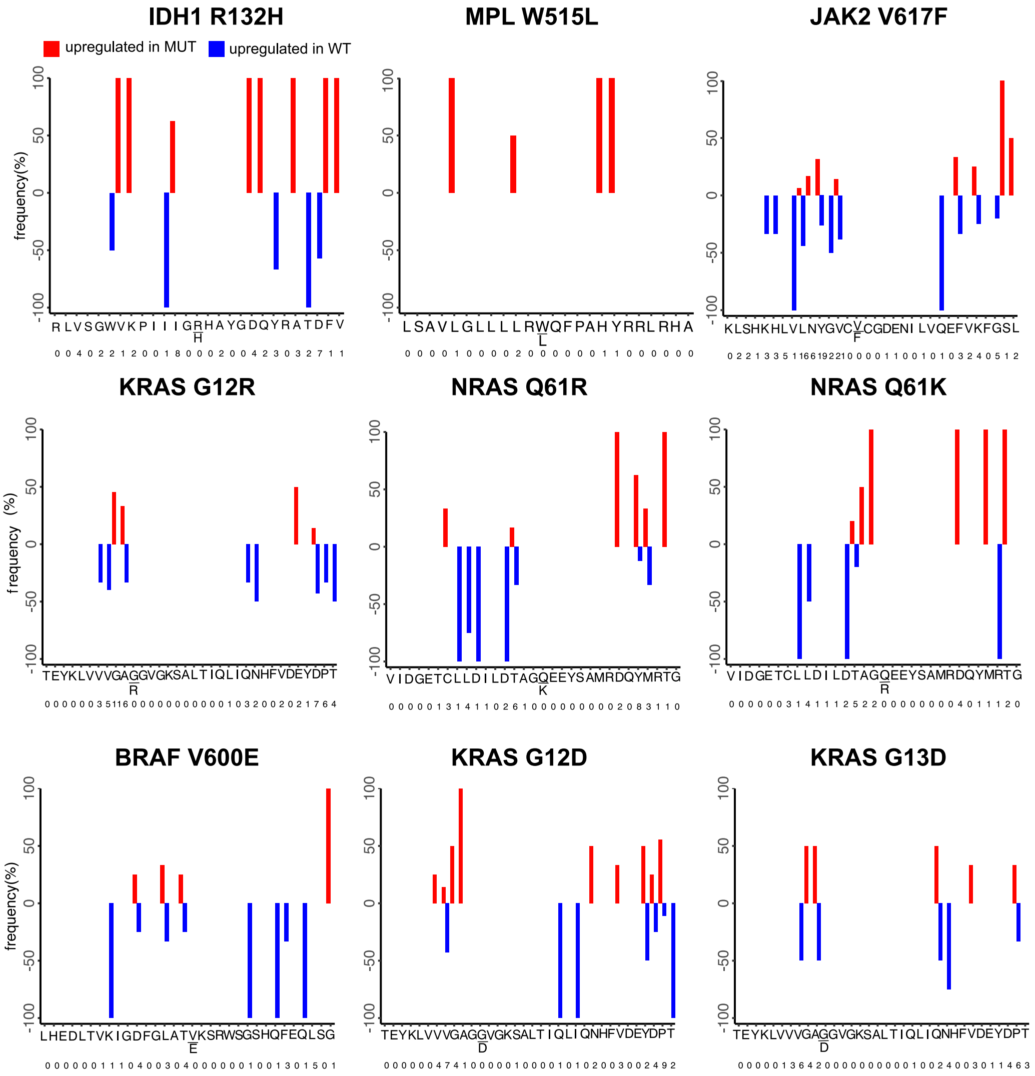


FIGURE 51: P1 frequency of upregulated spliced peptides
The barplots depict the frequencies of the sequence identical spliced peptides highly upregulated/downregulated in the digestions of WT and MUT polypeptides for each position in the polypeptides that is a p1 residue for the corresponding product peptides. The peptides were selected based on whether the log10 signal intensity ratios were above 75% percentile or below 25% percentile of the distribution of log10 signal intensity ratios. Each bar corresponds to the peptides with that p1 position in the polypeptide sequence that are upregulated in the digestions of MUT polypeptides (red) or upregulated in the digestions of WT polypeptides (blue). The numbers along the X axis denote the total number of product peptides that were found to have that position in the substrate as pC.

Finally, we wanted to determine if we could glean any additional insight into the impact of mutations on the proteasomal processing of polypeptides by extracting the sequence identical peptides with very low or very high signal intensity ratios. Our goal was to determine if among polypeptides of any given group there was a tendency to utilise residues of specific properties in the cleavage or splicing reactions of the most highly up-regulated product peptides. We were unable to observe any specific pattern of usage specific types of residues in the digestions of either WT or MUT polypeptides nor could there be inferred a tendency for the p1/pC residues of upregulated/downregulated residues to be located close or distanced to the position with the mutation (Figure 50, 51). We did take note that the majority of residues where upregulation/downregulation occurred were uncharged residues - both hydrophobic and hydrophilic in equal measure.

6.3.3 Frequencies of cleavage and splicing reactions

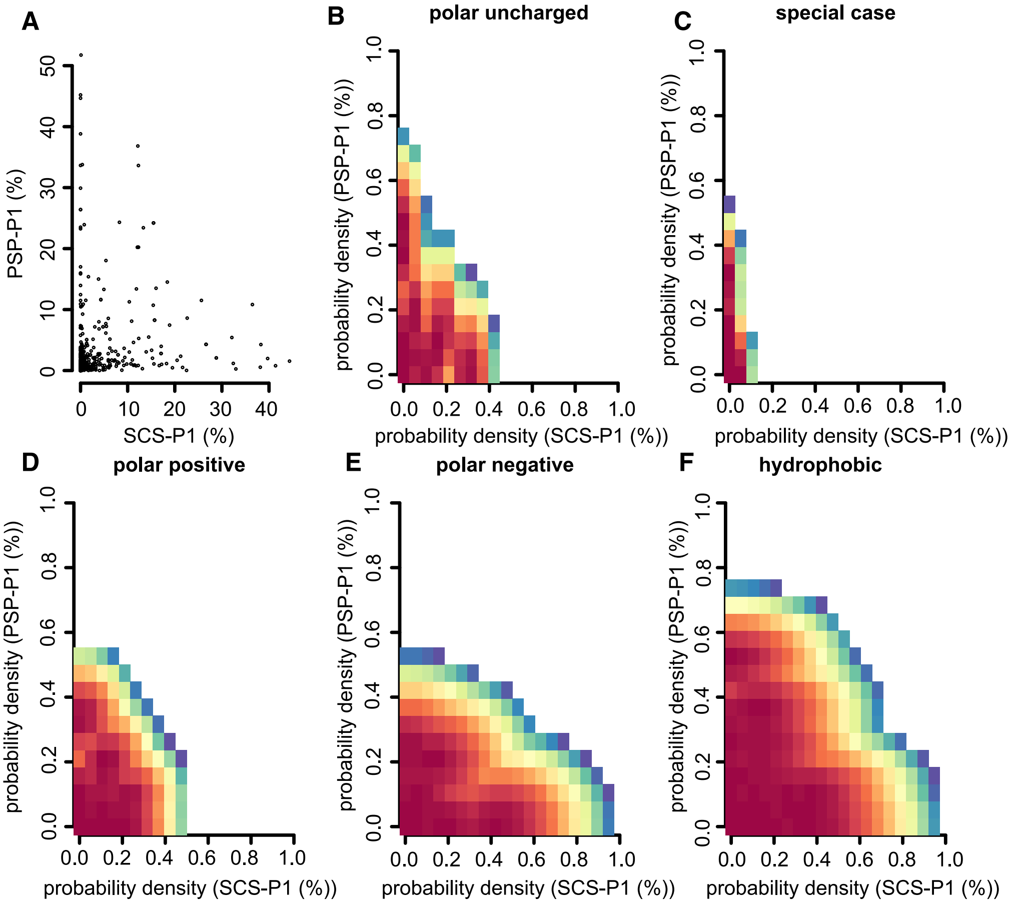


FIGURE 52: residues densities

(A) The scatter plot depicts the SCS-P1 (%) frequencies of cleavage after each position in the polypeptides plotted against the PSP-P1 (%) frequencies of splicing after each position in the polypeptides ggregated across all polypeptides in WT-MUT digestion set (B-F) Heat maps showing the probability of amino acid residues of specific property ((B) - polar uncharged, (C) - special case, (D) - polar positive, (E) - polar negative, (F) - hydrophobic residues) being frequently used in cleavage/splicing reactions for all of the polypeptides in WT-MUT digestion set.

Next, we aimed to determine if there were preferences for utilisation of specific amino acid residues for cleavage or splicing reactions overall for all of the polypeptides included into our WT-MUT set and whether there were any pronounced differences in residue usage by the proteasome in the digestion of WT or MUT substrates. Previously, several groups had claimed to determine the rules governing cleavage and splicing reactions [17, 18, 19].

To do so, for each position in the digested polypeptides we computed the frequency of usage of those residues for either canonical peptide bond breakage or for the ligation reaction of p1 residue of SR1 to p1' residue of SR2. Due to the current lack of data on the chemical amounts of the generated product peptides, the frequencies were computed based on the raw signal intensities. Previously, Mishto *et al*. had observed that the sites after cleavages occur most frequently don't necessarily overlap with splicing reactions suggesting distinct sequence preferences of the proteasome when it comes to the catalysis of either reaction [27]. As such first we compared the PSP-P1 frequencies with SCS-P1 frequencies for each given substrate. As others had noted previously, there was a large cluster of positions with low frequencies of either cleavage and splicing (Figure 52A). Yet, there were other positions with distinctly high frequencies of cleavage and low frequencies of splicing and vice versa thus confirming that the major cleavage/splicing sites often don't overlap.

Following that, we determined whether there were tendencies to use amino acid residues with particular physico-chemical properties in either cleavage or splicing reactions. This was determined by computing the probability densities of residues of specific category being frequently used either as p1 residues in splicing reactions or pC residues in cleavage reactions. What appeared to be the case was that the intensity of usage of polar uncharged residues (S, T, N, Q) and special case residues (C, S, P) was higher in the splicing reactions compared to peptide-bond cleavages (Figure 52B,C). Additionally, polar positive residues (R, H, K) were used slightly more frequently in splicing reactions (Figure 52D). On the other hand, there was a tendency to use polar negative residues (D, E) more frequently for cleavage reactions compared to splicing (Figure 38 E). We didn't however observe clear preferences in terms of utilisation of hydrophobic residues which constitute a plurality of all used amino acids (A, V, I, L, M, F, Y, W) (Figure 52F).

To further zero in on specific amino acids contributing mainly to cleavage and splicing reactions, we extracted the frequencies of cleavage and splicing reactions for each position across all the polypeptides in our digestion set and compared their distributions for either general classes of residues (*i.e* (Figure 53 and and Table S22). hydrophobic, special case, polar uncharged, polar positive and polar negative) or for each amino acid individually. When comparing groups of amino acids, there were significant differences in the distributions for each subset (Figure 53A). The median values for some hydrophobic residues were somewhat higher in the splicing reactions, while others were used more for cleavage reactions indicating comparable rate of usage of those residues (Figure 53B), despite overall similar frequency distribution, which was nevertheless significantly different between cleavage and splicing reactions (Figure 53A). Interestingly, there was a statistically significant higher usage of L and V for splicing reactions (Figure 53B). In addition, polar negative residues (D and E) were used significantly more frequently for cleavage reactions based on the distributions despite comparable medians (Figure 53A and Figure 53F). On the other hand, special case residues, polar uncharged residues and polar positive residues were overall used significantly more frequently in splicing reactions largely mirroring the analysis described above (Figure 53A), despite unequal frequencies of cleavage/splicing after individual residues (Figure 53C, 53D, 53E).

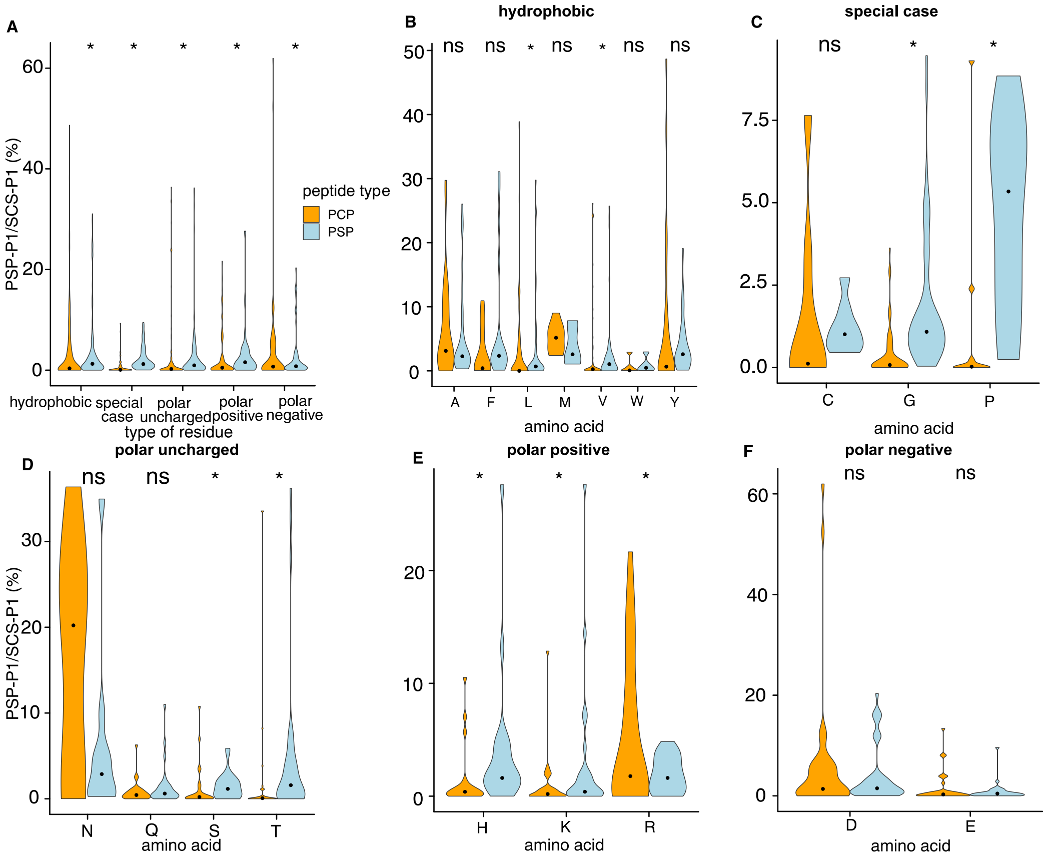


FIGURE 53: SCS-P1 and PSP-P1 frequency comparison
The violin plots depict the distributions of SCS-P1 (%) and PSP-P1 (%) frequencies across all of the WT-MUT polypeptides for each group of amino acid residues based on properties as well as for the separate amino acid residues. (A) Distributions of frequencies of cleavage/splicing for groups of residues based on their chemical properties. (B-F) Distributions of frequencies of cleavage/splicing after each amino acid residues. (B) Frequency distributions for hydrophobic residues (C) Frequency distributions for special case residues (D) Frequency distributions for polar uncharged residues (E) Frequency distributions for polar positive residues (F) Frequency distributions for negatively charged residues. Black dots represent the median. Orange color denotes non-spliced peptides. Light blue color denotes spliced peptides. Significant difference between SCS-P1 and PSP-P1 frequencies are labelled with * (ks.test; p-value < 0.05).

In terms of specific residues, among the hydrophobic residues Vs and Ls are used significantly more frequently in splicing reactions. Despite the clearly higher median frequency of A, M and Y there was no significant difference in usage of those residues for either of the two reactions likely due to their low incidence among the polypeptide sequences included into our dataset (Figure 53B). Among special case residues, G and in particular P are used dramatically more frequently in splicing reactions (Figure 53C). Among the polar uncharged residues, S and T are used significantly more frequently in splicing reactions (Figure 53D). Interestingly, N has a much higher median frequency of cleavages than splicing reactions which didn't, however, result in a significant difference in the distributions likely due to the low overall frequency of Ns in the digested polypeptide sequences. Among the positively charged residues, H and K are used significantly more frequently in splicing reactions while R on the other hand is utilised more in cleavage reactions (Figure 53E). Finally, there are no significant differences in either of the two negatively charged residues despite them being used overall more frequently in splicing reactions when considered together (Figure 53F).

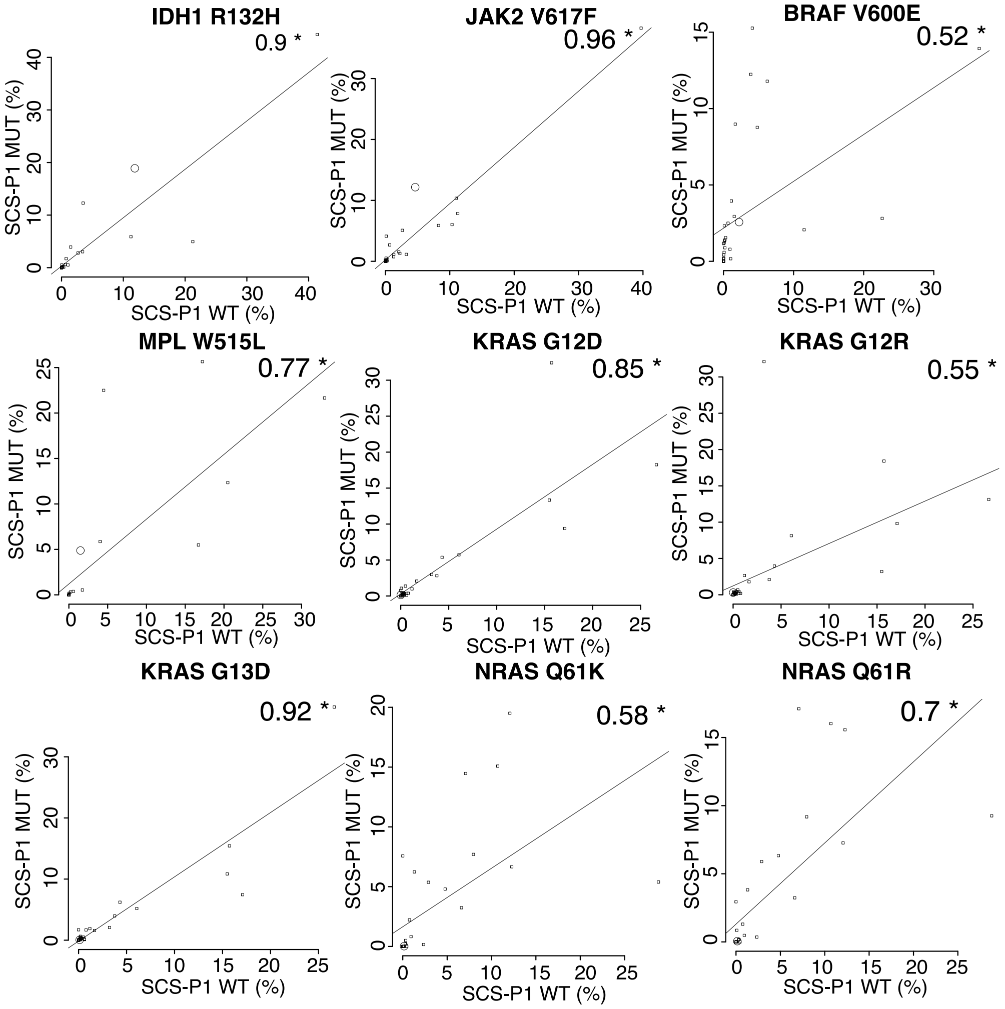


FIGURE 54: SCS-P1 comparison for the digestions of WT and MUT polypeptides

The scatter plot depicts the SCS-P1 (%) values of frequencies of cleavage after each amino acid position computed for the digestions of WT polypeptides plotted against the SCS-P1 (%) frequencies computed for the digestions of the analogous MUT polypeptides. The position with the mutation is labelled as large circle. Significant Pearson correlation coefficients between the SCS-P1 values are labelled with * (p-value < 0.05).

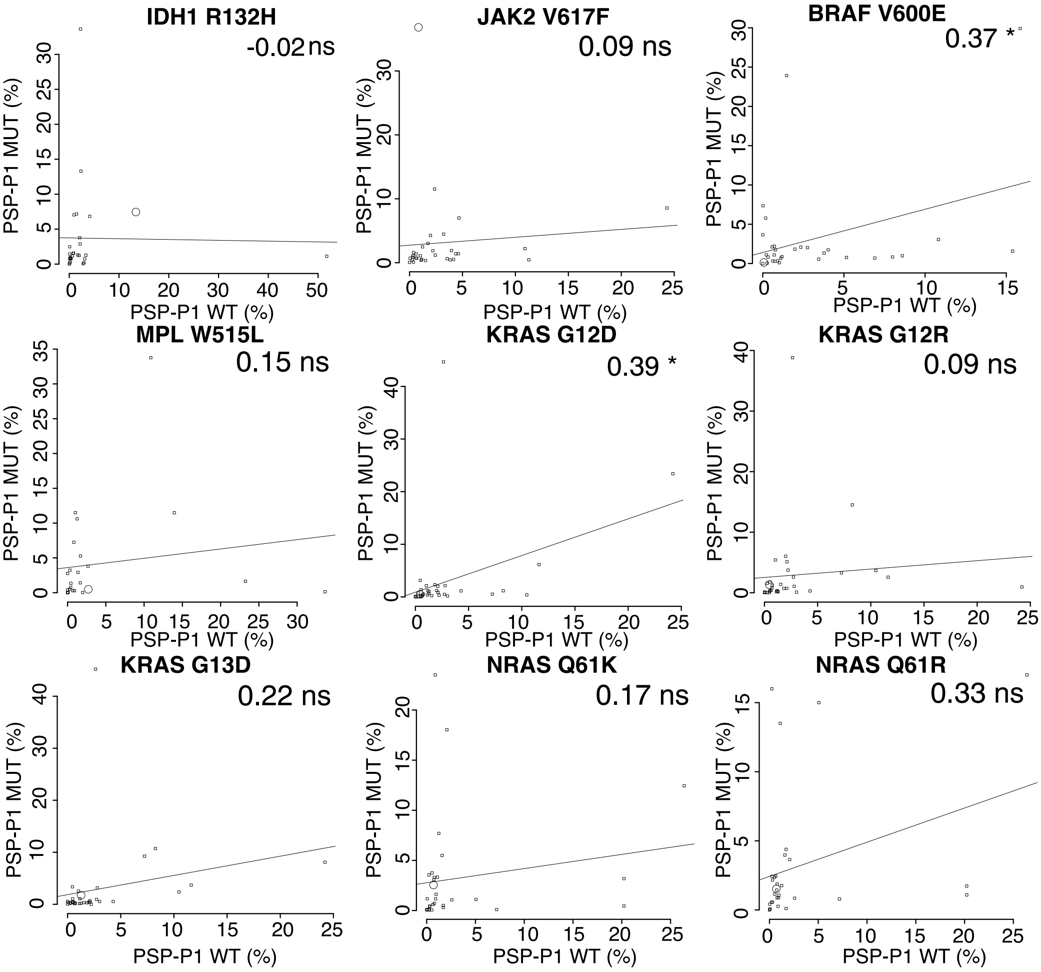


FIGURE 55: PSP-P1 comparison for the digestions of WT and MUT polypeptides

The scatter plot depicts the PSP-P1 (%) values of frequencies of splicing after each amino acid position computed for the digestions of WT polypeptides plotted against the PSP-P1 (%) frequencies computed for the digestions of the analogous MUT polypeptides. The position with the mutation is labelled as large circle. Significant Pearson correlation coefficients between the PSP-P1 values are labelled with * (p-value < 0.05).

Next, we aimed to determine if there was a concordance between the same residues used in cleavage/splicing reactions in the digestions of both WT and MUT substrates. To this end we compared the frequencies of cleavage/splicing after each amino acid residue for all WT-MUT pairs of polypeptides. There was a large discrepancy in correlations for peptide bond cleavage and peptide splicing reactions. Regardless of the WT-MUT pair there was a statistically significant moderate to strong correlation between cleavage frequency values for the same residues in WT or MUT polypeptide pairs suggesting that perhaps the main overall effect of the mutations is not elicited through changes in cleavage preferences and that the overall catalytic activity of the 20S proteasome is comparable for different polypeptides regardless of their mutation status (Figure 54 and Table S23). Remarkably, for the peptide splicing reactions there was no significant or strong correlation between the same residues found in both WT and MUT polypeptides.

The correlation was never higher than 40% which could potentially indicate that the the single point mutations primarily change the configuration of splicing but not canonical cleavage preferences (Figure 55 and Table S24).

Finally, we computed and compared the distributions of frequencies of cleavage and splicing after amino acids of each property group between WT and MUT polypeptides for each group of peptides based on the transitions of properties of amino acids being permuted (Figure 56 and Table S25). In accordance with the analysis described above, in terms of frequencies of peptide bond cleavages there was no statistically significant differences in frequencies of usage of amino acids of any type in the digestions of WT and MUT polypeptides with very comparable median values, with the exception of special case residues among a group of polypeptides in which hydrophobic residues were replaced with hydrophilic (Figure 56E).

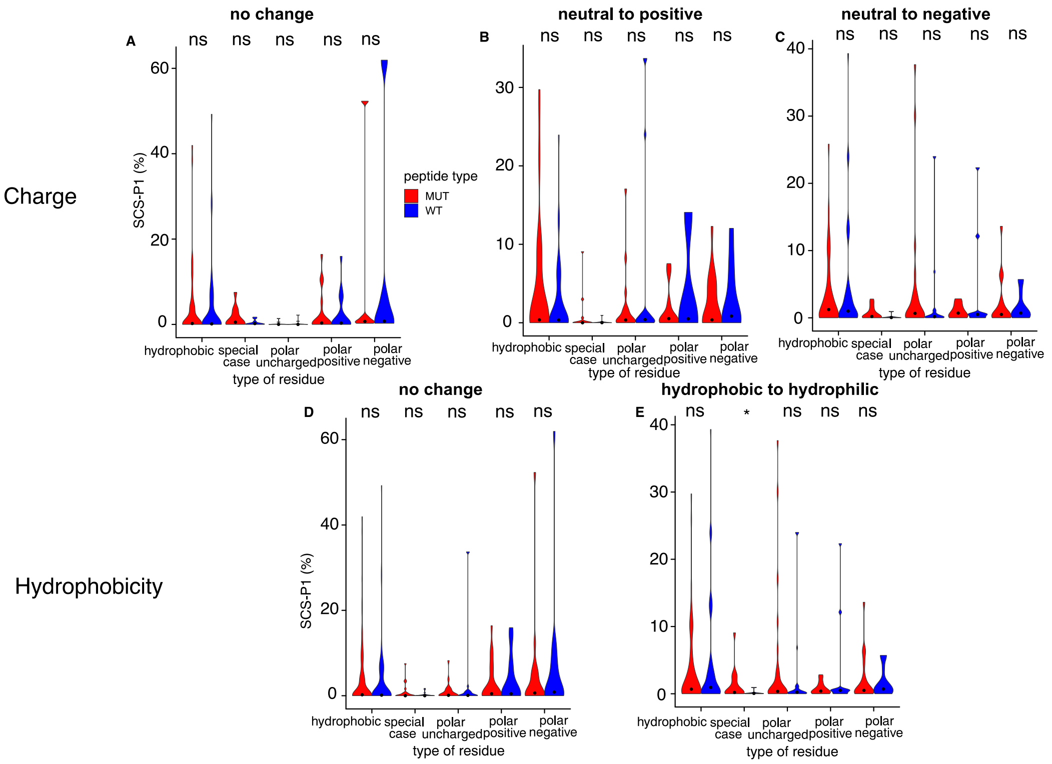


FIGURE 56: SCS-P1 frequency comparisons of the digestions of WT and MUT polypeptides per group of amino acid residues

The violin plots depict the distributions of SCS-P1 (%) frequencies for each group of amino acid residues based on properties for polypeptides (A) in which there was no change in charge, (B) in which a neutral amino acid was replaced with positively charged, (C) in which a neutral amino acid was replaced with negatively charged, (D) in which there was no change in hydrophobicity (E) in which a hydrophobic amino acid was replaced with hydrophilic amino acid. Black dots represent the median. Red color denotes MUT polypeptides. Blue color denotes WT polypeptides. Significant difference between SCS-P1 frequencies in the digestions of WT and MUT polypeptides are labelled with * (ks.test; p-value < 0.05).

When we considered frequencies of splicing after each residue for WT-MUT polypeptide pairs, despite the lack of statistical significance in the frequencies distributions, we did observe a noticeable differences in the median values for some groups of residues for each group of properties (Figure 57 and Table S26). The median frequency values were higher in all of the groups of residues for MUT substrates except for polar negative residues when considering polypeptides in which the change in the amino acid doesn't lead to the change in charge (Figure 57A). Among the group of polypeptides in which neutral residue was replaced with positively charged residue, there was a higher median frequency value for polar uncharged and polar positive residues in the digestions of MUT polypeptides (Figure 57B). When doing the opposite and replacing the the neutral amino acid with a negatively charged one, there was a significantly higher usage of hydrophobic and polar uncharged residues in the digestions of WT polypeptides compared to MUT polypeptides (Figure 57C). The median frequency value was also noticeably higher for polar positive residues in the digestions of WT residues, however the difference in the overall distribution wasn't statistically significant. The apparently overall more efficient frequency of splicing after those residues is in line with the general increase of the rate of production of sequence identical peptides that we observed in the comparison of signal intensity ratios for this group of polypeptides. For the polypeptides split based on the hydrophobicity of the altered residue, for the group of polypeptides, where there was no change in hydrophobicity, there was a higher median frequency value of hydrophobic, polar uncharged and polar positive residues despite the lack of statistical significance (Figure 57D). For the group of polypeptides, in which hydrophobic residue was replaced with hydrophilic, there was a statistically significant higher usage of hydrophobic and polar uncharged residues for splicing reactions in the digestions of WT polypeptides (Figure 57E). The general absence of the statistically significant differences in the frequency distributions can most likely be explained by the small number of polypeptide pairs (3-5) included in each charge/hydrophobicity group resulting in the insufficient number of product peptides and the resulting lack of data-points (amino acids) based on which the frequencies of cleavage/splicing were calculated. It's likely that the differences in the distributions would become more prominent provided more polypeptide digestions are included into each group.

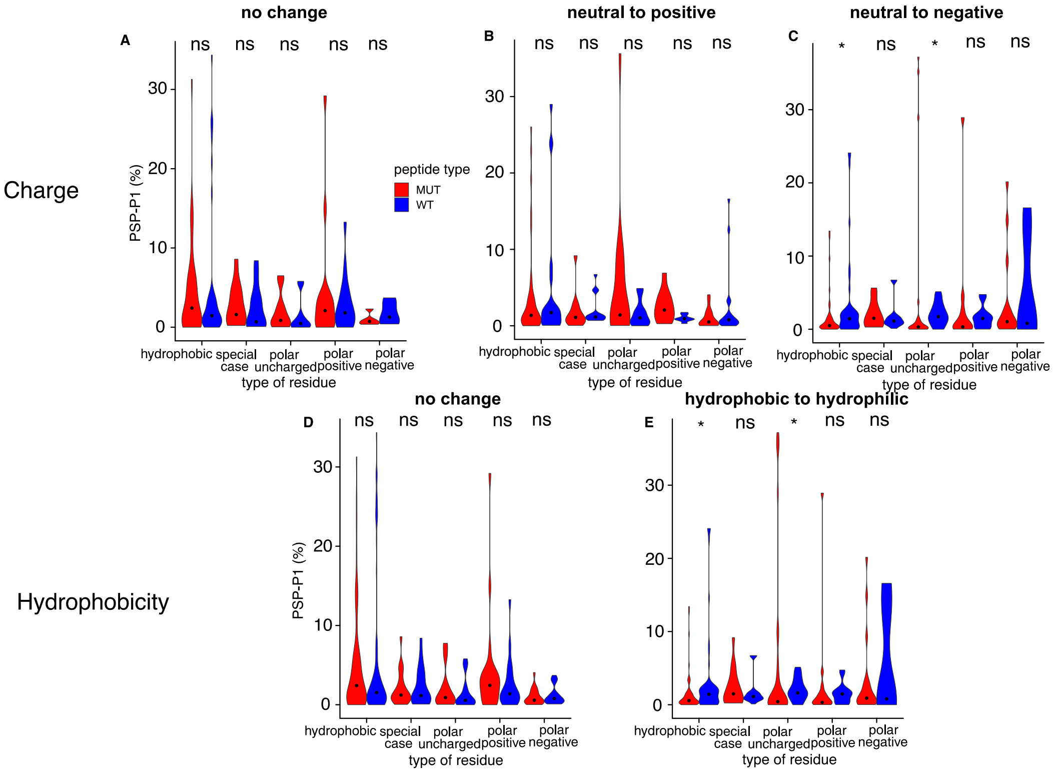


FIGURE 57: PSP-P1 frequency comparisons of the digestions of WT and MUT polypeptides per group of amino acid residues

The violin plots depict the distributions of PSP-P1 (%) frequencies for each group of amino acid residues based on properties for polypeptides (A) in which there was no change in charge, (B) in which a neutral amino acid was replaced with positively charged, (C) in which a neutral amino acid was replaced with negatively charged, (D) in which there was no change in hydrophobicity (E) in which a hydrophobic amino acid was replaced with hydrophilic amino acid. Black dots represent the median. Red color denotes MUT polypeptides. Blue color denotes WT polypeptides. Significant difference between PSP-P1 frequencies in the digestions of WT and MUT polypeptides are labelled with * (ks.test; p-value < 0.05).

6.3.4 Epitope candidate selection

As a proof of principle, we aimed to determine if any potential non-spliced or cis-spliced peptides that were identified in the 20h digestions of the MUT polypeptides substrates could also be quantified. This quantified epitope candidates could than be used in the further validation experiments aimed to determine the best candidates for the adoptive anti-tumor CD8+ T-cell therapies. To do this, we first predicted the binding affinities of all theoretical peptide products that would carry the mutation to the variety of HLA molecules. Then, we analysed 20h digestions of the MUT polypeptides and identified those putative binders that were detected in the 20h digestions. Following that, we surveyed the quantified non-spliced and cis-spliced peptides that passed the kinetic filtering and determined how many of the epitope candidates found in 20h digestions were indeed quantified. Finally, we checked how many of the peptides that could be quantified had an optimal kinetic behaviour. Crucially, we considered a given epitope candidate identified and quantified either if the minimal epitope was detected or if at least an N/C-terminal precursor of a given epitope no larger than by 5 amino acids compared to the minimal epitope was found. This is because even if a desired epitope isn't readily produced by the 20S proteasome, the N/C-terminally extended precursor may still be additionally trimmed afterwards either by the proteasome itself or by cytosolic/ER resident aminopeptidases in case of N-terminally extended precursors. Than means that while each given minimal epitope candidate is unique, one N/C-terminally extended precursor could be attributed to multiple different minimal epitopes due to the precursor carrying sequence of each of the possible minimal epitopes. If the precursor is further processed afterwards, any one of the possible minimal epitopes could be released as a result.

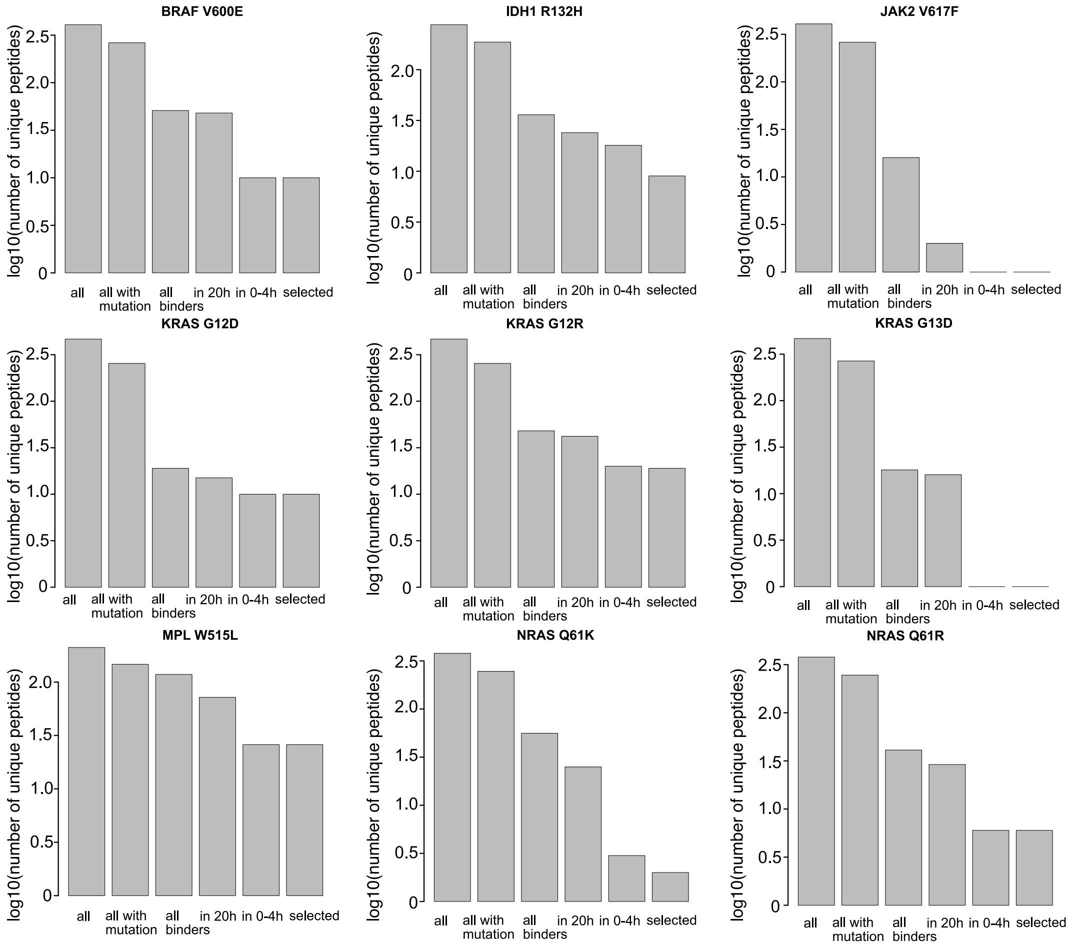


FIGURE 58: Comparison of the numbers of non-spliced that are predicted binders, identified in 20h digestions, quantified in the kinetics and selected based on the kinetic behaviour

The bar-plots depict the total number of unique peptides, those that carry the position with the mutation, those that could be HLA binders, those that were identified in the 20h digestions, those that were quantified and those that were selected on the basis of their generation dynamics. The numbers of unique peptides were counted for each MUT polypeptide individually.

As expected, the number of putative epitope candidates that were identified in the 20h digestions, was much lower than the theoretical number of putative binders and the number of epitope candidates that were quantified and selected based on their kinetic properties was considerably lower than the number of epitope candidates found in 20h digestions (Figure 58 and Figure 59; Table 5 and Table 6).

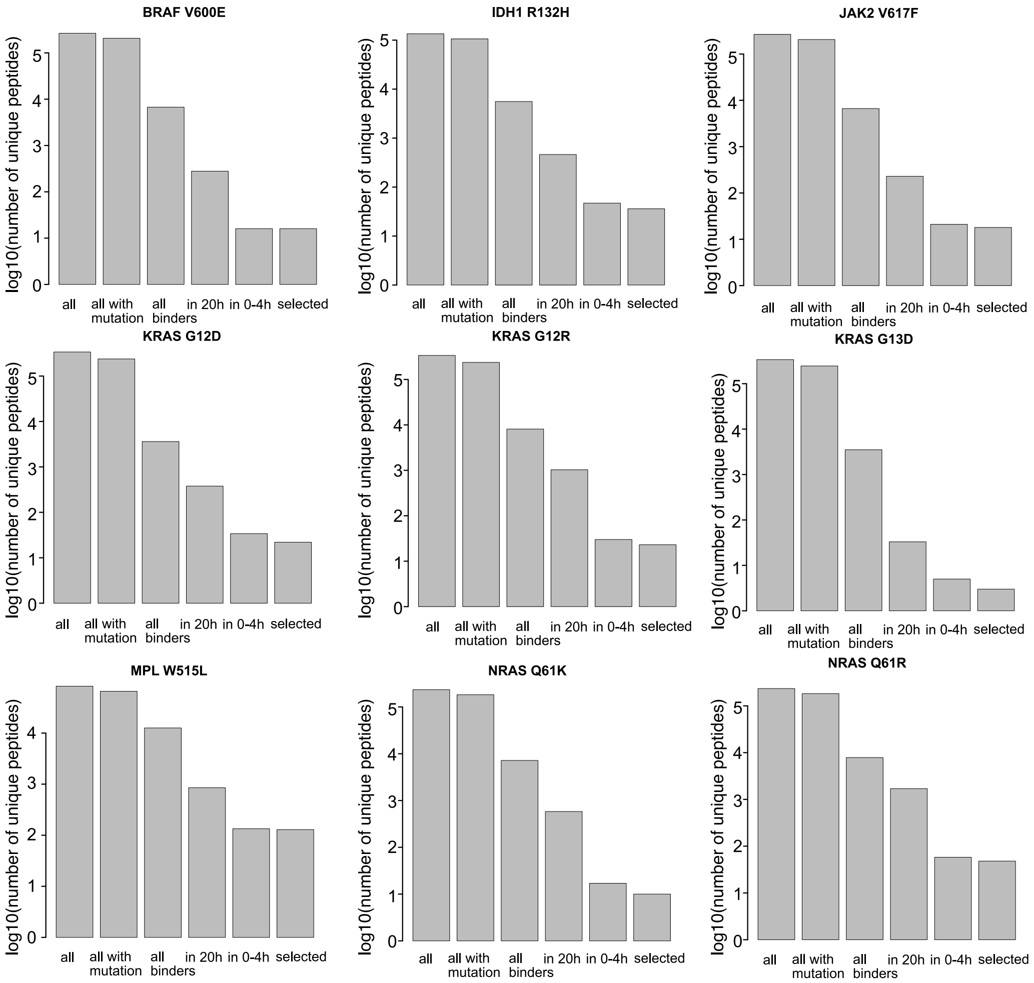


FIGURE 59: Comparison of the numbers of cis-spliced that are predicted binders, identified in 20h digestions, quantified in the kinetics and selected based on the kinetic behaviour

The bar-plots depict the total number of unique peptides, those that carry the position with the mutation, those that could be HLA binders, those that were identified in the 20h digestions, those that were quantified and those that were selected on the basis of their generation dynamics. The numbers of unique peptides were counted for each MUT polypeptide individually.

To understand the extend of the decrease of the number of peptides during the filtering, IDH1 R132H polypeptide is a good example. In total, 134604 peptide products, both non-spliced and spliced, could be derived from the digestion of this polypeptide by the proteasome. Out of those, 105942 peptides could contain position with the mutation. When we performed affinity prediction of all of the theoretically possible 8-14mers containing the mutation to a variety of HLA molecules, out of this great number of peptides, just 5583 could be considered theoretical binders at IC50 cut-off of 5000 nM. The sharp drop in the number of theoretical binders compared to all of the peptides with the mutation can be explained by the sequence length limit imposed by MHCI binding groves and by the overall sequence requirements by MHCI molecules towards potential epitopes. Next, we attempted to determine which of the theoretical HLA binders could in fact be identified in the 20h digestions. There was a drastic decrease in the number of the epitopes identified in the *in vitro* digestions - just 461 out of 5583. That is considering not just the minimal epitopes but also the N/C-terminally extended precursors. In the 0-4h digestion kinetics, only 46 out of those 461 peptides were detected and finally 35 of those quantified peptides passed the kinetic selection criteria that we imposed. This number would further decrease if we were to make the selection criteria more strict. It is notable that in most cases, a large fraction of quantified epitope candidates were cis-spliced and not non-spliced peptides, confirming our prior hypothesis that cis-spliced peptides could substantially enlarge the antigenic repertoire of MHCI molecules. Yet there are other instances when a relatively significant share of quantified peptides are non-spliced. In the IDH1 R132H, for example, 17 of the quantified peptides are non-spliced, while 46 are cis-spliced. Nevertheless, cis-spliced peptides still constitute 73% of all quantified peptides in the IDH1 R132H digestions. Finally, we checked how many of the quantified epitope candidates could be present in the human proteome as either non-spliced and cis-spliced peptides. This will be important in the further validation experiments since the peptides matching the human proteome regions could either theoretically elicit a strong autoimmune response or alternatively be ignored by the immune system due to the removal of T-cells that could recognise them during negative selection in the thymus, thus making them unfit for the T-cell therapies. The share of such peptides varied from polypeptide for polypeptide and for the most part constituted a relatively small share of all of the quantified peptides in total (Table 7). The only exception to this was BRAF V600E digestions, in which 40% of all putative epitope candidates were found in human proteome as cis-spliced peptides.

Table 5: Number of unique non-spliced peptides and epitope candidates that could be derived from each polypeptide

Substrate and mutation	N of all possible peptides	N of peptides with mutation	N of HLA binders	N of HLA binders in 20h	N of HLA binders in the kinetics	N of selected HLA binders
BRAF V600E	406	262	50	47	9	9
IDH1 R132H	276	186	35	23	17	5
JAK2 V617F	406	260	15	1	0	0
KRAS G12D	465	254	18	14	9	9
KRAS G12R	465	254	47	41	19	18
KRAS G13D	465	266	17	15	0	0
MPL W515L	210	146	117	71	25	25
NRAS Q61K	378	245	55	24	2	1
NRAS Q61R	378	245	40	28	5	5

Table 6: Number of unique cis-spliced peptides and epitope candidates that could be derived from each polypeptide

Substrate and mutation	N of all possible peptides	N of peptides with mutation	N of HLA binders	N of HLA binders in 20h	N of HLA binders in the kinetics	N of selected HLA binders
BRAF V600E	265653	207011	6733	278	15	15
IDH1 R132H	134604	105942	5583	461	46	35
JAK2 V617F	265505	205919	6651	228	20	17
KRAS G12D	338655	237853	3618	377	33	21
KRAS G12R	338655	237864	8108	1026	29	22
KRAS G13D	338764	245498	3505	32	4	2
MPL W515L	82718	65711	12596	852	133	128
NRAS Q61K	233895	182220	7192	581	16	9
NRAS Q61R	234488	182765	7847	1700	57	47

Table 7: Fraction (%) of quantified epitope candidates found in the human proteome

Substrate and mutation	% not found in human	% as non-spliced	% as cis-spliced with 25intv	% as cis-spliced without restriction	% as non-spliced	% as cis-spliced with 25intv	% as cis-spliced without restriction
BRAF V600E	57.4	0	0	42.9	0	0	42.8
IDH1 R132H	95.24	0	0	4.76	0	0	0
JAK2 V617F	100	0	0	0	0	0	0
KRAS G12D	100	0	0	0	0	0	0
KRAS G12R	83.3	11.1	16.7	16.7	11.1	16.7	16.7
KRAS G13D	100	0	0	0	0	0	0
MPL W515L	79.24	0	7.55	20.2	0	0	42.8
NRAS Q61K	100	0	0	0	0	0	0
NRAS Q61R	88.9	0	0	11.1	0	0	11.1

6.4 Discussion

In the recent years it has become clear that spliced peptides could substantially contribute to the diversity of MHCI immuno-peptidomes and thus the adaptive immune response [13, 14, 15] . As a consequence, the identification of spliced peptides that could bypass sequence limitations imposed on non-spliced peptides and that possess optimal MHCI binding properties became a tantalising possibility for enhancing the utility of the novel CD8+ T-cell immunotherapies and peptide-based vaccines. Several groups demonstrated the utility of the approaches aimed at the identification of non-spliced epitope candidates in the *in vitro* digestions of polypeptides containing for example tumor-specific mutations enhancing the MHCI binding affinity of an antigenic peptide and promoting a response of T-cells non-cross reactive with the WT antigens [12, 22, 44, 72]. Those identified epitopes could subsequently be used as targets in the adoptive T cell anti-tumor therapies. However, to this day little is known about the effect of such mutations on the proteasomal processing of antigens containing tumor-specific mutations. This, coupled to the lack of a complete understanding of the mechanisms and substrate sequence preferences driving canonical cleavage and splicing reactions makes the predictions of which antigenic peptides be produced untenable.

Our quantitative analysis based on signal intensities demonstrated that the effect of such single point mutations was wide ranging and highly heterogenous depending on the amino acid composition of each given polypeptide. The first obvious effects were observed in terms of the substrate degradation rates of different polypeptide pairs. While the relative degradation rate was similar in many cases, there were multiple instances in which the introduction of the mutation led to a noticeable decrease of the substrate degradation. The most likely explanation for that is that the strong cleavage sites within the polypeptide sequences that could be utilized by the proteasome early on were neutralised by the introduction of single point mutations. Next, we compared the generation kinetics of a subset of product peptides. Due to the unpredictable nature of even single amino acid substitutions on the ionisation of the product peptides during ESI, we elected to focus the first part of our comparative analysis on the sequence identical peptides. Our rationale for that was that since the sequence of the peptides weren't altered, we could compare the signal intensities of such peptides directly.

With this study we attempted to analyse the effect of such single-point mutations by analysing various aspects of the *in vitro* proteasomal digestions of a multitude of synthetic polypeptides derived from the known tumor associated antigens, possessing a variety of sequence characteristics. Importantly, it was possible to divide the polypeptide substrates in our complete set into groups based on the properties of amino acids that were replaced in the mutated substrates. To study the impact of mutations, we conducted both qualitative and semi-quantitative comparison of peptide products generated in the digestions of WT-MUT substrate pairs by utilising peptides' signal intensities obtained via LFQ quantification. In terms of the qualitative comparison of peptide products we observed similar median lengths of peptide products and SRs. In addition, we observed significant pairwise correlations between the number of unique peptide products detected in the digestions of both WT and MUT substrates. This suggests that in terms of the number of unique peptide products the digestions are quite comparable. This also suggests that the primary effects of the single-point substitutions were quantitative in nature. Instead of grouping all polypeptides together, we elected to split them into subgroups based on the transition of physico-chemical properties of the permuted amino acids. We decided to focus on two such properties - charge and hydrophobicity. We reasoned that even though each polypeptide pair that was included in our set had variable amino acid composition, barring cases where we were analysing the digestions of polypeptides, for which several single point mutations are known (*i.e.* KRAS and NRAS), and despite the decrease of the sequence space as a consequence of splitting of polypeptides into groups, it would be possible to draw some preliminary conclusions in regards to the effect of such constellation of mutations on the generation of peptide products. Overall, the presence of mutation affected the rate of generation of sequence identical peptides, indicating the effect of the mutation on the splicing preferecnes. Both charge and hydrophobicity of the amino acids are important determinants of the efficiency of cleavage/splicing reactions after P1 residues catalysed by the proteasome [17, 18, 19]. Interestingly, the most prominent changes in the digestions were observed when we split the polypeptides into groups based on the charge state. The prominent effects of the mutation observed in the digestions split by charge state, became less so when we performed the comparisons of polypeptide based on hydrophobicity. This suggests that the charge of the amino acids has a larger effect on the generation of peptide products. This raises questions regarding which property of the amino acids is the most important for determining the efficiency of cleavage/splicing reactions. Does this mean the hydrophobicity has a lesser impact on the efficiency of digestion than charge? This is likely not the case as it's known that different proteasome subunits preferentially cleave after different groups of amino acids depending on the catalytically active beta subunits. Beta5 subunit typically cleaves after hydrophobic residues, beta2 - after positively charge residues and beta1 - after acidic residues [18]. Thus both charge and hydrophobicity play a role in determining cleavage/splicing specificity. The alternative explanation for our observations, is that the presence/absence of charge is a more qualitative characteristic while hydrophobicity is a more quantitative feature. This is because different amino acids possess different hydrophobicity scores. As such any quantitive effect of on the efficiency of the digestions would be significantly different depending on the hydrophobicity value of a given amino acid substitution. As a consequence, when grouping broadly hydrophobic/hydrophilic residues together the effects that we might have observed were diluted due to significantly different impact of residues with slightly different hydrophobicity ratings on cleavage/splicing preferences. To better understand the impact of hydrophobicity on the generation of peptide products, it would be advisable to increase the number of digested polypeptide pairs and further split them into groups based on residues with similar hydrophobicity scores.

The other important amino acid characteristic that we didn't consider in this analysis was the size of altered amino acids. The size of the amino acids could be important for the determination of the binding properties of the substrate in the non-primed and primed sites in the catalytic chamber of the proteasome. Even one substitution could have a profound effect on the binding properties of the sequence motifs of the polypeptides, even those located distantly to the mutation site. Thus, the impact of the mutations would likely be heterogeneous depending on the size of any given amino-acids and grouping substrates based on the rough size estimate of the altered amino acids (*e.g.* small/medium/large) could dilute the effect of the mutation on the product generation dynamics due to highly unequal effect of small alterations of amino acids size on the configuration of binding of the substrates in the proteasome catalytic chamber.

This limitation didn't appear as crucial when considering the impact of the charge state. We didn't observe substantial changes in the kinetics of peptide products in the group of polypeptides in which the substitution didn't result in a change in charge. On the contrary, there was a strong positive effect of a neutral to positively charge substitutions on the generation of reverse-cis spliced peptides. Remarkably, we also observed a consistent negative impact of the group of neutral to negatively charged substitutions on both non-spliced and spliced peptides. This may seem counterintuitive, since negatively charged residues in P1 position were pinpointed by Berkers *et al.* as being beneficial for both cleavage and splicing reactions, while positively charge residues in P1 positions weren't the drivers of splicing reactions [18]. However, in this analysis we didn't consider the kinetics of product peptides containing the mutation, thus we weren't evaluating the generation dynamics of the peptides with those altered residues. What is possible, however, is that even if substitution in one position enhances/decreases the rate of cleavage or splicing reactions after that specific residue, it doesn't mean that this effect is uniformly applied across the entire polypeptide sequence. Rather, the effect of any given amino acid substitution is highly complex. The fact that we observed a consistent significant increase of the generation dynamics of reverse-cis spliced peptides, when the neutral amino acids were replaced with positively charged ones and a consistent significant decrease in the production of trans-spliced, reverse cis-spliced and non-spliced peptides, when the neutral amino acids were replaced with negatively charged ones, suggests that such substitutions likely lead to the altered utilisation of specific sequence motifs in the polypeptide sequences. Moreover, the utilisation of such sequence motifs could be tied to both the broad type of produced peptide (non-spliced or spliced) and moreover, the specific category of generated spliced peptides. Paes *et al*, pointed out that P1 residues used in creation of normal-cis spliced peptides and reverse cis-spliced peptides as well as the pairings of residues in SR1-SR2 didn't completely overlap [19]. As such, it would be prudent to investigate the specificities of P1 residues of peptides upregulated in the neutral to positive charge group and downregulated in theneutral to negative charge group. In addition, the sequence motifs of SR1 and SR2 will have to be investigated in order to pinpoint patterns in polypeptides sequence usage that enable such upregulation/downregulation. For example, introduction of a positively charged residue led to the utilization of specific P1 group of residues in reverse cis-splicing reactions that are otherwise underused in other splicing reactions. Changes in the polypeptide transport across proteasomal cavities could potentially occur in response to the introduction of the mutations and should also be investigated in more detail. Crucially, this analysis was conducted based only on a small subset of polypeptides relative to the overall diversity of products present in the digestions of the polypeptides. Thus it will be necessary to determine precise chemical amounts of all of the peptide products present in the digestions in order to perform the comparisons on the entire population of peptide products. Perhaps, we would no longer observe such clear preferences in generation of products of specific types. Ideally, more polypeptide digestions should be included in each group of polypeptides to confirm our observations.

We then considered the relationship between the proximity of p1 residues of non-spliced/spliced peptides and the degree of change of their intensities after the amino acid substitution. We observed that the peptides upregulated/downregulated in the digestions of MUT polypeptides were located both close and distant to the position with the mutation in all of the analysed polypeptides with little to no correlation between the extent or upregulation/downregulation and the proximity to the position with mutation. This suggest that the effects of the mutation could be elicited both at the neighbouring residues of the mutation and distant from it at least based on the kinetics of sequence identical peptides. Of note however, is that p1 residues of spliced peptides and pC residues of the non-spliced peptides tended to be situated further from the position with the mutation if they were located C-terminally relative to it. This could be explained by the fact that the minimal allowed peptide length in our analysis was 6, thus biasing the p1 residues of spliced and particularly non-spliced peptides towards the C-terminus of the substrate.

We reasoned that we could observe the clearest patterns among the most highly for thupregulated/downregulatied peptides. The hope was that the preference for certain residues/distances to mutations would be the most saliant among the most highly upregulated/downregulated peptide products. However, we didn't see any such patterns. It's worth pointing out however that the majority of residues that were upregulated/downregulated, were uncharged regardless of the group of polypeptides that we analysed or their WT or MUT status. This observation probably doesn't suggest a pattern of usage of specifically non-charged residues. Rather it is the consequence of the fact that the majority of the 20 amino acids don't carry a charge. In addition, due to the fact that in our analysis we were limited only to highly upregulated/downregulated sequence identical peptides only a very small fraction of peptides that could use a particular amino acid in the substrate for cleavage/splicing reactions (even smaller than all of the sequence identical peptides) were considered. In turn, this limitation severely curtailed the conclusions that we could draw. On the other hand, if we knew the precise chemical amounts of all of the peptide products in the digestions, both sequence identical and non-identical, that would enable us to perform a comprehensive analysis.

It could potentially reveal clear patterns of usage of specific residues in different groups of polypeptides.

Finally, we aimed to look at the frequencies of cleavage/splicing after p1 residues in all of the digested polypeptides. Previously, Berkers *et al.* and Paes *et al.* performed a series of *in vitro* digestions of synthetic polypeptides and suggested the first set of rules driving the splicing reactions. Berkers *et al.* found that that the primary driver of splicing reactions was the N-terminal splice reactants as it's sequence characteristics predict whether or not it would be able to form an acyl-enzyme intermediate with the catalycially active Thr1 of the proteasome. Specifically, they implicated hydrophobic, polar uncharged and negatively charged residues in position p1 as being the key determinants of the success of the splicing reactions. However the major drawback of that study was a small selection of polypeptides with narrow sequence characteristics likely biasing the authors' conclusions. Berkers *et al.* focused primarily on HLA-A*02:01 binders of length 9 amino acids (AAs) [18]. In general, peptides binding to HLA-A*02:01 molecules are known to be enriched in hydrophobic residues, particularly in their anchor regions. Paes *et al.* performed the *in vitro* digestions of a more diverse set of peptides such as HIV-1 derived polypeptides and a number of self-peptides ranging in length from 14 to 47 AAs, and highlighted a range of amino acids that could be optimal p1 residues in splicing reactions, including basic residues that were not suggested by Berkers *et al.* [19]. The other limitation of the approaches used by Berkers *et al.* and Paes *et al.* was that in their comparisons they relied on the relative intensities of the product peptides which can't be used as a reliable method of quantification due to the diversity of generated peptide sequences and their variable ionisation properties during MS. Nevertheless, we sought to compare their observations with ours. Firstly, as was shown by Mishto *et al.* and others in many cases where the proteasome cleaves isn't where it frequently splices and vice versa as suggested by the lack of correlation in SCS-P1 and PSP-P1 frequency values. Interestingly, in terms of the frequencies of cleavage/splicing after the individual amino acids we obtained the results not dissimilar to what was reported by Paes *et al.* and Berkers *et al.* Specifically, small and polar uncharged residues appeared to be used more frequently for splicing reactions. In contrast to Berkers *et al.*, however, we didn't observe clear preferences for splicing reactions after hydrophobic and polar negative residues. Nevertheless, the median frequency values for cleavage/splicing reactions were similar. This suggests that hydrolysis and splicing reactions are in large part driven by beta5 subunit proteasome known to prioritise cutting after hydrophobic residues. Berkers *et al.* focused on short antigenic peptides binding to HLA-A*02:01, likely biasing their results, which wasn't the case in our dataset [18]. In addition, there was a statistically significant higher frequency of splicing reactions after positively charged residues H and K, which also somewhat brings our results with what Paes *et al.* reported [19].

Firstly, when it comes to SCS-P1 frequencies, there was a strong to moderate statistically significant correlation between the digestions of WT and MUT polypeptides indicating that the catalytic activity of proteasome is quite comparable for the canonical peptides. This wasn't the case for PSP-P1, for which we observed little to no statistically significant correlation between the digestions of WT and MUT substrates. This suggests that regardless of the type of substitution, the presence of mutation considerably changes the splicing patterns of the polypeptides by perhaps changing the utilisation of previously underused/overused sequence motifs. Secondly, we performed a comparison of SCS-P1 and PSP-P1 frequencies between polypeptides grouped by the charge and hydrophobicity. In most cases, those comparisons didn't result in statistically significant differences in distributions. Nevertheless, the median frequency values for some groups of residues were noticeably higher in the digestions of WT or conversely MUT polypeptides. Due to the lack of statistical significance of most of those frequencies and the fact that we used strictly signal intensities to compute the frequencies it's currently difficult to judge if those alterations in frequencies for different groups residues are meaningful. This conundrum shall be resolved by increasing the number of polypeptide digestions in each group of WT-MUT sequence pairs and by computing the precise chemical amounts of all of the peptide products that would allow to perform direct comparisons of SCS-P1 and PSP-P1 not biased by different ionisation properties of different peptides. The issue of precise quantification of the *in vitro* proteasomal digestion products is currently being tackled in our group by Sarah Henze and Dr. Juliane Liepe who are developing a novel method of absolute quantification called Quantification of Peptides using Bayesian approach (QPuB). This method is based on the law of mass conservation in the proteasomal digestions. It allows to estimate of the conversion factors allowing to convert the raw signal intensities of the peptide products into the chemical amounts using Bayesian approaches.

In the future analysis it will also be important to determine the impact of the mutation not just on the utilisation of residues in p1 positions, but also broadly the sequence motifs of both SR1 and SR2 reactants binding in the non-primed and primed binding sites (*e.g.* p2,p3,p4 and p2',p3' and p4' residues) [18, 19, 27]. While p1 residues largely drive the cleavage/splicing reactions, the surrounding residues are also important in determining the efficiency of these reactions as they determine how well the sequence motifs of the polypeptides are accommodated in the non-primed and primed binding sites of the proteasome catalytic chamber. Mishto *et al.* hypothesised that the time the SR1 spends in the bound state largely determines a success of the reaction [27]. This conjecture however requires further experimental verification. Berkers *et al.*, on the other hand posited that a good fit of the substrate in the catalytic chamber of the proteasome would make the nucleophilic attack on the acyl-enzyme intermediate by water molecule more likely, hence leading to a canonical cleavage [18]. On the other hand, less optimal accommodation of the SR1 could disfavour reaction with water molecule and favour the trans-peptidation reaction with SR2. It's also possible that the a suboptimal fit of the SR1 in the non-primed binding site could disrupt the hydrogen-bonding networks necessary for hydrolysis reactions, thus making ligation reactions more likely. It will be interesting to investigate these mechanisms in context of the effect of mutations on the overall proteasomal dynamics. It will also be important to determine if the impact of the mutation on the proteasomal processing of the substrates is driven by specific catalytically active beta subunits. To answer this question it will be prudent to selectively inhibit specific beta subunits which will be followed by the comparison of the digestions of WT and MUT polypeptides. Moreover, for the peptides shown to be differentially produced in the WT and MUT digestions of the polypeptides by specific beta subunits, it would also be prudent to investigate interactions of these peptides in the proteasome catalytic chambers to get an insight into the structural properties of proteasomes facilitating differential efficiency of cleavage/splicing reactions. Finally, it would be important to select the several representative peptides with significantly different kinetic behaviour in the digestions of WT and MUT polypeptides and to track their amount in the proteasome chamber over time. This could be done with the *in silico* approach previously developed in our group based on the bayesian computation to obtain posterior parameter distributions [67].

In general, the differences in cleavage/splicing site usage were observed both close and distant to the position with the mutation, depending on the polypeptide. These effects could be explained by introduction of the strong cleavage/splicing sites via amino acid substitution as well as the changes in sequence motifs of the polypeptides most optimal for the interactions with the substrate binding sites within the proteasome triggered by the mutation - *i.e.* not just p1 and p1' positions but the residues surrounding them. Previously, several groups showed that the differences in catalytic activities of different proteasome isoforms were in part due to differential transport of the substrates along the proteasome towards it's catalytic chambers [66, 67]. This difference in transport of WT and MUT polypeptides could be the factor in the differential processing of WT-MUT substrates pairs by 20S standard proteasome. However this is yet to be determined. Only a small share of peptides that were detected and quantified were sequence identical between the digestions of WT and MUT polypeptides (15% on average). This suggests that regardless of the observed effect of the mutation in different substrate groups there was a substantial alteration of peptide pools produced in the digestions by 20S proteasome. In part these unique peptide pools could be attributed to the peptide products containing the position with the mutation. These tumor-specific somatic mutations may promote the generation of the novel antigenic peptides which both possess better MHCI binding properties compared to their WT counterparts and are capable of eliciting strong and tumor-specific CD8+ T-cell responses. On the other hand, among the peptides that could arise in the digestions of both versions of any given polypeptides, the lack of their detection in one or the other digestion doesn't necessarily mean that they weren't generated but rather that the efficiency of their generation was low. Due to the introduction of a strong cleavage/splicing site as a result of amino acid substitution, the efficiency of generation of such peptides could increase significantly. In some cases, for example in the study by Fidanza M. *et al* of the impact of G/Y mutation of the digestion of pepVIII peptide from GBL specific EGFR receptor, the presence of such somatic mutation, would enhance the generation of antigenic peptides of interest [29].

Finally, we verified the practical utility of our *in vitro* approach for the identification of putative epitope candidates for the adoptive T-cell therapies. We demonstrated that despite the reduction in the overall number of epitope candidates that were quantified compared to the 20h digestions of the same polypeptides, we could nevertheless identify a number of potential epitope candidates. A large fraction of those candidates were cis-spliced peptides in most substrates. This is in line with the results previously obtained in our group [13, 14]. The fact that the peptides that were detected in 20h digestions but not kinetics, it doesn't necessarily mean that those peptides are not generated but rather that their abundances are so low that they are not picked up during MS measurements. A good example of it is the epitope carrying timor-specific mutation that was identified in 20h digestions but not quantified in the digestion kientics. This epitope is derived from KRAS G12V [72].

It's important to stress that a given polypeptide is physiologically more likely to be digested by the proteasome for 4 hours as opposed to 20 hours *in vivo*. That is because, it's improbable that any single substrate will occupy the proteasome for 20 hours while there are many other proteins/polypeptides present in the cytosol that have to processed by the proteasome. The main reason, we use 20 hour digestions in our epitope identification pipeline, is because we want to assess the full scope of peptide products that could be produced by the proteasome. On the other hand, just because, the peptide is produced by the proteasome and detected, doesn't mean that it will be able to actually pass all of the downstream steps in the APP and be presented on the cell surface [17]. Moreover, our analysis is based on the *in silico* measurements of the binding affinity of the peptides to HLA molecules. We will have to determine the binding affinities of the peptides experimentally with the binding assays to understand which of these peptides are in fact binders. Importantly, the hight afiinity of the peptide for an MHCI molecule doesn't guarantee it will trigger an immunogenic response. It's known that the likelihood of a peptide activating CD8+ T-cells depends on the multitude of factors, binding affinity to HLA despite being important, being just one of them. In addition to that, the abundance of a given peptide on a cell surface plays a big role, as well as the presence of TCRs that could efficiently interact with the presented peptide [52].

The other important consideration is that in our analysis we counted N/C-terminally extended precursors of the epitopes of interest. We chose to do so because it's known that the proteasome can additionally trim the C-terminus of the generated precursor peptides while cytosolic and ER resident amino-peptidases can trim the precursor at its N-terminus thus producing the actual epitope [64]. The inclusion of N/C-terminally extended precursors increased our estimates. At this stage however, we don't know if any of the identified and quantified precursors can in fact be trimmed up to the desired epitopes, or if the precursors themselves could be presented by MHCI molecule and elicit the immune response. The elusions of the peptides from MHCI molecules of the target cells, followed by MS/MS will allow us to determine which of the peptides can be derived from the identified and quantified peptide products.

Of note however, is that in some cases, such as KRAS G12R there was a large share of quantified non-spliced epitope candidates. This could be explained by the amino acid composition of the peptide and the influence of the mutation on the utilisation of certain p1 cleavage/splicing residues and sequence motifs favouring generation of peptides that could be used to elicit immune responses. It is known that non-spliced peptides are more abundant in the *in vitro* proteasomal digests [72]. That means that despite a significantly lower number of the potential non-spliced HLA binding peptides, containing the tumor-associated mutation compared to cis-spliced peptides, those peptides that are produced, could be generated more efficiently that cis-spliced peptides. This translates into a large number of theoretically possible non-spliced peptides being produced relative to the small number of all possible non-spliced HLA-binders as compared to the small numbers of produced cis-spliced peptides compared to the large number of theoretically possible cis-spliced peptide binders. The reason we didn't consider trans-spliced peptides is due to the fact that while their generation was demonstrated *in vitro* and they were suggested to constitute a large fraction of the HLAI immunopeptidome by some groups, it's still being debated whether trans-spliced peptides are produced in sufficient quantities *in vivo* to be relevant for the antigen presentation [12, 70].

The amino-acid composition of any given polypeptide will favour or disfavour the generation of the potential MHCI binders. For example, MPL W515L derived substrate has a very high number of identified and quantified non-spliced and cis-spliced peptides compared to all other substrates. This is likely because MPL W515L has a long stretch of hydrophobic residues in its sequence, mainly Leucins. It's known that hydrophobic amino acids are the preferred anchor sites for some HLA molecules, for example HLA-A*02:01 [18]. This means that many peptide products derived from the digestion of MPL W515L could be good HLA binders due to their enrichment in hydrophobic residues. Other factors, such as presence of position with the mutation could affect not just the recognition of the peptide by TCRs but also it's binding to HLA molecules. To understand the impact of the mutation on the interaction of the peptides with both HLAs and TCRs, it will be important to generate the crystal structures of HLA-peptide-TCR complexes to decipher the molecular properties of their interaction. The other important consideration for the numbers of the identified peptides is the quality of proteasome preparation and the purity of the synthetic polypeptide (Dr. Michele Mishto, personal communication).

We presently don't know the precise concentrations of the detected and quantified peptides. This will require further investigation, including the precise quantification of peptides' concentration, using our currently developed QPuB approach. Knowing the chemical amount of the peptide of interest will be important for determining how likely it is to be presented on the cell surface. In addition, we observed that most of the peptides or their precursor that are considered the most promising epitope candidates could not be found in human proteome either as such or with I/L redundancies which will certainly be beneficial for the downstream verification of the identified and quantified epitope candidates. Filtering of such sequence matching peptides is very important in any downstream experiments as such peptides could trigger unwanted autoimmune responses. The big exception in this analysis was BRAF V600E derived polypeptides, in which a very large fraction the quantified epitope candidates were found in the human proteome. This could be explained by the amino acid composition of that particular polypeptide that increases the probability of identical stretches of sequence found in the human proteome.

The next steps in the verification of such peptides for the therapeutic use will be binding assays to HLA molecules to confirm that they are indeed good binders, ERAP assays to demonstrate that the precursors of the epitopes of interest could be trimmed for binding to the HLA molecules and the T-cell immunological assays using PBMCs to confirm that the TCRs can indeed recognise such presented peptides and elicit immunogenic response (*e.g.* IFNgamma and TFNalpha secretion). The final step will be the experiments using cell culture or mice genetically engineered to express HLAs of interest to confirm *in cellulo* and *in vivo* that the peptides of interest are presented and are capable of inducing strong CD8+ T-cell responses [71].

Chapter 7

Conclusions

Ever since the discovery of PCPS in 2004, its true purpose and function has been elusive. That was the case until the last several years when giant strides in mass spectrometry and the tools for the MS data analysis revealed that spliced peptides are produced by the proteasome significantly more frequently than originally though and that PCPS is tightly controlled by the proteasome, demonstrating that it's not just a random byproduct of the proteasome's primary catalytic activities [2, 13, 14, 15, 27]. Remarkably, the patterns of splicing are distinct from the canonical peptide bong cleavage as the sites of most frequent cleavages and splicing reactions often don't overlap [27]. The potential importance of proteasome generated spliced peptides became clear, once it was demonstrated by several groups via MS of tpeptides eluted from MHCI molecules of a variety of cell lines, that spliced peptides constitute a substantial fraction of MHCI presented peptides and thus could be critical for a proper immune responses [13, 14, 15]. Moreover, on several occasions spliced peptides were shown to trigger specific immune responses by distinct populations of CD8+ T-cells that were non-cross-reactive with non-spliced peptides [12, 20, 22]. Despite this impressive progress in understanding of the nature of PCPS and its biological function, many questions are still remaining that concern the biochemical mechanisms of PCPS and PCPS's role in various aspects of adaptive immune response.

In this thesis, I investigated several aspects of the PCPS that required further clarification. I focused on three topics: the contribution of spliced peptides into immune evasion by pathogens, role of spliced peptides in the viral-triggered autoimmunity and impact of single amino acid substitutions on a quantitative profile of the proteasomal digestions of polypeptides, particularly in regards to PCPS.

Despite the fact that spliced peptides are now thought to significantly enlarge the antigenic repertoire of cells that could be targeted by CD8+ T-lymphocytes, another more unsettling role of spliced peptides was suggested [24]. It was hypothesised that the prominence of the spliced peptides in the MHCI immunopeptidome, could increase the frequency of non-self peptides that are indistinguishable from self from adaptive immune system point of view. This is due to the fact that one of the potential immune evasion strategies of intracellular pathogens is mimicking of the self-antigens via a frequent presentation of short antigenic peptides identical or very similar in sequence to self derived peptides. The previous estimates of the extent of such molecular mimicry conducted for non-spliced peptides suggested that only a small share of non-spliced non-self peptides could interfere with immune response [24]. Those estimates however increased considerably when T-cell cross-reactivity was taken into account. When we incorporated cis-spliced peptides into the estimate of the frequency of viral-human zwitter peptides we found out that even when only completely matching peptides were evaluated, a fraction of non-self cis-spliced peptides that would be identical to self was orders of magnitude larger than for non-spliced peptides. Moreover, when we considered TCR cross-reactivity, these estimates increased dramatically and suggested that the majority of non-self cis spliced peptides would escape immune surveillance. These estimates however assumed that all of the theoretically possible cis-spliced peptides that could be generated, are in fact produced by proteasome and that all of the putative HLA-A*02:01 binders are presented on the cell surface. This is however not the case, as it was shown that in the *in vitro* digestions of a variety of synthetic polypeptides only a minor fraction (less than 1%) of all theoretically possible spliced peptides are produced at the MS detectable level [17]. Moreover, those peptides that are generated have to pass through extensive selection steps in APP to be presented which further decreases the number of immunologically relevant peptides [13, 14]. When we took this information into account and re-examined the theoretical frequency of viral-human zwitter peptides, we found that the majority of such zwitter peptides wouldn't be presented. Our estimates of the numbers of viral-human zwitter peptides in fact dropped from tens and hundreds of thousands per individual virus to single digits in most cases. While the inclusion of TCR cross-reactivity increased the expected frequency of the presented viral-human zwitter peptides, it was still less than 1%. This suggests that at least based on our estimates, cis-spliced peptides are not expected to significantly impinge upon the repertoire of virus-specific CD8+ T-cells. Our analysis was limited to peptide of length 9, while the epitopes as long as 14 amino acids could be presented by MHCI molecules and it will be interesting to investigate wether such longer peptides could assist in the immune evasion. Additionally, it will be important to better understand the rules of the immunogenicity of the MHCI presented peptides as this will aid indiscrimination between immunogenic and non-immunogenic epitopes as well as between zwitter and non-zwitter peptides. TCR cross-reactivity is a complex topic due to the fact that the rules that determine it aren't clear yet and there are examples of very similar peptides producing non-cross reactive responses and vice versa the peptides having little similarity triggering one population of CD8+ T-cells [131, 133, 137, 140]. Finally, a better understanding of the rules governing PCPS and conclusive estimations of the frequencies of MHCI presented spliced peptides will be necessary to fully elucidate the role of zwitter peptides in the immune evasion.

Despite this, even those lone zwitter peptides could pose a different challenge for immune system. The recognition of these MHCI presented peptides by self-reactive CD8+ T-cells could trigger strong autoimmune responses. While the negative selection of T-lymphocytes in the thymus is stringent, it isn't perfect. As a result, some self-reactive CD8+ T-cells can make their way into periphery and coupled to the defects in the peripheral tolerance mechanisms, this could result in the autoimmunity. It has long been speculated that viral infections could be triggers of some autoimmune disease such as Multiple Sclerosis and T1D and in some cases cancer [38, 40, 41]. This is due to the fact that upon viral infection, some of the self-reactive zwitter peptides could be presented at a high rate, which would in turn lead to the increased stimulation of self-reactive CD8+ T-cells. For example, an association was found between T1D and infections with some enteroviruses and EBV in some T1D patients [41]. Thus, we set out to determine if there were possibilities for the emergence of viral-human zwitter cis-spliced peptides that could be derived from both viral proteins and T1D associated pancreatic antigens. These zwitter cis-spliced peptides could then be presented by MHCI molecules and recognised by TCRs of pancreas infiltrating self-reactive CD8+ T-cells. We implemented a set of stringent selection criteria based on whether the peptides could be derived from the highly expressed T1D associated antigens and be MHCI binders. As a result of this analysis, no zwitter non-spliced peptides passed the selection steps but by contrast we managed to zero in on a number of cis-spliced peptides that could be relevant in context of the viral induction of T1D. Nevertheless, we are still yet to confirm whether any of these potential candidates are produced at the detectable level and are presented on the cell surface. Moreover, we will have to find out whether there exist subsets of self-reactive CD8+ T-cells in T1D patients that would respond to our candidates. The picture is also obfuscated by the T-cell receptor cross-reactivity which would likely increase the number of the relevant candidates.

Finally, we leveraged data obtained from a variety of polypeptides derived from tumor-associated antigens and carrying tumor-specific mutations to understand the impact of such mutations on a proteasomal processing of polypeptides. Understanding what such changes might entail on a systematic level is not just crucial for gaining a deeper insight into the mechanisms of PCPS but are also important in practical applications, particularly for predicting which potential epitopes carrying tumor-specific mutations would be produced by the proteasome [28, 29]. The isolation of T-lymphocytes specific for the epitopes that could only be produced from the tumor-associated antigens that could then be administered to the patients has been suggested to be one of the effective strategies for tumor-specific immunotherapies [12, 72]. This is because such tumor-specific lymphocytes would ensure targeted tumor destruction without the damage to the periphery. Several groups previously demonstrated that introduction of such single substitutions drastically changed cleavage and splicing patterns of the polypeptides' digestion and led to the increased generation of the peptides that could be used for the immuno-therapies [28, 29]. In our analysis we primarily focused on a quantitative impact of the mutations on the proteasomal dynamics by comparing the digestion profiles and the kinetics of the peptide products derived from the digestions of WT and analogous MUT substrates. We found that the effect of the mutations was highly variable and was dependent on an amino acid composition of a given polypeptide and the type of the amino acid substitution that was introduced into the sequence. We primarily focused on comparisons of the signal intensities of sequence identical peptides found in both the digestions of the WT and MUT polypeptides, due to the unpredictable nature of the ionisation properties of the peptide products differing by even one substitution. We also discovered, that in the case of the introduction of positively charged amino acid residue there was an increase of the generation of the subset of peptide products in the digestions of MUT polypeptides, particularly reverse-cis spliced peptide. On the other hand, the introduction of negatively charged residue resulted in the decreased generation of multiple peptide types in the digestion of the MUT polypeptides. This could mean that the introduction of the amino acid residues from a particular category led to the increased use of certain sequence motifs in the case of positively charged amino acids and vice versa in case of negatively charged residues. These effects were observed for the peptides, which P1 residues were located both in close proximity or distant to the mutation. The other characteristics of the amino acid residues such as their size and the aforementioned hydrophobicity likely also play an important role in determining sequence specificities, however to fully appreciate the impact of these characteristics, we would require digestions of additional polypeptides to be able to split the substrates into groups with similar sizes of the amino acids and similar hydrophobicity scores. In addition, the analysis was limited by us using only the kinetics of the sequence identical peptides not containing the mutation which constitute a small share of all of the generated peptide products. The other interesting observation that we made, was that the frequencies of peptide splicing were affected more strongly by the mutations compared to frequencies of cleavage, which might suggest that the effect of single mutations may be more profound in determining the proteasome splicing sequence preferences compared to cleavage. In addition, we investigated the overall cleavage and splicing patterns after P1 residues across all of the polypeptides to compare them with observations previously made by us and others [17, 18, 19]. While we observed that frequencies of splicing after some residues, were similarly high such as small and non-polar residues to what was reported previously, there weren't clear preferences for splicing after others such as hydrophobic and negatively charged residues. This could be explained by different nature of the data-set that we were using compared to the previously published data as well as the fact that we relied on signal intensity values in our analysis that can't be used as a reliable indicated of the actual peptide concentration. These limitations will have to be rectified in the future analysis by including the larger number of polypeptide digestions into our data-set and employing strict quantitive approaches for determining the precise chemical amounts of all of the peptide products, such as our in-house QPuB approach. This will enable us to perform comprehensive comparisons of the generation dynamics of all of the peptide products including those that contain the mutated residues. Finally, we utilised our *in vitro* digestions to identify and quantify potential epitope candidates carrying tumor specific mutations that could be used in the anti-cancer immunotherapies. Despite the fact that many of the theoretically possible MHCI binders weren't detected in the digestions, we still managed to identify a number of both the minimal epitopes and the precursors that could be used for a further verification. Importantly, the large fraction of those candidates were cis-spliced peptides which demonstrates that peptide splicing diversifies an antigenic repertoire of cells and could promote the increased presentation rate of the peptides carrying the mutations in line with what was reported previously [13, 14, 15]. Despite this, we would still have to precisely determine the amounts of the generated peptides and verify that they could pass all of the APP steps, be presented and trigger the specific non-cross-reactive immune responses. This will be demonstrated in the further *in vitro*, *in cellulo* and *in vivo* experiments.

Overall, the research described in this thesis provides new insight into the role of the spliced peptides in the different aspects of adaptive immune response. We gained additional insight on the role proteasome-generated spliced peptides in self/non-self discrimination by immune system in context of both viral immune evasion and autoimmunity. Crucially, in the analysis we considered a type of peptides that is located in the boundary between non-self pathogenic and immunological self peptides - the antigenic peptides derived from the tumor associated antigens. On the one hand, these unique peptides are derived from self antigens, but the presence of the mutations that lead to the abnormal cellular growth simultaneously makes them non-self peptides. We showed that the introduction of such mutations drastically change cleavage and splicing profiles and could potentially lead to the production of antigenic peptides that could be utilised in a targeted immunotherapies against cancer.

Supplementary material

Table S1: Viruses and viral strains

Virus species	Strain
Adeno associated virus 2	isolate Srivastava/1982
Aichi virus 1	isolate A846/88
Australian bat lyssavirus	isolate Bat/AUS/1996
Banna virus strain Indonesia JKT 6423	strain Indonesia/JKT-6423/1980
Barmah forest virus BFV	strain BH2193
BK polyomavirus	strain Dunlop
Bunyamwera virus	-
Bunyavirus La Crosse US L78 1978	isolate Human/United States/L78/1978
Cercopithecine alphaherpesvirus 2	strain B264
Chandipuravirus	strain I653514
Chikungunya virus	strain S27-African prototype
Cowpox virus (CPV)	strain Brighton Red
Coxsakievirus A16	strain G-10
Crimean-Congo hemorrhagic fever virus	strain Nigeria/IbAr10200/1970
Dengue virus type 1	strain Nauru/West Pac/1974
Dhori virus	strain Indian/1313/61
Dugbe virus	isolate ArD44313
Duvenhage virus (DUVV)	-
Eastern equine encephlitis virus	strain ssp. North American variant
Encephalomyocarditis_virus_strain_Rueckert	strain Ruckert
Epstein Barr virus	strain B95-8 HHV-4
European bat lyssavirus	strain Bat/Germany/RV9/1968
GB virus C/Hepatitis G virus	Isolate PNF2161
Hantaan virus	strain 76-118
Hendra virus	isolate Horse/Autralia/Hendra/1994
Hepatitis B virus genotype C	isolate Human/Japan/Okamoto/-
Hepatitis_E virus genotype 1	isolate Human/China/HeBei/1987
Hepatitis C virus genotype 1a	isolate H
Hepatitis delta virus I	isolate D380
HIV 1 group M subtype B	isolate HXB2
Horsepox virus	strain MNR-76
Human adenovirus C serotype 2	strain H2ts125
Human astrovirus 1	Isolate Newcastle
Human betaherpesvirus 7	strain RK
Human coronavirus 229E	strain 229E
Human cytomegalovirus strain AD169	strain AD169
Human enterovirus 70	strain J670/71
Human hepatitis A virus genotype IB	isolate HM175
Human herpesvirus 1	strain 17
Human papillomavirus type 1	-
Human parainfluenza virus 1	strain Washington/1964
Human parechovirus 2	strain Williamson
Human parvovirus B19	strain HV
Human respiratory syncytial virus B	strain B1

Human rhinovirus A serotype 89	strain 41467-Gallo
Human SARS coronavirus	Isolate Tor2
Human SARS CoV 2	Wuhan-Hu-1
Human spumaretrovirus SFVcpz	-
Human T cell leukemia virus 1	isolate Caribbea HS-35 subtype A
Influenza A virus	strain A/Puerto Rico/8/1934 H1N1
Isfahan virus ISFV	-
Japanese encephalitis virus	strain Jaoars982
JC polyomavirus	strain Mad1
Junin mammarenvirus JUNV	strain XJ13
KI polyomavirus	isolate Stockholm 60
Kunjin virus	strain MRM61C
Lagos bat virus	-
Lake victoria marburgvirus	strain Musoke-80
Langat virus	strain TP21
Lassa virus strain Mouse Sierra Leone	strain Mouse/Sierra Leone/Josiah/1976
Lordsdale virus	strain GII/Human/United Kingdom/Lordsdale/1993
Louping ill virus	strain 369/T2
Lymphocytic choriomeningitis virus	strain Armstrong
Machupo virus	strain Carvallo
Mayaro virus strain Brazil	strain Brazil
Measles_virus_strain_Ichinose_B95a	strain Ichinose-B95a
Merkel cell polyomarvirus	strain R17b
Mokola virus	-
Molluscum contagiosum virus subtype 1	-
Monkeypox virus	strain Zaire-96-I-16
Mums virus	strain Miyahara vaccine
Murray_valley_encephalitis_virus_strain_MVE_1_51	strain MVE-1-51
Nipah virus	-
Norwalk virus	strain GI/Human/United States/Norwalk/1968
Onyong nyong virus	strain Gulu
Orf virus	strain Goat/Texas/SA00/2000
Oropouche virus	strain BeAn19991
Parainfluenza virus 5	strain W3
Pichinde mammarenavirus	strain AN3739
Poliovirus type 1	strain Mahoney
Puumala virus	strain Sotkamo/V-2969/81
Rabies virus	strain Pasteur vaccins / PV
Rift valley fever virus	strain ZH-548 M12
Rosavirus A2	isolate Human/Gambia/GA7403/2008
Ross river virus	strain NB5092
Rotavirus C	isolate RVC/Human/United Kingdom/Bristol/1989
Rubella virus	strain Therien
Sagiyama virus	-
Salivirus A	strain NG-J1
Sapovirus	strain GII/Human/Japan/Sakai C12/2001
Seoul virus	strain 80-39
Simian foamy virus	isolate chimpanzee
Sinbis virus	strain HRsp
Southampton virus	strain GI/Human/United Kingdom/Southampton/1991
St Louis encephalitis virus	strain Kern217
Tick borne powassan virus strain LB	strain LB
Torque teno virus	strain VT416
Toscana virus	strain Toscana
Vaccinia virus	strain Western Reserve
Varicella zoster virus	strain Dumas

Variola_virus_Smallpox	isolate Human/India/Ind3/1967
Vesicular stomatitis Indiana virus	strain San Juan
Western equine encephalitis virus	strain 71V-1658
West Nile virus	strain 956
WU polyomavirus	strain B0
Yaba monkey tumor virus	strain VR587
Yellow fever virus	strain 17D vaccine
Zaire ebolavirus	strain Mayinga-76
Zika virus	strain Mr 766

Table S2: Statistical test values (Kolmogorov-Smirnov test (difference between groups of zwitter peptides))

Figure	Group A	Group B	p-value
2A	Zwitter non-spliced peptides	Zwitter cis-spliced peptides	2.2e-16
2A	Zwitter non-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
2A	Zwitter cis-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
2B	Zwitter non-spliced peptides	Zwitter cis-spliced peptides	2.2e-16
2B	Zwitter non-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
2B	Zwitter cis-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
4B	Zwitter non-spliced peptides	Zwitter cis-spliced peptides	2.2e-16
4B	Zwitter non-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
4B	Zwitter cis-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
6A	Zwitter non-spliced peptides	Zwitter cis-spliced peptides	2.2e-16
6A	Zwitter non-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
6A	Zwitter cis-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
6B	Zwitter non-spliced peptides	Zwitter cis spliced peptides	2.2e-16
6B	Zwitter non-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
6B	Zwitter cis-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1

7A	Zwitter non-spliced peptides	Zwitter cis spliced peptides	1.992e-11
7A	Zwitter non-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
7A	Zwitter cis-spliced peptides	Zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.22e-11

Table S3: Statistical test values (Correlation coefficients between virus length and number of different types of zwitter peptides)

Type of zwitter peptide	Pearson Correlation Coefficient (C)	p-value
zwitter peptides	0.904	2.2e16
HLA-A*02:01-restricted zwitter peptides	0.943	2.2e-16
HLA-A*02:01-restricted zwitter peptides based on RNA-based proteome database	0.955	2.2e-16
HLA-A*02:01-restricted zwitter peptides assuming $\sim$15% frequency of cis-spliced peptides in HLA-I immunopeptidomes	0.172	7,40e-02

Table S4: List of crystal structures of 9mer-HLA-A*02:01 complexes with TCR analyzed for the implementation of the degenerate recognition model

PDB ID	Peptide sequence	Peptide name	Protein of origin (UNIPROT ID)	Source organism
3GSN	NLVPMVATV	pp65 495-503	A0A0K1W3P3	Human cytomegalovirus (HCMV)
1AO7	LLFGYPVYV	Tax peptide	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
5HHM	GILGLVFTL	Synthetic construct M1-F5L	-	-
5HHO	GILEFVFTL	Synthetic construct M1-G4E	-	-
1BD2	LLFGYPVYV	Tax peptide	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
1LP9	ALWGFFPVL	Self-peptide P1049	Q9NPA0	Human
2VLR	GILGFVFTL	Flu matrix peptide 58-66	P03485	Influenza A virus
3H9S	MLWGYLQYV	Tet1p peptide	P38110	Saccharomyces cerevisiae
3QEQ	AAGIGILTV	Melanoma antigen peptide recognized by T-cells 1 27-35	Q16655	Human

3PWP	LGYGFVNYI	HuD peptide	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
3QFJ	LLFGFPVYV	Modified Tax(Y5F) peptide	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
3QDJ	AAGIGILTV	Melanoma antigen peptide recognized by T-cells 1 27-35	Q16655	Human
4FTV	LLFGYPVYV	Tax peptide	P0C213	Human T-cell leukemia virus type 1 (HTLV-1)
5TEZ	GILGFVFTL	Flu matrix peptide 58-66	P03485	Influenza A virus
3D3V	LLFGFPVYV	Modified Tax (Y5(3,4-difluoro)F) peptide	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
5NME	SLYNTVATL	Gag protein peptide	P04591	Human Immunodeficiency virus 1 (HIV-1)
5ISZ	GILGFVFTL	Flu matrix peptide 58-66	P03485	Influenza A virus
2GJ6	LLFGKPVYV	Modified Tax (Y5K-IBA) peptide, chain C	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
1QRN	LLFGYAVYV	Modified Tax peptide P6A	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
5NMF	SLYNTIATL	Gag protein peptide	P04591	Human immunodeficiency virus 1 (HIV-1)
4EUP	ALGIGILTV	Melanoma antigen recognized by T-cells 1	Q16655	Human
2UWE	ALWGFFPVL	Self-peptide P1049	Q9NPA0	Human
2J8U	ALWGFFPVL	Self-peptide P1049	Q9NPA0	Human
5MEN	ILAKFLHWL	Self-peptide from Telomerase reverse transcriptase	O14746	Human
1QSF	LLFGYPVAV	Modified Tax peptide Y8A	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
5JHD	GILGFVFTL	Flu matrix peptide 58-66	P03485	Influenza A virus
2P5E	SLLMWITQC	Cancer/testis antigen 1B peptide	P78358	Human
2BNR	SLLMWITQC	Cancer/testis antigen 1B peptide	P78358	Human
2F53	SLLMWITQC	Cancer/testis antigen 1B peptide	P78358	Human
2BNQ	SLLMWITQV	Cancer/testis antigen 1B peptide	P78358	Human
2F54	SLLMWITQC	Cancer/testis antigen 1B peptide	P78358	Human
1QSE	LLFGYPRYV	Modified Tax peptide V7R	P14079	Human T-cell leukemia virus type 1 (HTLV-1)

3O4L	GLCTLVAML	BSLF2/BMLF1 protein derived peptide	Q3KSU1	Human gammaherpesvirus 4
5D2N	NLVPMVATV	pp65 fragment 495-503	P18139	Human cytomegalovirus (HCMV)
4MNQ	ILAKFLHWL	Self-peptide from Telomerase reverse transcriptase	O14746	Human
2VLJ	GILGFVFTL	Flu matrix peptide 58-66	-	Unidentified influenza virus
5E6I	GILGFVFTL	Flu matrix peptide 58-66	P03485	Influenza A virus
6D78	AAGIGILTV	Melanoma antigen peptide recognized by T-cells 1 27-35	Q16655	Human
3D39	LLFGFPVYV	Modified Tax (Y5(4fluoro)F) peptide	P14079	Human T-cell leukemia virus type 1 (HTLV-1)
2JCC	ALWGFFPVL	self-peptide P1049	Q9NPA0	Human
6EQA	AAGIGILTV	Melanoma antigen peptide recognized by T-cells 1 27-35	Q16655	Human
5EU6	YLEPGPVTV	Synthetic self-peptide	P40967	Human
6RSY	RMFPNAPYL	Synthetic self-peptide	P19544	Human
6RPA	SLLMWITQV	Heteroclitic NY-ESO-1 157-165 peptide	P78358	Human
6Q3S	SLLMWITQV	Heteroclitic NY-ESO-1 157-165 peptide	P78358	Human
6EQB	AAGIGILTV	Melanoma antigen peptide recognized by T-cells 1 27-35	Q16655	Human
5NMG	SLYFNTIAVL	Gag protein peptide	P04591	Human immunodeficiency virus 1 (HIV-1)
6R2L	SLSKILDTV	Synthetic self-peptide	Q9BXX3	Human

Table S5: Statistical test values (Kolmogorov-Smirnov test (difference between groups of degenerate zwitter peptides))

Figure	Group A	Group B	p-value
12	Degenerate zwitter non-spliced peptides	degenerate zwitter cis-spliced peptides	2.2e-16
12	Degenerate zwitter non-spliced peptides	degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
12	Degenerate zwitter cis-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
13A	Degenerate zwitter non-spliced peptides	Degenerate zwitter cis-spliced peptides	2.2e-16

13A	Degenerate zwitter non-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
13A	Degenerate zwitter cis-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
13B	Degenerate zwitter non-spliced peptides	Degenerate zwitter cis-spliced peptides	2.2e-16
13B	Degenerate zwitter non-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
13B	Degenerate zwitter cis-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	1
15	Degenerate zwitter non-spliced peptides	Degenerate zwitter cis-spliced peptides	2.2e-16
15	Degenerate zwitter non-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16
15	Degenerate zwitter cis-spliced peptides	Degenerate zwitter non-spliced and cis-spliced peptides (combined zwitter peptides)	2.2e-16

Table S6: Statistical test values (Correlation coefficients between virus length and number of different types of degenerate zwitter peptides)

Type of zwitter peptide	Pearson Correlation Coefficient (C)	p-value
HLA-A*02:01-restricted degenerate zwitter peptides	0.995	2.2e-16
HLA-A*02:01-restricted degenerate zwitter peptides based on RNA-based proteome database	0.995	2.2e-16
HLA-A*02:01-restricted degenerate zwitter peptides assuming ∼15% frequency of cis- spliced peptides in HLA-I immunopeptidomes	0.942	2.2e-16

Table S7: List of virus strains included in the study

Virus	Acronym	Strain
Coxsakievirus B1	CVB1	Japan
Coxsakievirus B4	CVB4	E2
Epstein Barr virus	EBV	B95 8 HHV 4
Human cytomegalovirus	HCMV	AD169
Human herpesvirus 6A	HHV-6A	6A
Human herpesvirus 6B	HHV-6B	6B
Human parechovirus 2	HPeV2	Williamson
Rotavirus C	RVC	Isolate RVC

Table S8: Peptide-HLA-I binding affinity, reported as IC50 and used as cut-offs.

HLA complex	IC50 cut-off (nM)
HLA-A*02:01	500.0
HLA-A*01:01	1486.5
HLA-A*03:01	482.3
HLA-A*11:01	132.2
HLA-A*23:01	263.7
HLA-A*24:02	519.9
HLA-B*07:02	239.3
HLA-B*08:01	1687.7
HLA-B*15:01	357.7
HLA-B*35:01	153.9
HLA-B*39:06	5574.7
HLA-B*40:01	171.0
HLA-B*44:02	550.0
HLA-B*44:03	650.2

Table S9: Statistical test values (Kolmogorov-Smirnov test (difference between lengths of different peptide types))

Figure	group A	group B	p-value
37A	non-spliced	normal cis-spliced	4e-15
37A	non-spliced	reverse cis-spliced	1.47e-05
37A	non-spliced	trans spliced	2.2e-16
37A	normal cis-spliced	reverse cis-spliced	5.903e-13
37A	normal cis-spliced	trans spliced	2.2e-16
37A	reverse cis-spliced	trans spliced	2.2e-16
37C	non-spliced	normal cis-spliced	4.57-e06
37C	non-spliced	reverse cis-spliced	0.012
37C	non-spliced	trans spliced	2.2e-16
37C	normal cis-spliced	reverse cis-spliced	1.5688e-07
37C	normal cis-spliced	trans spliced	2.2e-16
37C	reverse cis-spliced	trans spliced	2.2e-16
37D	non-spliced	normal cis-spliced	7.56e-16
37D	non-spliced	reverse cis-spliced	0.0008
37D	non-spliced	trans spliced	2.2e-16
37D	normal cis-spliced	reverse cis-spliced	1.57e-07
37D	normal cis-spliced	trans spliced	2.2e-16
37D	reverse cis-spliced	trans spliced	2.2e-16

Table S10: Statistical test values (Kolmogorov-Smirnov test (difference between lengths of WT and MUT derived different peptide types))

Figure	type of peptide	p-value
37B	non-spliced	0.02
37B	normal cis spliced	0.0067
37B	reverse-cis spliced	1.678-e06
37B	trans spliced	0.56

Table S11: Statistical test values (Pearson correlation coefficients and p.values (length of polypeptides vs number of peptide products of different types)

Figure	type of peptide	correlation coefficient	p-value
38	non-spliced	0.85	5.1e-05
38	normal cis-spliced	0.83	0.00012
38	reverse cis-spliced	0.27	0.34
38	trans spliced	0.51	0.051

Table S12: Statistical test values (Kolmogorov-Smirnov test (difference between lengths of SR1 and SR2 of different peptide types)

Figure	type of peptide	p-value
39A	normal cis-spliced	2.49e-12
39A	reverse cis-spliced	9.6e-07
39A	trans spliced	1.33e-05
39B	normal cis-spliced	5.18e-14
39B	reverse cis-spliced	0.047
39B	trans spliced	0.48
39C	normal cis-spliced	0.0088
39C	reverse cis-spliced	2.12e-10
39C	trans spliced	2.35e-06

Table S13: Statistical test values (Kolmogorov-Smirnov test (difference between lengths of different peptide types in WT-MUT data-set and the DB published by Specht et al.)

Figure	type of peptide	p-value
40A	non-spliced	2.2e-16
40A	normal cis-spliced	2.2e-16
40A	reverse cis-spliced	2.2e-16
40A	trans spliced	2.2e-16

Table S14: Statistical test values (Kolmogorov-Smirnov test (difference between SR1 and SR2 lengths of different peptide types in WT-MUT data-set and the DB published by Specht et al.)

Figure	type of peptide	SR	p-value
40B	normal cis-spliced	SR1	2.2e-16
40B	reverse cis-spliced	SR1	8.4e-07
40B	trans spliced	SR1	1.8e-11
40B	normal cis-spliced	SR2	0.08
40B	reverse cis-spliced	SR2	2.5e-10
40B	trans spliced	SR2	2.3e-10

Table S15: Statistical test values (Kolmogorov-Smirnov test (difference between lengths of different peptide types in WT-MUT data-set and the random DB)

Figure	type of peptide	p-value
41A	non-spliced	0.0061
41A	normal cis-spliced	4.51e-10
41A	reverse cis-spliced	9.3e-8
41A	trans spliced	2.2e-16

Table S16: Statistical test values (Kolmogorov-Smirnov test (difference between SR1 and SR2 lengths of different peptide types in WT-MUT data-set and the random DB)

Figure	type of peptide	SR	p-value
41B	normal cis-spliced	SR1	1.17e-12
41B	reverse cis-spliced	SR1	0.07
41B	trans spliced	SR1	2.2e-16
41B	normal cis-spliced	SR2	0.003
41B	reverse cis-spliced	SR2	6.3e-10
41B	trans spliced	SR2	2.2e-16

Table S17: Statistical test values (Kolmogorov-Smirnov test (difference between intensity distributions of all non-spliced and spliced peptides)

Figure	timepoint(hours)	p-value
43A	1h	2.2e-16
43A	2h	2.2e-16
43A	3h	2.2e-16
43A	4h	2.2e-16

Table S18: Statistical test values (Kolmogorov-Smirnov test (difference between intensity distributions of different types of spliced peptides)

Figure	timepoint	group A	group B	p-value
43B	1h	normal cis-spliced	reverse cis-spliced	3.2e-06
43B	2h	normal cis-spliced	reverse cis-spliced	0.009
43B	3h	normal cis-spliced	reverse cis-spliced	0.0375
43B	4h	normal cis-spliced	reverse cis-spliced	9.91e-13
43B	1h	normal cis-spliced	trans spliced	8.76e-08
43B	2h	normal cis-spliced	trans spliced	6.2e-06
43B	3h	normal cis-spliced	trans spliced	0.00012
43B	4h	normal cis-spliced	trans spliced	0.0002
43B	1h	reverse cis-spliced	trans spliced	0.011
43B	2h	reverse cis-spliced	trans spliced	0.076
43B	3h	reverse cis-spliced	trans spliced	0.00016
43B	4h	reverse cis-spliced	trans spliced	1.98e-12

Table S19: Statistical test values (Pearson correlation coefficients and p.values (number of peptide products in the digestions of WT and MUT polypeptides)

Figure	group	contrast	correlation coefficient	p-value
45A	charge	no change	0.85	0.0000431
45A	charge	neutral to positive	0.72	0.00821
45A	charge	neutral to negative	0.785	0.0025
45B	hydrophobicity	no change	0.652	0.006
45B	hydrophobicity	hydrophobic to hydrophilic	0.78	4.5e-05

Table S20: Statistical test values (Wilcoxon test (difference of the mean of log10 ratio distribution from 0)

Figure	group	contrast	timepoint	peptide type	p-value
46A	charge	no change	1h	all peptides	0.25
46A	charge	no change	2h	all peptides	0.68
46A	charge	no change	3h	all peptides	0.49
46A	charge	no change	4h	all peptides	0.56
46A	charge	no change	1h	non-spliced	0.65
46A	charge	no change	2h	non-spliced	0.79
46A	charge	no change	3h	non-spliced	0.176
46A	charge	no change	4h	non-spliced	0.835
46A	charge	no change	1h	all spliced	0.26
46A	charge	no change	2h	all spliced	0.470
46A	charge	no change	3h	all spliced	0.910
46A	charge	no change	4h	all spliced	0.442
46A	charge	no change	1h	normal cis-spliced	0.256
46A	charge	no change	2h	normal cis-spliced	0.42
46A	charge	no change	3h	normal cis-spliced	0.48
46A	charge	no change	4h	normal cis-spliced	0.96
46A	charge	no change	1h	reverse cis-spliced	0.27
46A	charge	no change	2h	reverse cis-spliced	0.8
46A	charge	no change	3h	reverse cis-spliced	0.258
46A	charge	no change	4h	reverse cis-spliced	0.251
46A	charge	no change	1h	trans spliced	0.044
46A	charge	no change	2h	trans spliced	0.012
46A	charge	no change	3h	trans spliced	0.67
46A	charge	no change	4h	trans spliced	0.67
46B	charge	neutral to positive	1h	all peptides	0.001
46B	charge	neutral to positive	2h	all peptides	0.163
46B	charge	neutral to positive	3h	all peptides	0.076
46B	charge	neutral to positive	4h	all peptides	0.018
46B	charge	neutral to positive	1h	non-spliced	0.024
46B	charge	neutral to positive	2h	non-spliced	0.31
46B	charge	neutral to positive	3h	non-spliced	0.415
46B	charge	neutral to positive	4h	non-spliced	0.049

46B	charge	neutral to positive	1h	all spliced	0.026
46B	charge	neutral to positive	2h	all spliced	0.264
46B	charge	neutral to positive	3h	all spliced	0.11
46B	charge	neutral to positive	4h	all spliced	0.14
46B	charge	neutral to positive	1h	normal cis-spliced	0.87
46B	charge	neutral to positive	2h	normal cis-spliced	0.256
46B	charge	neutral to positive	3h	normal cis-spliced	0.11
46B	charge	neutral to positive	4h	normal cis-spliced	0.11
46B	charge	neutral to positive	1h	reverse cis-spliced	0.014
46B	charge	neutral to positive	2h	reverse cis-spliced	0.0024
46B	charge	neutral to positive	3h	reverse cis-spliced	5.77e-05
46B	charge	neutral to positive	4h	reverse cis-spliced	0.014
46B	charge	neutral to positive	1h	trans spliced	0.057
46B	charge	neutral to positive	2h	trans spliced	0.28
46B	charge	neutral to positive	3h	trans spliced	0.51
46B	charge	neutral to positive	4h	trans spliced	0.2
46C	charge	neutral to negative	1h	all peptides	1.98e-06
46C	charge	neutral to negative	2h	all peptides	1.36e-08
46C	charge	neutral to negative	3h	all peptides	9.6e-10
46C	charge	neutral to negative	4h	all peptides	2.74e-06
46C	charge	neutral to negative	1h	non-spliced	0.00112
46C	charge	neutral to negative	2h	non-spliced	0.0016
46C	charge	neutral to negative	3h	non-spliced	0.00011
46C	charge	neutral to negative	4h	non-spliced	0.032
46C	charge	neutral to negative	1h	all spliced	5.16e-05
46C	charge	neutral to negative	2h	all spliced	1.65e-06
46C	charge	neutral to negative	3h	all spliced	2.33e-06
46C	charge	neutral to negative	4h	all spliced	1.22e-05

46C	charge	neutral to negative	1h	normal cis-spliced	0.07
46C	charge	neutral to negative	2h	normal cis-spliced	0.236
46C	charge	neutral to negative	3h	normal cis-spliced	0.114
46C	charge	neutral to negative	4h	normal cis-spliced	0.363
46C	charge	neutral to negative	1h	reverse cis-spliced	0.12
46C	charge	neutral to negative	2h	reverse cis-spliced	0.02
46C	charge	neutral to negative	3h	reverse cis-spliced	0.189
46C	charge	neutral to negative	4h	reverse cis-spliced	0.04
46C	charge	neutral to negative	1h	trans spliced	0.00015
46C	charge	neutral to negative	2h	trans spliced	4.74e-06
46C	charge	neutral to negative	3h	trans spliced	1.46e-06
46C	charge	neutral to negative	4h	trans spliced	2.71e-06
47A	hydrophobicity	no change	1h	all peptides	0.77
47A	hydrophobicity	no change	2h	all peptides	0.96
47A	hydrophobicity	no change	3h	all peptides	0.22
47A	hydrophobicity	no change	4h	all peptides	0.21
47A	hydrophobicity	no change	1h	non-spliced	0.86
47A	hydrophobicity	no change	2h	non-spliced	0.87
47A	hydrophobicity	no change	3h	non-spliced	0.61
47A	hydrophobicity	no change	4h	non-spliced	0.88
47A	hydrophobicity	no change	1h	all spliced	0.86
47A	hydrophobicity	no change	2h	all spliced	0.95
47A	hydrophobicity	no change	3h	all spliced	0.26
47A	hydrophobicity	no change	4h	all spliced	0.18
47A	hydrophobicity	no change	1h	normal cis-spliced	0.26
47A	hydrophobicity	no change	2h	normal cis-spliced	0.79
47A	hydrophobicity	no change	3h	normal cis-spliced	0.134
47A	hydrophobicity	no change	4h	normal cis-spliced	0.135
47A	hydrophobicity	no change	1h	reverse cis-spliced	0.01
47A	hydrophobicity	no change	2h	reverse cis-spliced	0.0067
47A	hydrophobicity	no change	3h	reverse cis-spliced	0.00034
47A	hydrophobicity	no change	4h	reverse cis-spliced	0.0019
47A	hydrophobicity	no change	1h	trans spliced	0.29
47A	hydrophobicity	no change	2h	trans spliced	0.027
47A	hydrophobicity	no change	3h	trans spliced	0.91
47A	hydrophobicity	no change	4h	trans spliced	0.98
47B	hydrophobicity	hydrophobic to hydrophilic	1h	all peptides	0.011
47B	hydrophobicity	hydrophobic to hydrophilic	2h	all peptides	0.00013
47B	hydrophobicity	hydrophobic to hydrophilic	3h	all peptides	1.08e-05
47B	hydrophobicity	hydrophobic to hydrophilic	4h	all peptides	0.0056

47B	hydrophobicity	hydrophobic to hydrophilic	1h	non-spliced	0.36
47B	hydrophobicity	hydrophobic to hydrophilic	2h	non-spliced	0.105
47B	hydrophobicity	hydrophobic to hydrophilic	3h	non-spliced	0.017
47B	hydrophobicity	hydrophobic to hydrophilic	4h	non-spliced	0.551
47B	hydrophobicity	hydrophobic to hydrophilic	1h	all spliced	0.011
47B	hydrophobicity	hydrophobic to hydrophilic	2h	all spliced	0.00025
47B	hydrophobicity	hydrophobic to hydrophilic	3h	all spliced	0.00018
47B	hydrophobicity	hydrophobic to hydrophilic	4h	all spliced	0.00013
47B	hydrophobicity	hydrophobic to hydrophilic	1h	normal cis-spliced	0.25
47B	hydrophobicity	hydrophobic to hydrophilic	2h	normal cis-spliced	0.19
47B	hydrophobicity	hydrophobic to hydrophilic	3h	normal cis-spliced	0.059
47B	hydrophobicity	hydrophobic to hydrophilic	4h	normal cis-spliced	0.67
47B	hydrophobicity	hydrophobic to hydrophilic	1h	reverse cis-spliced	0.124
47B	hydrophobicity	hydrophobic to hydrophilic	2h	reverse cis-spliced	0.928
47B	hydrophobicity	hydrophobic to hydrophilic	3h	reverse cis-spliced	0.48
47B	hydrophobicity	hydrophobic to hydrophilic	4h	reverse cis-spliced	0.054
47B	hydrophobicity	hydrophobic to hydrophilic	1h	trans spliced	0.103
47B	hydrophobicity	hydrophobic to hydrophilic	2h	trans spliced	0.0033
47B	hydrophobicity	hydrophobic to hydrophilic	3h	trans spliced	0.00015
47B	hydrophobicity	hydrophobic to hydrophilic	4h	trans spliced	0.0023

Table S21: Statistical test values (Pearson correlation coefficients and p.values (log10 ratio of MUT to WT vs the distance of the P1 residue to mutation)

Figure	group	contrast	peptide type	up in WT or MUT	correlation coefficient	p-value
48A	charge	no change	non-spliced	MUT	-0.25	0.058
48A	charge	no change	non-spliced	WT	-0.27	0.066
48B	charge	neutral to positive	non-spliced	MUT	-0.23	0.08
48B	charge	neutral to positive	non-spliced	WT	-0.064	0.75
48C	charge	neutral to negative	non-spliced	MUT	0.136	0.348

48C	charge	neutral to negative	non-spliced	WT	0.136	0.23
48D	hydrophobicity	no change	non-spliced	MUT	-0.19	0.085
48D	hydrophobicity	no change	non-spliced	WT	-0.146	0.26
48E	hydrophobicity	hydrophobic to hydrophilic	non-spliced	MUT	0.086	0.423
48E	hydrophobicity	hydrophobic to hydrophilic	non-spliced	WT	0.082	0.143
49A	charge	no change	spliced	MUT	0.238	0.031
49A	charge	no change	spliced	WT	-0.18	0.128
49B	charge	neutral to positive	spliced	MUT	0.122	0.31
49B	charge	neutral to positive	spliced	WT	0.23	0.123
49C	charge	neutral to negative	spliced	MUT	-0.296	0.1
49C	charge	neutral to negative	spliced	WT	0.154	0.18
49D	hydrophobicity	no change	spliced	MUT	0.243	0.0064
49D	hydrophobicity	no change	spliced	WT	-0.094	0.365
49E	hydrophobicity	hydrophobic to hydrophilic	spliced	MUT	-0.162	0.211
49E	hydrophobicity	hydrophobic to hydrophilic	spliced	WT	0.206	0.039

Table S22: Statistical test values (Kolmogorov-Smirnov test (difference in distributions of PSP-P1 and SCS-P1 for all substrates combined)

Figure	residue group or residue	median of SCS-P1 (%)	median of PSP-P1 (%)	p-value
53A	hydrophobic	0.36	1.26	1.65e-09
53A	special case	0.07	1.2	1.02e-14
53A	polar uncharged	0.23	0.95	0.000134
53A	polar positive	0.49	1.57	0.000301
53A	polar negative	0.71	0.77	0.024
53B	A	3.12	2.27	0.42
53B	V	0.25	1.04	0.0025
53B	L	3.7e-07	0.68	4.24e-10
53B	M	5.2	2.58	0.27
53B	F	0.41	2.35	0.07
53B	Y	0.65	2.6	0.056
53B	W	0.06	0.48	0.87
53C	C	0.12	1.01	0.088
53C	G	0.08	1.1	1.17e-10
53C	P	0.025	5.34	0.0033
53D	S	0.21	1.16	0.02
53D	T	0.075	1.59	0.00065
53D	N	20.2	2.88	0.164
53D	Q	0.441	0.61	0.541
53E	R	0.39	1.61	0.037
53E	H	0.39	1.61	0.00064
53E	K	0.19	0.39	0.037
53F	D	1.37	1.5	0.097
53F	E	0.29	0.46	0.172

Table S23: Statistical test values (Pearson correlation coefficients and p.values (SCS-P1 in WT vs SCS-P1 in MUT polypeptides)

Figure	substrate and mutation	correlation coefficient	p-value
54	IDH1 R132H	0.9	2.021e-10
54	JAK2 V617F	0.96	2.2e-16
54	BRAF V600E	0.52	0.00253
54	MPL W515L	0.77	1.031e-05
54	KRAS G12D	0.85	2.49e-10
54	KRAS G12R	0.55	0.0008
54	KRAS G13D	0.92	1.258e-14
54	NRAS Q61K	0.58	0.00061
54	NRAS Q61R	0.7	1.08e-05

Table S24: Statistical test values (Pearson correlation coefficients and p.values (PSP-P1 in WT vs PSP-P1 in MUT polypeptides)

Figure	substrate and mutation	correlation coefficient	p-value
55	IDH1 R132H	-0.02	0.933
55	JAK2 V617F	0.097	0.634
55	BRAF V600E	0.37	0.043
55	MPL W515L	0.15	0.48
55	KRAS G12D	0.39	0.022
55	KRAS G12R	0.09	0.595
55	KRAS G13D	0.22	0.21
55	NRAS Q61K	0.17	0.3567
55	NRAS Q61R	0.33	0.068

Table S25: Statistical test values (Kolmogorov-Smirnov test (difference in distributions of SCS-P1 between WT and MUT polypeptides)

Figure	group	contrast	residue group	median of SCS-P1 in WT (%)	median of SCS-P1 in MUT (%)	p-value
56A	charge	no change	hydrophobic	0.13	0.22	0.99
56A	charge	no change	special case	0.2	0.52	0.46
56A	charge	no change	polar uncharged	0.0064	0.02	0.988
56A	charge	no change	positively charged	0.31	0.33	0.999
56A	charge	no change	negatively charged	0.75	0.7	0.999
56B	charge	neutral to positive	hydrophobic	0.33	0.36	0.56
56B	charge	neutral to positive	special case	0.009	5e-06	0.78
56B	charge	neutral to positive	polar uncharged	0.4	0.35	0.884
56B	charge	neutral to positive	positively charged	0.49	0.52	0.67

56B	charge	neutral to positive	negatively charged	0.84	0.35	0.964
56C	charge	neutral to negative	hydrophobic	0.98	1.23	0.87
56C	charge	neutral to negative	special case	0.04	0.195	0.123
56C	charge	neutral to negative	polar uncharged	0.17	0.66	0.215
56C	charge	neutral to negative	positively charged	0.5	0.7	0.4611
56C	charge	neutral to negative	negatively charged	0.7	0.5	0.952
56D	hydrophobicity	no change	hydrophobic	0.13	0.22	0.71
56D	hydrophobicity	no change	special case	0.001	9.86e-06	0.57
56D	hydrophobicity	no change	polar uncharged	0.008	0.09	0.94
56D	hydrophobicity	no change	positively charged	0.45	0.46	0.99
56D	hydrophobicity	no change	negatively charged	0.84	0.63	0.972
56E	hydrophobicity	hydrophobic to hydrophilic	hydrophobic	0.94	0.67	0.86
56E	hydrophobicity	hydrophobic to hydrophilic	special case	0.05	0.2	0.027
56E	hydrophobicity	hydrophobic to hydrophilic	polar uncharged	0.23	0.34	0.37
56E	hydrophobicity	hydrophobic to hydrophilic	positively charged	0.49	0.4	0.43
56E	hydrophobicity	hydrophobic to hydrophilic	negatively charged	0.7	0.8	0.822

Table S26: Statistical test values (Kolmogorov-Smirnov test (difference in distributions of PSP-P1 between WT and MUT polypeptides)

Figure	group	contrast	residue group	median of PSP-P1 in WT (%)	median of PSP-P1 in MUT (%)	p-value
57A	charge	no change	hydrophobic	1.445	2.42	0.28
57A	charge	no change	special case	0.68	1.59	0.48
57A	charge	no change	polar uncharged	0.473	0.86	0.79
57A	charge	no change	positively charged	1.82	2.1	0.999
57A	charge	no change	negatively charged	1.275	0.73	0.873
57B	charge	neutral to positive	hydrophobic	1.72	1.36	0.48
57B	charge	neutral to positive	special case	1.16	1.1	0.492
57B	charge	neutral to positive	polar uncharged	1.04	1.41	0.214
57B	charge	neutral to positive	positively charged	0.93	2.1	0.103
57B	charge	neutral to positive	negatively charged	0.77	0.5	0.491

57C	charge	neutral to negative	hydrophobic	1.43	0.5	0.038
57C	charge	neutral to negative	special case	1.11	1.5	0.72
57C	charge	neutral to negative	polar uncharged	1.72	0.312	0.0082
57C	charge	neutral to negative	positively charged	1.47	3.21	0.461
57C	charge	neutral to negative	negatively charged	0.81	1.03	0.8
57D	hydrophobicity	no change	hydrophobic	1.54	2.42	0.425
57D	hydrophobicity	no change	special case	1.15	1.22	0.972
57D	hydrophobicity	no change	polar uncharged	0.55	0.91	0.36
57D	hydrophobicity	no change	positively charged	1.38	2.44	0.2
57D	hydrophobicity	no change	negatively charged	0.77	0.58	0.794
57E	hydrophobicity	hydrophobic to hydrophilic	hydrophobic	1.43	0.56	0.0072
57E	hydrophobicity	hydrophobic to hydrophilic	special case	1.12	1.5	0.33
57E	hydrophobicity	hydrophobic to hydrophilic	polar uncharged	1.62	0.42	0.014
57E	hydrophobicity	hydrophobic to hydrophilic	positively charged	1.47	0.32	0.26
57E	hydrophobicity	hydrophobic to hydrophilic	negatively charged	0.81	0.9	0.552

References

[1] Kisselev A., Linden W., and Overkleeft H. “Proteasome Inhibitors: An Expanding Army Attacking a Unique Target”. In: *Cell Chemistry & Biology* 19 (2012), pp. 99–115. DOI: https://doi.org/10.1016/j.chembiol.2012.01.003.

[2] Ebstein F., Textoris-Taube K., Keller C., Golnik R., Vigneron N., Van den Eynde J., and *et al.* “Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes”. In: *Nature Scientific Reports* 6.24032 (2016). DOI: https://doi.org/10.1038/srep24032.

[3] Shwartz A. and Ciechanover A. “Targeting proteins for destruction by the ubiquitin system: implications for human pathobiology”. In: *Annual review of pharmacology and toxicology* 49 (2009), pp. 73–96. DOI: https://doi.org/10.1146/annurev.pharmtox.051208.165340.

[4] Ben-Nissan G. and Sharon M. “Regulating the 20S proteasome ubiquitin-independent degradation pathway”. In: *Biomolecules* 4.3 (2014), pp. 862–884. DOI: https://doi.org/10.1126/scitranslmed.aax4100.

[5] Kloetzel P. “Antigen processing by the proteasome”. In: *Nature reviews. Molecular and Cellular Biology* 2.3 (2001), pp. 179–187. DOI: https://doi.org/10.1038/35056572.

[6] Platteel A., Liepe J., Eden W., Mishto M., and Sitjs A. “An Unexpected Major Role for Protrasome-Catalyzed Peptide Splicing in Generation of T Cell Epitopes: Is There Relevance for Vaccine Development”. In: *Frontiers in Immunology* 8.1441 (2017), pp. 1–6. DOI: https://doi.org/10.3389/fimmu.2017.01441.

[7] Mishto M. and Liepe J. “Post-Translational Peptide Splicing and T Cell Responses”. In: *Trends in Immunology* 38 (2017), pp. 904–915. DOI: https://doi.org/10.1016/j.it.2017.07.011.

[8] Hanada K., Ywedell J., and Yand J. “Immune recognition of a human renal cancer antigen through post-translational protein splicing”. In: *Nature* 427 (2004), pp. 252–256. DOI: https://doi.org/10.1038/nature02240.

[9] Dalet A., Vignerorn N., Stroobant V., Hanada K., and Van den Eynde B. “Splicing of Distant Peptide Fragments Occurs in the Proteasome by Transpeptidation and Produces the Spliced Antigenic Peptide Derived from Fibroblast Growth Factor-5”. In: *The Journal of Immunology* 184 (2010), pp. 3016–3024. DOI: https://doi.org/10.4049/jimmunol.0901277.

[10] Vigneron N., Stroobant V., Chapiro J., Ooms A., Degiovanni V., Morel S., and *et al.* “An Antigenic Peptide Produced by Peptide Splicing in the Proteasome”. In: *Science* 304 (2004), pp. 587–589. DOI: https://doi.org/10.1126/science.1095522.

[11] Warren E., Vigneron N., Gavin M., Coulie P., Stroobant V., Dalet A., and *et al*. "An Antigen Produced by Splicing of Noncontiguous Peptides in the Reverse Order". In: *Science* 313 (2006), pp. 1444–1447. DOI: https://doi.org/10.1126/science.1130660.

[12] Dalet A., Robbins PF., Stroobant V., Vigneron N.., Li YF., El-Gamil M., and *et al*. "An antigenic peptide produced by reverse splicing and double asparagine deamidation". In: *PNAS* 108 (2011), E323–E331. DOI: https://doi.org/10.1073/pnas.1101892108.

[13] Liepe J., Marino F., Sidney J., Jeko A., Bunting D., Sette A., and *et al*. "A large fraction of HLA class I ligands are proteasome-generated spliced peptides". In: *Science* 354.6310 (2016), pp. 354–358. DOI: https://doi.org/10.1126/science.aaf4384.

[14] Liepe J., Sidney J., Lorenz F., Sette A., and Mishto M. "Mapping the MHC Class I-Spliced Immunopeptidome of Cancer Cells". In: *Cancer Immunology Research* 7.1 (2019), pp. 62–76. DOI: https://doi.org/10.1158/2326-6066.CIR-18-0424.

[15] Woods K., Faridi P.., Ostrouska S., Deceneux C., Wong S., Chen W., and *et al*. "The diversity of the immunogenic components of the melanoma immunopeptidome". In: *bioRxiv* (2019). DOI: https://doi.org/10.1101/623223.

[16] Xu J. and Jo J. "Broad cross-reactivity of the T-cell repertoire achieves specific and sufficiently rapid target searching". In: *Journal of Theoretical Biology* 466 (2019), pp. 119–127. DOI: https://doi.org/10.1016/j.jtbi.2019.01.025.

[17] Specht G., Roetschke H., Mansurkhodzhaev A., Textoris-Taube K., Henklein P., Urlaub H., and *et al*. "Large database for the analysis and prediction of spliced and non-spliced peptide generation by proteasomes". In: *Nature Scientific Data* (2020). DOI: https://doi.org/10.6084/m9.figshare.12205274.

[18] Berkers C., de Jong A., Schuurman K., Linnemann C., Miring H., Janssen L., and *et al*. "Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules". In: *The Journal of Immunology* 195 (2015), pp. 4085–4095. DOI: https://doi.org/10.4049/jimmunol.1402455.

[19] Paes W., Leonov G., Partridge T., Nicastri A., Ternette N., Borrow P., and *et al*. "Elucidation of the Signatures of Proteasome-Catalysed Peptide Splicing". In: *Frontiers in Immunology* 11 (2020), pp. 1–13. DOI: https://doi.org/10.3389/fimmu.2020.563800.

[20] Platteel A., Liepe J., Textoris-Taube K., Keller C., Henklein P., Schalkwijk H., and *et al*. "Multi-level strategy for identifying proteasome-catalyzed spliced epitopes targeted by CD8+ T cells during bacterial infection". In: *Cell reports* 20.5 (2017), pp. 1242–1253. DOI: https://doi.org/10.1016/j.celrep.2017.07.026.

[21] Paes W., Leonov G., Partridge T., Chikata T., Murakoshi H., Frangou A., and et a. “Contribution of proteasome-catalyzed peptide cis-splicing to viral targeting by CD8+ T cells in HIV-1 infection”. In: *PNAS* 116.49 (2019), pp. 24748–24759. DOI: https://doi.org/10.1073/pnas.1911622116.

[22] Robbins PF., el Gamil M., Kawakami Y., Stevens E., Yannelli JR., and Rosenberg SA. “Recognition of tyrosinase by tumor-infiltrating lymphocytes from a patient responding to immunotherapy”. In: *Cancer Research* 54 (1994), pp. 3124–3126.

[23] Faridi P., Woods K., Ostrouska S., Deceneux C., Aranha R., Duscharla D., and *et al.* “Spliced peptides and cytokine-driven changes in the immunopeptidome of melanoma”. In: *Cancer Immunology Research* 8 (2020), pp. 1322–1334. DOI: https://doi.org/10.1158/2326-6066.CIR-19-0894.

[24] Calis J., Boer R., and Kesmir C. “Degenerate T-cell Recognition of Peptides on MHC Molecules Creates Large Holes in the T-cell Repertoire”. In: *Plos Computational Biology* 8.3 (2012). DOI: https://doi.org/10.1371/journal.pcbi.1002412.

[25] Grignolio A., Mishto M., Faria A., Garagnani P., Franceschi C., and Tieri P. “Towards a Liquid Self: How Time, Geography, and Life Experiences Reshape the Biological Identity”. In: *Frontiers in Immunology* 5 (2014), p. 153. DOI: https://doi.org/10.3389/fimmu.2014.00153.

[26] Mishto M., Mansurkhodzhaev A., Rodriguez-Calvo T., and Liepe J. “Potential mimicry of viral and pancreatic b cell antigens through non-spliced and cis-spliced zwitter epitope candidates in Type 1 Diabetes”. In: *Frontiers in Immunology* 10 (2021), p. 2572. DOI: https://doi.org/10.3389/fimmu.2021.656451.

[27] Mishto M., Goede A., Textoris-Taube K., Keller C., Janek K., Henklein P., and *et al.* “Driving Forces of Proteasome-catalyzed Peptide Splicing in Yeast and Humans”. In: *Molecular and Cellular Proteomics* 11.10 (2012), pp. 1008–1023. DOI: https://doi.org/10.1074/mcp.M112.020164.

[28] Textoris-Taube K., Keller C., Liepe J., Henklein P., Sideny J., Sette A., and *et al.* “The T210M substitution in the HLA-A*02:01 gp100 epitope strongly affects overall proteasomal cleavage site usage and antigen processing”. In: *Journal of Biological Chemistry* 290.51 (2015), pp. 3041730428. DOI: https://doi.org/10.1074/jbc.M115.695189.

[29] Fidanza M., Gupta P., Sayana A., Shanker V., Pahlke S., Vu B., and *et al.* “Enhancing proteasomal processing improves survival for a peptide vaccine used to treat glioblastoma”. In: *SCIENCE TRANSLATIONAL MEDICINE* 13.598 (2021), eaax4100. DOI: https://doi.org/10.1126/scitranslmed.aax4100.

[30] Marshall J., Warrington R., Watson W., and Lim H. “An introduction to immunology and immunopathology”. In: *Allergy, asthma and clinical immunology* 14.49 (2018). DOI: https://doi.org/10.1186/s13223-018-0278-1.

[31] Chaplin D. “Overview of the Immune Response”. In: *Journal of Allergy and Clinical Immunology* 125 (2010), S2–S3. DOI: https://doi.org/10.1016/j.jaci.2009.12.980.

[32] Abbas A., Lichtman A., and Pillai S. “Cellular and Molecular Immunology”. In: *Elsevier Sanders* (2015).

[33] Raskov H., Orhan A., Christensen J., and Gogenur I. “Cytotoxic CD8+ T cells in cancer and cancer immunotherapy”. In: *Nature* 124 (2021), pp. 359–367. DOI: https://doi.org/10.1038/s41416-020-01048-4.

[34] Xie Z., Chang C., and Zhou Z. “Molecular mechanisms in autoimmune type 1 diabetes: a critical review”. In: *Clinical Review In Allergy and Immunology* 47.2 (2014), pp. 174–192. DOI: https://doi.org/10.1007/s12016-014-8422-2.

[35] Anton L. and Yewdell J. “Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors”. In: *Journal of Leukocyte Biology* 95.4 (2014), pp. 551–562. DOI: https://doi.org/10.1189/jlb.1113599.

[36] Yewdell J., Reits E., and Neefjes J. “Making sense of mass destruction: quantitating MHC class I antigen presentation”. In: *Nature Reviews, Immunology* 3.12 (2003), pp. 952–961. DOI: https://doi.org/10.1038/nri1250.

[37] Murata S., Sasaki K., Kishimoto T., Niwa S., Hayashi H., Takahama Y., and Tanaka K. “Regulation of CD8+ T cell development by thymus-specific proteasomes”. In: *Science* 316.5829 (2007), pp. 1349–1353. DOI: https://doi.org/10.1126/science.1141915.

[38] Wong P. and Pamer E. “CD8 T cell responses to infectious pathogens”. In: *Annual Review of Immunology* 21 (2003), pp. 29–70. DOI: https://doi.org/10.1146/annurev.immunol.21.120601.141114.

[39] Schmidt M. and Varga S. “The CD8 T Cell Response to Respiratory Virus Infections”. In: *Frontiers in Immunology* 9 (2018), p. 678. DOI: https://doi.org/10.3389/fimmu.2018.00678.

[40] Houen G., Trier N., and Frederiksen J. “Epstein-Barr Virus and Multiple Sclerosis”. In: *Frontiers in Immunology* 11 (2020). DOI: https://doi.org/10.3389/fimmu.2020.587078.

[41] Fujiya A., Ochiai H., Mizukoshi T., Kiyota A., Shibata T., Suzuki A., Ohashi N., and Sobajima H. “Fulminant type 1 diabetes mellitus associated with a reactivation of Epstein-Barr virus that developed in the course of chemotherapy of multiple myeloma”. In: *Journal of diabetes investigation* 1.6 (2010), pp. 286–289. DOI: https://doi.org/10.1111/j.2040-1124.2010.00061.x..

[42] van der Leun A., Thommen D., and Schumacher T. “CD8 + T cell states in human cancer: insights from single-cell analysis”. In: *Nature Reviews. Cancer* 20.4 (2020), pp. 218–232. DOI: https://doi.org/10.1038/s41568-019-0235-4.

[43] Jiang W., He Y., He W., Wu G., Zhou X., Sheng Q., and *et al*. "Exhausted CD8+T Cells in the Tumor Immune Microenvironment: New Pathways to Therapy". In: *Frontiers in Immunology* 11 (2021), p. 622509. DOI: https://doi.org/10.3389/fimmu.2020.622509.

[44] Kubuschok B., Neumann F., Breit R., Sester M., Schormann C, Wagner C., and *et al*. "Naturally Occurring T-Cell Response against Mutated p21 Ras Oncoprotein in Pancreatic Cancer". In: *Clinical Cancer Resarch* 12.4 (2006), pp. 1365–1372. DOI: https://doi.org/10.1158/1078-0432.CCR-05-1672.

[45] Maoz A., Rennert G., and Gruber S. "T-Cell Transfer Therapy Targeting Mutant KRAS". In: *The New England Journal of Medicine* 367.7 (2018), e11. DOI: https://doi.org/10.1056/NEJMc1616637.

[46] Blankenstein T., Leisegang M., Uckert W., and Schreiber H. "Targeting cancer-specific mutations by T cell receptor gene therapy". In: *Current Opinion in Immunology* 33 (2015), pp. 112–119. DOI: https://doi.org/10.1016/j.coi.2015.02.005.

[47] Coulie P., van den Eynde B., vand den Bruggen P., and Boon T. "Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy". In: *Nature Reviews. Cancer* 14.2 (2014), pp. 135–146. DOI: https://doi.org/10.1038/nrc3670.

[48] Rosenberg S. and Restifo N. "Adoptive cell transfer as personalized immunotherapy for human cancer". In: *Science* 348.6230 (2015), pp. 62–68. DOI: https://doi.org/10.1126/science.aaa4967.

[49] Klein L., Kyewski B., Allen P., and Hogquist K. "Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see)". In: *Nature reviews, Immunology* 14.6 (2014), pp. 377–391. DOI: https://doi.org/10.1038/nri3667.

[50] ElTanbouly M. and Noelle R. "Rethinking peripheral T cell tolerance: checkpoints across a T cells journey". In: *Nature Immunology* (2020). DOI: https://doi.org/10.1038/s41577-020-00454-2.

[51] Culina S., Lalanne G., Cerosaletti K., Pinto S., Sebastiani G., Kuranda K., and *et al*. "Islet-reactive CD8+ T-cell frequencies in the pancreas but not blood distinguish type 1 diabetes from healthy donors". In: *Science Immunology* 2.3 (2018), eaao4013. DOI: https://doi.org/10.1126/sciimmunol.aao4013.

[52] Calis J., Maybenno M., Grennbaum J., Weiskopf D., De Silva A., Sette A., and *et al*. "Properties of MHC Class I Presented Peptides That Enhance Immunogenicity". In: *Plos computational biology* 9.10 (2013). DOI: https://doi.org/10.1371/journal.pcbi.1003266.

[53] Pickering A. and Davies K. "Degradation of damaged proteins: the main function of the 20S proteasome". In: *Progress in Molecular Biology and Translational Science* 109 (2012), pp. 227–248. DOI: https://doi.org/10.1016/B978-0-12-397863-9.00006-7.

[54] Schmidt M. and Finley D. “Regulation of proteasome activity in health and disease”. In: *Biochimica et Biophysica Acta* 1843.1 (2013), pp. 13–25. DOI: https://doi.org/10.1016/j.bbamcr.2013.08.012.

[55] Vingeron N. and Van den Eynde B. “Proteasome Subtypes and Regulators in the Processing of Antigenic Peptides Presented by Class I Molecules of the Major Histocompatibility Complex”. In: *Biomolecules* 4.4 (2014), pp. 9941025. DOI: https://doi.org/10.3390/biom4040994.

[56] Sharon M., Witt S., Felderer K., Rockel B., Baumeister W., and Robinson V. “20S Proteasomes Have the Potential to Keep Substrates in Store for Continual Degradation”. In: *Journal of Biological Chemistry* 281.14 (2006), pp. 9569–9575. DOI: https://doi.org/10.1074/jbc.M511951200.

[57] Borissenko L. and Groll M. “Diversity of proteasomal missions: fine tuning of the immune response”. In: *Biological Chemistry* 388.9 (2007), pp. 947–955. DOI: https://doi.org/10.1515/BC.2007.109.

[58] Liepe J., Ovaa H., and Mishto M. “Why do proteases mess up with antigen presentation by re-shuffling antigen sequences”. In: *Current opinion in immunology* 52 (2018), pp. 81–86. DOI: https://doi.org/10.1016/j.coi.2018.04.016.

[59] Kniepert A. and Groettrup M. “The unique functions of tissue-specific proteasomes”. In: *Trends in Biochemical Sciences* 39.1 (2013), pp. 17–24. DOI: https://doi.org/10.1016/j.tibs.2013.10.004.

[60] Ebstein F., Kloetzel P., Kruger E., and Seifert U. “Emerging roles of immunoproteasomes beyond MHC class I antigen processing”. In: *Cellular and Molecular Life Sciences* 69.15 (2012), pp. 2543–2558. DOI: https://doi.org/10.1007/s00018-012-0938-0.

[61] Emmerich N., Nussbaum A., Stevanovic S., Priemer M., Toes R., Rammensee H., and Schild H. “The human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate”. In: *Journal of Biological Chemistry* 275.28 (2000), pp. 21140–21148. DOI: https://doi.org/10.1074/jbc.M000740200.

[62] Erales J. and Coffino P. “Ubiquitin-independent proteasomal degradation”. In: *Molecular Cell Research* 1843.1 (2014), pp. 216–221. DOI: https://doi.org/10.1016/j.bbamcr.2013.05.008.

[63] Berkers C., de Jong A., Schuurman K., Linnemann C., Meiring H., Janssen L., and *et al.* “Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules”. In: *Jounral of Immunology* 195 (2015), pp. 4085–4095. DOI: https://doi.org/10.4049/jimmunol.1402455.

[64] Michaux A., Larrieu P., Stroobant V., Fonteneau J., Jotereau F.., Van den Eynde B., and *et al.* “A Spliced Antigenic Peptide Comprising a Single Spliced Amino Acid Is Produced in the Proteasome by Reverse Splicing of a Longer Peptide Fragment followed by Trimming”. In: *Journal of Immunology* 192 (2014), pp. 1962–1971. DOI: https://doi.org/10.4049/jimmunol.1302032.

[65] Mishto M., Liepe J., Textoris-Taube K., Keller C., Henklein P., Weberrus M., and *et al*. “Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation”. In: *European Journal of Immunology* 44 (2014), pp. 3508–3521. DOI: https://doi.org/10.1002/eji.201444902.

[66] Liepe J., Holzhutter H., Bellvista E., Kloetzel P., Stumpf F., and Mishto M. “Quantitative time-resolved analysis reveals intricate, differential regulation of standard- and immuno-proteasomes”. In: *eLife* 4 (2015), e07545. DOI: https://doi.org/10.7554/eLife.07545.

[67] Kuckelkorn U., Stubler S., Textoris-Taube K., Kilian C., Niewienda A., Henklein P., and *et al*. “Proteolytic dynamics of human 20S thymoproteasome”. In: *Journal of Autoimmunity* 294.19 (2019), pp. 7740–7754. DOI: https://doi.org/10.1074/jbc.RA118.007347.

[68] Winter M., Greca F., Arastu-Kapur S., Caiazza F., Cimermancic P., Bucholz T., and *et al*. “Immunoproteasome functions explained by divergence in cleavage specificity and regulation”. In: *eLife* 6 (2017), e27364. DOI: https://doi.org/10.7554/eLife.27364.

[69] Chen H., Li L., Weimershaus M., Evnouchidou I., Endert P., and Bouvier M. “ERAP1-ERAP2 dimers trim MHC I-bound precursor peptides; implications for understanding peptide editing”. In: *Nature Scientific Reports* 6.1 (2016), p. 28902. DOI: https://doi.org/10.1038/srep28902.

[70] Faridi P., Li C., Ramarathinam S., Vivian J., Illing P., Mifsud N., and *et al*. “A subset of HLA-I peptides are not genomically templated: Evidence for cis-and trans-spliced peptide ligands”. In: *Science Immunology* 3.28 (2018). DOI: https://doi.org/10.1126/sciimmunol.aar3947.

[71] Platteel A., Mishto M., Textoris-Taube K., Keller C., Liepe J., Busch D., and *et al*. “CD8+ T cells of Listeria monocytogenes-infected mice recognize both linear and spliced proteasome products”. In: *European Journal of Immunology* 46 (2016), pp. 1109–1118. DOI: https://doi.org/10.1002/eji.201545989.

[72] Mishto M., Mansurkhodzhaev A., Ying G., Bitra A., Cordfunke R., Henze S., and *et al*. “An in silicoin vitro Pipeline Identifying an HLA-A*02:01+ KRAS G12V+ Spliced Epitope Candidate for a Broad Tumor-Immune Response in Cancer Patients”. In: *Frontiers in immunology* 10.2572 (2019). DOI: https://doi.org/10.3389/fimmu.2019.02572.

[73] Bassani-Sternberg M., Mylonas R., Beer I., Iseli C., Chong C., Pak H., and *et al*. “Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome”. In: *Molecular and Cellular Proteomics* 17 (2018), pp. 2347–2357. DOI: https://doi.org//mcp.RA118.000877.

[74] Rolfs Z., Solntsev S., Shortreed M., Frey B., and Lloyd S. “Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion”. In: *Journal of Proteome Research* 18.1 (2018), pp. 349–358. DOI: https://doi.org/10.1021/acs.jproteome.8b00651.

[75] Kelly M., Rayner M., Mujovic C., and Barnett A. “Molecular aspects of type 1 diabetes”. In: *Molecular Pathology* 56.1 (2003), pp. 1–10. DOI: https://doi.org/10.1136/mp.56.1.1.

[76] Roep B., Thomaidou S., Tienhoven R., and Zaldumbide A. “Type 1 diabetes mellitus as a disease of the -cell (do not blame the immune system?)” In: *Nature Reviews. Endocrinology* 17.3 (2020), pp. 150–161. DOI: https://doi.org/10.1038/s41574-020-00443-4.

[77] Sinha A. and Mann M. “A beginners guide to mass spectrometrybased proteomics”. In: *The Biochemist* 42.5 (2020), pp. 64–69. DOI: https://doi.org/10.1042/BIO20200057.

[78] “https://www.matrixscience.com/help.html”. In: ().

[79] Zhang Y., Fonslow B., Shang B., Baek M., and Yates J. “Protein Analysis by Shotgun/Bottom-up Proteomics”. In: *Chemical Reviews* 113.4 (2013), 23432394. DOI: https://doi.org/10.1021/cr3003533.

[80] Dupree E., Jayathirtha M., Yorkey H., Mihasan M., Pete B., and Darie C. “A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field”. In: *Proteomes* 8.14 (2020), pp. 1–26. DOI: https://doi.org/10.3390/proteomes8030014.

[81] Zubarev R. and Makarov A. “Orbitrap Mass Spectrometry”. In: *Analytical Chemistry* 85.11 (2013), 52885296. DOI: https://doi.org/10.1021/ac4001223.

[82] Mansurkhodzhaev A., Barbosa CRR., Mishto M., and Liepe J. “Proteasome-Generated cis-Spliced Peptides and Their Potential Role in CD8+ T Cell Tolerance”. In: *Frontiers in Immunology* 12 (2021), p. 614276. DOI: https://doi.org/10.3389/fimmu.2021.614276.

[83] Huseby E., White J., Crawford F., Vass T., Becker D., Pinilla S., and *et al.* “How the T cell repertoire becomes peptide and MHC specific”. In: *Cell* 122 (2005), pp. 247–260. DOI: https://doi.org/10.1016/j.cell.2005.05.013.

[84] Yu W., Jiang N., Ebert P., Kidd B., Muller S., Lund P., and *et al.* “Clonal Deletion Prunes but Does Not Eliminate Self-Specific alpha-beta CD8(+) T Lymphocytes”. In: *Immunity* 42 (2015), pp. 929–941. DOI: https://doi.org/10.1016/j.immuni.2015.05.001.

[85] Klein L., Kyewski B., Allen PM., and Hogquist KA. “Positive and negative selection of the T cell repertoire: what thymocytes see (and dont see)”. In: *Nature Reviews Immunology* 14 (2014), pp. 377–391. DOI: https://doi.org/10.1038/nri3667.

[86] Bouneaud C., Kourilsky P., and Bousso P. “Impact of Negative Selection on the T Cell Repertoire Reactive to a Self-Peptide: A Large Fraction of T Cell Clones Escapes Clonal Deletion”. In: *Immunity* 13 (2000), pp. 829–840. DOI: https://doi.org/10.1016/S1074-7613(00)00080-7.

[87] Ogishi M. and Yotsuyanagi H. "Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space". In: *Frontiers in Immunology* 10.827 (2019). DOI: https://doi.org/10.3389/fimmu.2019.00827.

[88] Kanduc D., Stufano A., Lucchese G., and Kusalik A. "Massive peptide sharing between viral and human proteomes". In: *Peptides* 29 (2008), pp. 1755–1766. DOI: https://doi.org/10.1016/j.peptides.2008.05.022.

[89] Kusalik A., Bickis M., Lewis C., Li Y., Luchhese G., Marincola FM., and *et al.* "Widespread and ample peptide overlapping between HCV and Homo sapiens proteomes". In: *Peptides* 28 (2007), pp. 1260–1267. DOI: https://doi.org/10.1016/j.peptides.2007.04.001.

[90] Ricco R. and Kanduc D. "Hepatitis B virus and Homo sapiens proteome-wide analysis: A profusion of viral peptide overlaps in neuron-specific human proteins". In: *Biologics:Targets&Therapy* 4 (2010), pp. 75–81. DOI: https://doi.org/10.2147/btt.s8890.

[91] Trost B., Kusalik A., Lucchese G., and Kanduc D. "Bacterial peptides are intensively present throughout the human proteome". In: *Self/Nonself* 1.1 (2010), pp. 71–74. DOI: https://doi.org/10.4161/self.1.1.9588.

[92] Trost B., Lucchese G., Stufano A., Bickis M., Kusalik A., and Kanduc D. "No human protein is exempt from bacterial motifs, not even one". In: *Self/Nonself* 1.4 (2010), pp. 328–334. DOI: https://doi.org/10.4161/self.1.4.13315.

[93] Burroughs N., De Boer R., and Kesmir C. "Discriminating self from nonself with short peptides from large proteomes". In: *Immunogenetics* 56 (2004), pp. 311–320. DOI: https://doi.org/10.1007/s00251-004-0691-0.

[94] Rolland M., Nickle D., Deng W., Frahm N., Brander C., Learn G., and *et al.* "Recognition of HIV-1 peptides by host CTL is related to HIV-1 similarity to human proteins." In: *PLos ONE* e823 (2007). DOI: https://doi.org/10.1371/journal.pone.0000823.

[95] Assarsson E., Sidney J., Oseroff C., Pasquetto V., Bui H., Frahm N., and *et al.* "A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection". In: *Journal of Immunology* 178 (2007), pp. 7890–7901. DOI: https://doi.org/10.4049/jimmunol.178.12.7890.

[96] Croft A., Smith S., Pickering J., Sidney J., Peters B., Faridi P., and *et al.* "Most viral peptides displayed by class I MHC on infected cells are immunogenic". In: *PNAS* 116.8 (2019), pp. 3112–3117. DOI: https://doi.org/10.1073/pnas.1815239116.

[97] Wolf M., Rutebemberwa A., Mosbruger T., Mao Q., Li HM., and Netski D. *et al.* "Hepatitis C virus immune escape via exploitation of a hole in the T cell repertoire." In: *Journal of Immunology* 181 (2008), pp. 6435–6446. DOI: https://doi.org/10.4049/jimmunol.181.9.6435.

[98] Bairoch A. and Apweiler R. “The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999”. In: *Nucleic Acids Research* 27.1 (1999), pp. 49–54. DOI: https://doi.org/10.1093/nar/27.1.49.

[99] Hulo C., Castro E., Masson P., Bougueleret L., Bairoch A., Xenarios I., and *et al*. “ViralZone: a knowledge resource to understand virus diversity”. In: *Nucleic Acids Research* 39 (2011), pp. 576–582. DOI: https://dx.doi.org/10.1093%2Fnar%2Fgkq901.

[100] Peters B., Tong W., Sidney J., Sette A., and Weng Z. “Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules”. In: *Bioinformatics* 19 (2003), pp. 1765–1772.

[101] Pinto S., Michel C., Schmidt-Glenewinkel H., Harder N., Rohr K., Wild S., and *et al*. “Overlapping gene co-expression patterns in human medullary thymic epithelial cells generate self-antigen diversity”. In: *PNAS* 110.37 (2013), pp. 3497–3505. DOI: https://doi.org/10.1073/pnas.1308311110.

[102] Zeng Y., Liu C., Gong Y., Bai Z., Hou S., He J., and *et al*. “Single-Cell RNA Sequencing Resolves Spatiotemporal Development of Pre-thymic Lymphoid Progenitors and Thymus Organogenesis in Human Embryos”. In: *Immunity* 51.9 (2019), pp. 930–948. DOI: https://doi.org/10.1016/j.immuni.2019.09.008.

[103] Satija R., Farrell Ja., Gennert D., Schier A., and Regev A. “Spatial reconstruction of single-cell gene expression data”. In: *Nature Biotechnology* 33.5 (2015). DOI: https://doi.org/10.1038/nbt.3192.

[104] Stuart T., Butler A., Hoffman P., Hafemeister C., Papalexi E., Mauck W., and *et al*. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177.7 (2019), pp. 1888–1902. DOI: https://doi.org/10.1016/j.cell.2019.05.031.

[105] Chapiro J., Claverol S., Piette F., Ma W., Stroobant T., Guillaume B., and *et al*. “Destructive Cleavage of Antigenic Peptides Either by the Immunoproteasome or by the Standard Proteasome Results in Differential Antigen Presentation”. In: *Journal of immunology* 176 (2006), pp. 1053–1061. DOI: https://doi.org/10.4049/jimmunol.176.2.1053.

[106] Deol P., Zais D., Monaco J., and Sitjs A. “Rates of Processing Determine the Immunogenicity of Immunoproteasome-Generated Epitopes”. In: *Journal of immunology* 178 (2007), pp. 7557–7562. DOI: https://doi.org/10.4049/jimmunol.178.12.7557.

[107] Guillaume B., Stroobant V., Bousquet-Dubouch M., Colau D., Chapiro J., Parmentier N., and *et al*. “Analysis of the Processing of Seven Human Tumor Antigens by Intermediate Proteasomes”. In: *Journal of immunology* 189 (2012), pp. 3538–3547. DOI: https://doi.org/10.4049/jimmunol.1103213.

[108] Tenzer S., Wee E., Burgevin A., Stewart-Jones G., Friis L., Lamberth K., and *et al*. "Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance". In: *Nature* 10.6 (2009), pp. 636–646. DOI: https://doi.org/10.1038/ni.1728.

[109] Zanker D., Waithman J., Yewdell J., and Chen W. "Mixed Proteasomes Function To Increase Viral Peptide Diversity and Broaden Antiviral CD8+ T Cell Responses". In: *Journal of Immunology* 10 (2013), pp. 52–59. DOI: https://doi.org/10.4049/jimmunol.1300802.

[110] Dalet A., Stroobant V., Vigneron N., and Van den Eynde B. "Differences in the production of spliced antigenic peptides by the standard proteasome and the immunoproteasome". In: *European Journal of Immunology* 41 (2011), pp. 39–46. DOI: https://doi.org/10.1002/eji.201040750.

[111] Perez CL., Larsen MV., Gustafsson R., Norstrom MM., Atlas A, Nixon DF., and *et al*. "Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV- 1 subtypes". In: *Journal of Immunology* 180 (2008), pp. 5092–5100. DOI: https://doi.org/10.4049/jimmunol.180.7.5092.

[112] Molero-Abraham M., Lafuente EM., and Reche P. "Customized predictions of peptide-MHC binding and T-cell epitopes using EPIMHC". In: *Methods in Molecular Biology* 1184 (2014), pp. 319–332. DOI: https://doi.org/10.1007/978-1-4939-1115-8_18.

[113] Waterhouse A., Bertoni M., Bienert S., Studer G., Tauriello G., Gumienny R., and *et al*. "SWISS-MODEL: homology modelling of protein structures and complexes". In: *Nucleic Acids Research* 46 (2018), pp. 296–303. DOI: https://doi.org/10.1093/nar/gky427.

[114] Weinzierl A., Lemmel C., Schoor O., Muller M., Jruger T., Wernet D., and *et al*. "Distorted Relation between mRNA CopyNumber and Corresponding MajorHistocompatibility Complex LigandDensity on the Cell Surface". In: *Molecular and Cellular Proteomics* 6.1 (2007), pp. 102–113. DOI: https://doi.org/10.1074/mcp.M600310-MCP200.

[115] Pearson H., Daouda T., Granados D., Durette C., Bonneil E., Courcelles M., and *et al*. "MHC class I-associated peptides derive from selective regions of the human genome". In: *The Journal of Clinical Investigation* 126.12 (2016), pp. 4690–4701. DOI: https://doi.org/10.1172/JCI88590.

[116] Li L., de Batliwala M., and Bouvier M. "ERAP1 enzyme-mediated trimming and structural analyses of MHC I-bound precursor peptides yield novel insights into antigen processing and presentation". In: *Jounral of Biological Chemistry* 294.49 (2018), pp. 18534–18544. DOI: https://doi.org/10.1074/jbc.RA119.010102.

[117] Toes R., Nussbaum A., Degermann J., Schirle M., Emmerich N., Kraft M., and *et al*. "Discrete Cleavage Motifs of Constitutive and Immunoproteasomes Revealed by Quantitative Analysis of Cleavage Products". In: *The Journal of Immunology* 194.1 (2001), pp. 1–12. DOI: https://doi.org/10.1084/jem.194.1.1.

[118] Guillaume B., Chapiro J., Stroobant V., Colau D., Van Holle B., and Bousquet-Dubouch G. “Two abundant proteasome subtypes that uniquely process some antigens presented by HLA class I molecules and *et al.*” In: *PNAS* 107.43 (2010), pp. 18599–18604. DOI: https://doi.org/10.1073/pnas.1009778107.

[119] Kuckelkorn U., Stubler S., Textoris-Taube K., Killian C., Niewienda A., Henklein P., and et a. “Proteolytic dynamics of human 20S thymoproteasome”. In: *Journal of Biological Chemistry* 294.19 (2019), pp. 7740–7754. DOI: https://doi.org/10.1074/jbc.RA118.007347.

[120] Dianzani C., Domizia V., Clemente N., Chiocchetti A., Boneschi F., Galimberti D., and *et al.* “Untangling Extracellular Proteasome-Osteopontin Circuit Dynamics in Multiple Sclerosis”. In: *Cells* 8 (2019), pp. 262–269. DOI: https://doi.org/10.3390/cells8030262.

[121] Fabre B., Lambrour T., Garrigues L., Amalric F., Vingeron N, Menneteau T., and *et al.* “Deciphering preferential interactions within supramolecular protein complexes: the proteasome case”. In: *Molecular Systems Biology* 11.1 (2015), pp. 771–787. DOI: https://doi.org/10.15252/msb.20145497.

[122] Apavaloaei A., Brochu S., Dong M., Rouette A., Marie-Pierre H., Villafano G., and *et al.* “PSMB11 Orchestrates the Development of CD4 and CD8 Thymocytes via Regulation of Gene Expression in Cortical Thymic Epithelial Cells”. In: *The Journal of Immunology* 202 (2018), pp. 966–978. DOI: https://doi.org/10.1074/jbc.RA118.007347.

[123] Brinbraum M., Mendoza J., Sethi D., Dong S., Glanville J., Dobbins J., and *et al.* “Deconstructing the peptide-MHC specificity of T cell recognition”. In: *Cell* 157.5 (2014), pp. 1073–1087. DOI: https://doi.org/10.1016/j.cell.2014.03.047.

[124] Singh N., Riley T., Baker S., Borrman T., Weng Z., and Baker B. “Emerging concepts in T cell receptor specificity: rationalizing and (maybe) predicting outcomes”. In: *Journal of Immunology* 199.7 (2017), pp. 2203–2213. DOI: https://doi.org/doi:10.4049/jimmunol.1700744.

[125] Welsh R., Che J., Brehm M., and Selin L. “Heterologous immunity between viruses”. In: *Immunological Reviews* 235 (2010), pp. 244–266. DOI: https://doi.org/10.1111/j.0105-2896.2010.00897.x.

[126] Bakker A., van der Burg S., Huijbens R., Drijfhout J., Melief C. Adema G., and *et al.* “Analogues of CTL epitopes with improved MHC class-I binding capacity elicit anti-melanoma CTL recognizing the wild-type epitope”. In: *International Journal of Cancer* 70 (1997), pp. 302–309. DOI: https://doi.org/10.1002/(sici)1097-0215(19970127)70:3<302::aid-ijc10>3.0.co;2-h.

[127] Adams J., Narayanan S., Birnbaum M., Sidhu S., Blevins S., Gee M., and *et al.* “Structural interplay between germline and adaptive recognition determines TCR-peptide-MHC cross-reactivity”. In: *Nature Immunololgy* 17.1 (2016), pp. 87–94. DOI: https://doi.org/10.1038/ni.3310.

[128] Riley T., Hellman L., Gee M., Mendoza J., Alonso J., Foley K., and *et al.* "T cell receptor cross-reactivity expanded by dramatic peptide/MHC adaptability". In: *Nature Chemial Biology* 14.10 (2018), pp. 934–942. DOI: https://doi.org/10.1038/s41589-018-0130-4.

[129] Lythe G., Callard R., Hoare R., and Molina-Paris C. "How many TCR clonotypes does a body maintain?" In: *Journal of Theoretical Biology* 389 (2016), pp. 214–224. DOI: https://doi.org/10.1016/j.jtbi.2015.10.016.

[130] Greef P., Oakes T., Gerritsen B., Ismail M., Heather J., Hermsen R., Chain B., and de Boer R. "The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes". In: *eLife* 9 (2020), e49900. DOI: https://doi.org/10.7554/eLife.49900.

[131] Szomolay B., Liu J., Brown P., Miles J., Clement M., Llewellyn-Lacey S., and *et al.* "Identification of human viral protein-derived ligands recognized by individual MHCI-restricted T-cell receptors". In: *Immunology & Cell Biology* 94 (2016), pp. 573–582. DOI: https://doi.org/10.1038/icb.2016.12.

[132] Sewell A. "Why must T cells be cross-reactive?" In: *Nature Reviews. Immunology* 12 (2012), pp. 669–677. DOI: https://doi.org/10.1038/nri3279.

[133] Ishizuka J., Grebe K., Shenderov E., Peters B., Chen Q., Peng Y., and *et al.* "Quantitating T cell cross-reactivity for unrelated peptide antigens". In: *Journal of Immunology* 183 (2009), pp. 4337–4345. DOI: https://doi.org/10.4049/jimmunol.0901607.

[134] Franklid S., De Boer R., Lund O., Nielesen M., and Kesmir C. "Amino acid similarity accounts for T cell cross-reactivity and for holes in the T cell repertoire". In: *Plos ONE* 3 (2008), e1831. DOI: https://doi.org/10.1371/journal.pone.0001831.

[135] Whalley J., Dolton S., Brown PE., Wall A., Wooldrigde L., van den Berg H., and *et al.* "GPU-accelerated discovery of pathogen-derived molecular mimics of a T-cell insulin epitope". In: *Frontiers in Immunology* 11 (2020), p. 296. DOI: https://doi.org/10.3389/fimmu.2020.00296.

[136] Cole D., Bulek A., Dolton G., Schauenberg A., Szomolay B., Rittase W., and *et al.* "Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity". In: *The Journal of Clinical Investigation* 126.6 (2016), pp. 2191–2204. DOI: https://doi.org/10.1172/JCI85679.

[137] Rossjohn J., Gras S., Miles JJ., Turner SJ., Godfrey DI., and McCluskey J. "T cell antigen receptor recognition of antigen-presenting molecules". In: *Annual Review of Immunology* 33 (2015), pp. 169–200. DOI: https://doi.org/10.1146/annurev-immunol-032414-112334.

[138] Berman H., Westbrook J., Feng Z., Gilliland G., Bhat T., Weiising H., and *et al.* "The Protein Data Bank". In: *Nucleic Acid Research* 28 (2000), pp. 235–242. DOI: https://doi.org/10.1093/nar/28.1.235.

[139] Kim Y., Sidney J., Pinilla C., Sette A., and Peters B. “Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior”. In: *BMC Bioinfomratics* 10.394 (2009). DOI: https://doi.org/10.1186/1471-2105-10-394.

[140] Assmus LM., Guan J., Wu T., Farenc C., Sing XYX., Zareie P., and *et al.* “Overlapping peptides elicit distinct CD8+ T cell responses following influenza a virus infection”. In: *Journal of Immunology* 205 (2020), pp. 1731–1742. DOI: https://doi.org/10.4049/jimmunol.2000689.

[141] Gonzalez-Duque S., Azoury M., Coili M., Afonso G., Georgia A., Turatsinze J., Nigi L., and *et al.* “Conventional and Neo-antigenic Peptides Presented by Cells Are Targeted by Circulating Naïve CD8+ T Cells in Type 1 Diabetic and Healthy Donors”. In: *Cell* 28.6 (2018), pp. 946–960. DOI: https://doi.org/10.1016/j.cmet.2018.07.007.

[142] Wan X., Vomund A., Peterson O., Chervonsky A., Lichti C., and Unanue E. “The MHC-II peptidome of pancreatic islets identifies key features of autoimmune peptides”. In: *Nature* 21 (2020), pp. 455–463. DOI: https://doi.org/10.1038/s41590-020-0623-7.

[143] Delong T., Wiles T. A., Baker R. L., Bradley B., Barbour G., Reisdorph R., and *et al.* “Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion”. In: *Science* 351 (2016), pp. 711–714. DOI: https://doi.org/10.1126/science.aad2791.

[144] Baker R., M. Rihanek, Hohenstein A., Nakayama M., Michels A., Gottlieb P., and *et al.* “Hybrid Insulin Peptides Are Autoantigens in Type 1 Diabetes”. In: *Diabetes* 68.9 (2019), pp. 1830–1840. DOI: https://doi.org/10.2337/db19-0128.

[145] Layton-Arribas D., Guyer D., Delong T., Dang M., Chow T., Speake C., and *et al.* “Hybrid Insulin Peptides Are Recognized by Human T Cells in the Context of DRB1*04:01”. In: *Diabetes* 69.7 (2020), pp. 1492–1502. DOI: https://doi.org/10.2337/db19-0620.

[146] Babon J., DeNicola M., Blodgett D., Crevecoeur I., Buttrick T., Maehr R., and *et al.* “Analysis of self-antigen specificity of islet-infiltrating T cells from human donors with type 1 diabetes”. In: *Nature Medicine* 22.12 (2016), pp. 1482–1487. DOI: https://doi.org/10.1038/nm.4203.

[147] Wang Y., Sosinowski T., Novikov A., Crawford F., White J., Jin N., and *et al.* “How C-terminal additions to insulin B-chain fragments create superagonists for T cells in mouse and human type 1 diabetes”. In: *Science* 4.34 (2019), pp. 1–12. DOI: https://doi.org/10.1126/sciimmunol.aav7517.

[148] Dotta F., Censini S., van Halteren A., Marselli L., Masini M., Dionisi S., and *et al.* “Coxsackie B4 virus infection of beta cells and natural killer cell insulitis in recent-onset type 1 diabetic patients”. In: *PNAS* 104 (2007), pp. 5115–5120. DOI: https://doi.org/10.1073/pnas.0700442104.

[149] Variela-Calvino R., Skowera A., Arif S., and Peakman M. “Identification of a naturally processed cytotoxic CD8 T-cell epitope of coxsackievirus B4, presented by HLA-A2.1 and located in the PEVKEK region of the P2C nonstructural protein”. In: *Journal of Virology* 78 (2004), pp. 13399–13408. DOI: https://doi.org/10.1128/JVI.78.24.13399-13408.2004.

[150] Honeyman M., Stone N., and Harrison L. “T-cell epitopes in type 1 diabetes autoantigen tyrosine phosphatase IA-2: potential for mimicry with rotavirus and other environmental agents”. In: *Molecular Medicine* 4 (1998), pp. 231–239.

[151] Jones D. and Crosby I. “Proliferative lymphocyte responses to virus antigens homologous to GAD65 in IDDM”. In: *Diabetologia* 39 (1996), pp. 1318–1324. DOI: https://doi.org/10.1007/s001250050576.

[152] Kolehmainen P., Koskiniemi M., Oikarinen S., Veijola R., Simell O., Lionen J., and *et al*. “Human parechovirus and the risk of type 1 diabetes”. In: *Journal of Medical Virology* 85 (2013), pp. 1619–1623. DOI: https://doi.org/10.1002/jmv.23659.

[153] Pak C., Eun H., McArthur R., and Yoon J. “Association of cytomegalovirus infection with autoimmune type 1 diabetes”. In: *Lancet* 2 (1988), pp. 1–4. DOI: https://doi.org/10.1016/s0140-6736(88)92941-8.

[154] Yoneda S., Imagawa A., Fukui K., Uno S., Kozawa J., Sakai M., and *et al*. “A Histological Study of Fulminant Type 1 Diabetes Mellitus Related to Human Cytomegalovirus Reactivation”. In: *Journal of Clinical Endocrinology & Metabolism* 102 (2017), pp. 2394–2400. DOI: https://doi.org/10.1210/jc.2016-4029.

[155] Rodriguez-Calvo T., Krogvold L., Amirian N., Dahl-Jogensen K., and von Herrath M. “One in Ten CD8+ Cells in the Pancreas of Living Individuals With Recent-Onset Type 1 Diabetes Recognizes the Preproinsulin Epitope PPI15-24”. In: *Diabetes* 70 (2021), pp. 752–758. DOI: https://doi.org/10.2337/db20-0908.

[156] Aarnisalo J., Veijola R., Vainionpaa R., Simell O., Knip M., and Ilonene J. “Cytomegalovirus infection in early infancy: risk of induction and 1232 progression of autoimmunity associated with type 1 diabetes”. In: *Diabetologia* 51.5 (2008), pp. 769–772. DOI: https://doi.org/10.1007/s00125-008-0945-8.

[157] Ekman I., Vuorinen R., Knip M., Veijola R., Toppari J., Hyoty H., Kinnunen T., and *et al*. “Early childhood CMV infection may decelerate the progression to clinical type 1 diabetes”. In: *paediatric diabetes* 20.1 (2018), pp. 73–77. DOI: https://doi.org/10.1111/pedi.12788.

[158] Bian X., Wallstrom G., Davis A., Wang J., Park J., Throop A., and *et al*. “Immunoproteomic Profiling of Antiviral Antibodies in New-Onset Type 1 Diabetes Using Protein Arrays”. In: *Diabetes* 65.1 (2015), pp. 285–296. DOI: https://doi.org/10.2337/db15-0179.

[159] Sabouri S., Benkahla M., Kiosses W., Rodriguez-Calvo T., Zapardiel-Gonzalo J., Castillo E., and *et al*. “Human herpesvirus-6 is present at higher levels in the pancreatic tissues of donors with type 1 diabetes”. In: *Journal of Autoimmunity* 107 (2020), p. 102378. DOI: https://doi.org/10.1016/j.jaut.2019.102378.

[160] Vita R., Mahajan S., Overton J., Dhanda S., Martini S., Cantrell J., and *et al*. “The Immune Epitope Database (IEDB)”. In: *Nucleic Acid Research* 47 (2018), pp. 339–343. DOI: https://doi.org/10.1093/nar/gky1006.

[161] Jurtz V., Paul S., Andreatta M., Marcatili P., Peters B., and Nielsen M. “NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity”. In: *Journal of Immunology* 199 (2017), pp. 3360–3368. DOI: https://doi.org/10.4049/jimmunol.1700893.

[162] Yang J., Yan R., Roy A., Xu J., Poisson J., and Zhang Y. “The I-TASSER Suite: protein structure and function prediction”. In: *Nature Methods* 12.1 (2015), pp. 7–8. DOI: https://doi.org/10.1038/nmeth.3213.

[163] Ouyang Q., Standifer N., Qin H., Gottlieb P., Verchere C., Nepom G., and *et al*. “Recognition of HLA class I-restricted beta-cell epitopes in type 1 diabetes”. In: *Diabetes* 12 (2006), pp. 3068–3074. DOI: https://doi.org/10.2337/db06-0065.

[164] Sidney J., Vela J., Friedrich D., Kolla R., von Herrath M., Wesley J., and *et al*. “Low HLA binding of diabetes-associated CD8+ T-cell epitopes is increased by post translational modifications”. In: *BMC Immunology* 19 (2018), p. 12. DOI: https://doi.org/10.1186/s12865-018-0250-3.

[165] Paes W., Leonov G., Partridge T., Nicastri A., Ternette N., and Borrow P. “Elucidation of the Signatures of Proteasome-Catalyzed Peptide Splicing”. In: *Frontiers in Immunology* 11 (2020), p. 563800. DOI: https://doi.org/10.3389/fimmu.2020.563800.

[166] Coli M., Ramos-Rodriguez M., Nakayasu E., Alvelos M., Lopes M., Hill J., and *et al*. “An integrated multi-omics approach identifies the landscape of interferon-alpha-mediated responses of human pancreatic beta cells”. In: *Nature Communictaions* 11 (2020), p. 2584. DOI: https://doi.org/10.1038/s41467-020-16327-0.

[167] Apaolaza P., Balcacean D., Zapardiel-Gonzalo J., Nelson G., Lenchik N., Akhbari P., and *et al*. “Islet expression of type I interferon response sensors is associated with immune infiltration and viral infection in type 1 diabetes”. In: *Science Advances* 7.9 (2021), eabd6527. DOI: https://doi.org/10.1126/sciadv.abd6527.

[168] Tenzer S., Wee E., Burgevin A., Stewart-Jones G., Friis L, Lamberth K., and *et al*. “Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance”. In: *nature Immunology* 10 (2009), pp. 636646. DOI: https://doi.org/10.1038/ni.1728.

[169] Val M. Schlicht H., Ruppert M., Reddehase M., Koszinowski U., and *et al.* “Efficient processing of an antigenic sequence for presentation by MHC class I molecules depends on its neighboring residues in the protein”. In: *Cell* 66.6 (1991), pp. 1145–1153. DOI: https://doi.org/10.1016/0092-8674(91)90037-y.

[170] Eggers M., Boes-Fabian B., Ruppert T., Kloetzel P., Koszinowski U., and *et al.* “The cleavage preference of the proteasome governs the yield of antigenic peptides”. In: *Journal of Experimental Medicine* 182.6 (1995), pp. 1865–1870. DOI: https://doi.org/10.1084/jem.182.6.1865.

[171] Velders M., Weijzen S., Eiben G., Elimshad A., Kloetzel P., Higgins T., and *et al.* “Defined flanking spacers and enhanced proteolysis is essential for eradication of established tumors by an epitope string DNA vaccine”. In: *Journal of Immunology* 166.9 (2001), pp. 5366–5373. DOI: https://doi.org/10.4049/jimmunol.166.9.5366.

[172] Theobald M., Ruppert T., Kuckelkorn U., Henrandez J., Haussler A. Ferreira E., and *et al.* “The Sequence Alteration Associated with a Mutational Hotspot in p53 Protects Cells From Lysis by Cytotoxic T Lymphocytes Specific for a Flanking Peptide Epitope”. In: *Journal of Experimental Medicine* 188.6 (1998), pp. 1017–1028. DOI: https://doi.org/10.1084/jem.188.6.1017.

[173] Beekman N., van Veelen P., van Hall T., Neisig A., Sijts A., Camps M., and *et al.* “Abrogation of CTL epitope processing by single amino acid substitution flanking the C-terminal proteasome cleavage site”. In: *Journal of Immunology* 164.4 (2000), pp. 1898–1905. DOI: https://doi.org/10.4049/jimmunol.164.4.1898.

[174] Seifert U., Liermann H., Racanelli V., Halenius A., Wiese M., Wedemeyer H., and *et al.* “Hepatitis C virus mutation affects proteasomal epitope processing”. In: *The Journal of Clinical Investigation* 114.2 (2004), pp. 250–259. DOI: https://doi.org/10.1172/JCI20985.

[175] Ossendrop F., Egger M., Neisig A., Ruppert T., Groettrup M., Sijts A., and *et al.* “A single residue exchange within a viral CTL epitope alters proteasome-mediated degradation resulting in lack of antigen presentation”. In: *Immunity* 5.2 (1996), pp. 115–124. DOI: https://doi.org/10.1016/s1074-7613(00)80488-4.

[176] Calvert A., Chalastains A., Wu Y., Hurley L., Kouri F., Bi Y., and *et al.* “Cancer-associated IDH1 promotes growth and resistance to targeted therapies in the absence of mutation”. In: *Cell Reports* 19.9 (2017), pp. 1858–1873. DOI: https://doi.org/10.1016/j.celrep.2017.05.014.

[177] Marranci A., Jiang Z., Vitiello M., Guzzolino E., Comelli L., Sarti S., and *et al.* “The landscape of BRAF transcript and protein variants in human cancer”. In: *Molecular Cancer* 16.1 (2017), p. 85. DOI: https://doi.org/10.1186/s12943-017-0645-4.

[178] Hu Y., Tao S., Deng J., Hou Z., Liang J., Huang Q., and *et al.* "Prognostic Value of NRAS Gene for Survival of Colorectal Cancer Patients: A Systematic Review and Meta-Analysis". In: *Asian Pacific Journal of Cancer Prevention* 19.11 (2018), pp. 3001–3008. DOI: https://doi.org/10.31557/APJCP.2018.19.11.3001.

[179] Rozovski U., Manshouri T., Dembitz V., Bozinovic K., Pierce S., Kantarjian H., and *et al.* "JAK2, Calr and MPL Mutation Status Predicts the Survival Outcome of Patients with Primary Myelofibrosis". In: *Blood* 124.21 (2014), p. 1829. DOI: https://doi.org/10.1182/blood.V124.21.1829.1829.