Advancing immunopeptidomics: validation of the method, improved epitope prediction, peptide-based HLA typing and discrimination of healthy and malignant tissue

Weiterentwicklung der Immunpeptidomik: Validierung der Methode, Verbesserung der Epitopvorhersage, peptidbasierte HLA-Typisierung und Unterscheidung von gesundem und bösartigem Gewebe

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M. Sc. Technische Biologie Michael Ghosh

aus Kempten

Tübingen

2020

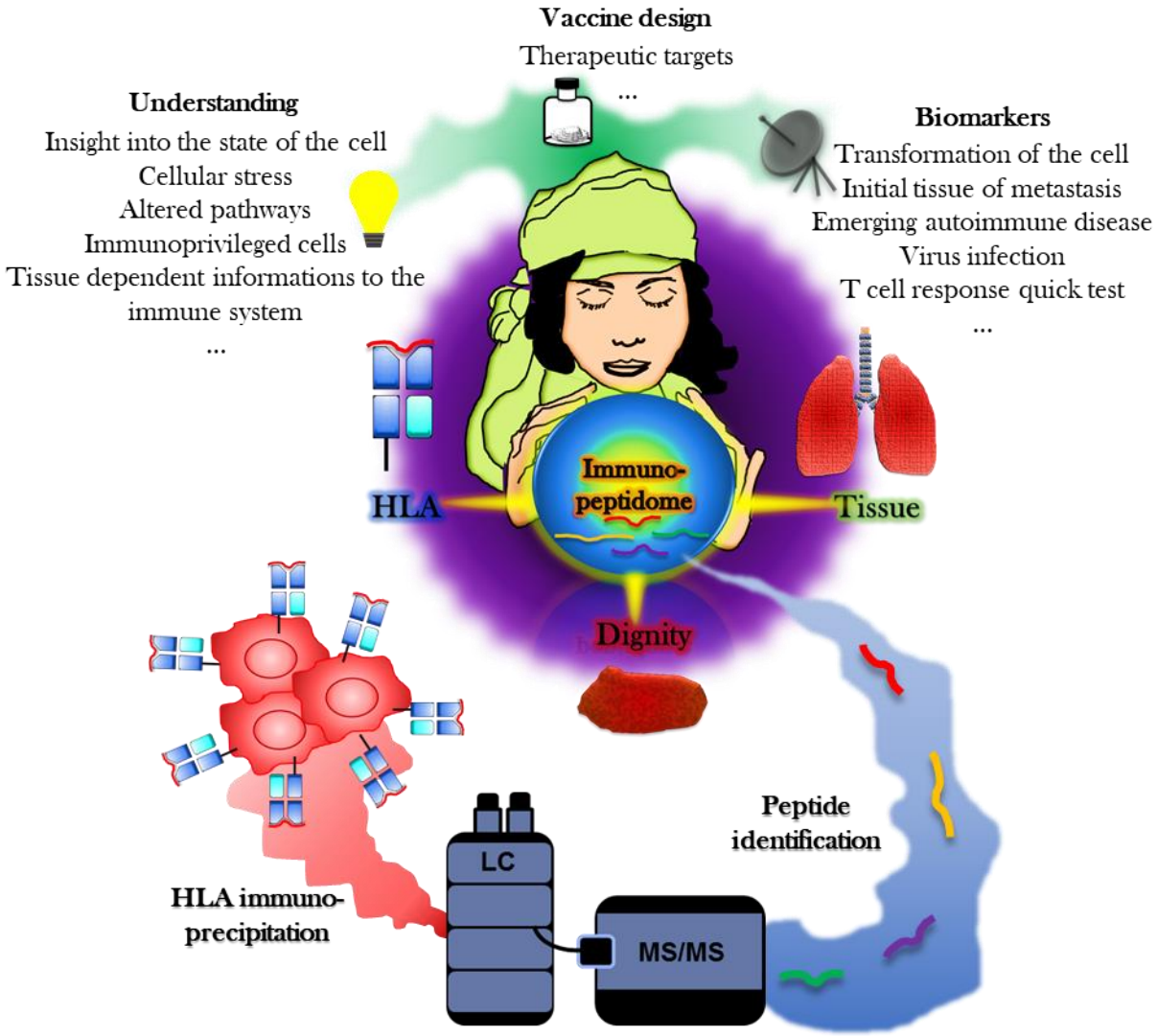
Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	30.07.2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Stefan Stevanović
2. Berichterstatter:	Prof. Dr. Hans-Georg Rammensee

Contents

1 Abstract

1.1 Graphical Abstract



1.2 Summary

For almost 30 years now, the immunopeptidome has been analyzed by eluting peptides from HLA molecules. This method has already been established in several institutes and companies worldwide and is now used for a wide range of investigations from the simple identification of HLA peptide motifs for different organisms to the detection of cryptic disease-specific peptides. The field of immunopeptidomics is more popular than ever as drug development has focused on the positive modulation of the immune system in recent years. Since the approval of the first checkpoint antibodies, the era of immunotherapy has been running and specific immunotherapies with fewer side effects are in the focus. There is a wide range of applications, yet, the immunopeptidome still contains a great wealth of information waiting to be deciphered. Currently, immunopeptidomics is limited in the identification of the large number of peptides with different affinities and stabilities of the peptide-HLA complexes. Therefore, amongst many other factors, only limited recovery rates are possible. When this doctoral thesis started, there were several unresolved questions in the field of immunopeptidomics that should be approached in this thesis:

Is it possible to validate immunopeptidomics and use it reliably for clinical studies and drug development? Is there nowadays a reliable method to identify the peptide motif for peptide presenting MHC class I allotypes, the cornerstone for epitope predictions or active substance identification? Is it possible to use peptides to classify HLA allotypes or differentiate between healthy and malignant tissue? Can tumor-specific peptides be reliably characterized with this omic technology?

In this doctoral thesis the immunopeptidomic method was validated to ensure the reliability of LC-MS/MS peptide identification and all required parameters of the European Medicines Agency (EMA) and Food and Drug Administration (FDA) were investigated. In addition, an updated protocol for the identification of MHC ligands, deconvolution of peptide motifs and generation of matrices for epitope prediction was established, which can be used for monoallelic cells as well as multiallelic tissue. Finally, a method was developed to identify allotypic peptides that allow HLA typing. These peptides can also be used as an internal standard for semi-quantitative investigation of the tumor specificity of peptides. The developed method was also successfully implemented to identify tissue and dignity specific patterns in the immunopeptidome and to determine the dignity of immunopeptidomic samples.

1.3 Zusammenfassung

Seit fast 30 Jahren wird das Immunpeptidom durch Elution von Peptiden aus HLA-Molekülen analysiert. Weltweit nutzen mittlerweile mehrere Institute und Unternehmen diese Methode für ein breites Spektrum an Untersuchungen, die von der simplen Identifizierung von HLA-Peptidmotiven für verschiedene Organismen bis hin zum Nachweis kryptischer krankheitsspezifischer Peptide reichen. Die Immunpeptidomik ist populärer denn je, seit sich die Medikamentenentwicklung in den letzten Jahren auf die positive Modulation des Immunsystems fokussiert hat. Die Zulassung der ersten Checkpoint-Antikörper leitete die Ära der Immuntherapie ein und spezifische Immuntherapien mit weniger Nebenwirkungen stehen nun im Blickpunkt. Das Anwendungsspektrum der Immunpeptidomik ist mittlerweile breit gefächert, dennoch enthält das Immunpeptidom immer noch eine große Fülle von Informationen, die darauf warten, entschlüsselt zu werden. Aktuell ist die Immunpeptidomik darin eingeschränkt, dass die große Anzahl von Peptiden, mit unterschiedlichen Affinitäten und Stabilitäten der Peptid-HLA-Komplexe, nicht optimal erfasst werden kann und daher unter anderem nur begrenzte Wiederfindungsraten möglich sind. Zu Beginn dieser Doktorarbeit gab es ungelöste Fragestellungen auf dem Gebiet der Immunpeptidomik, die in dieser Arbeit untersucht werden sollten:

Ist es möglich, die Immunpeptidomik zu validieren und diese zuverlässig für klinische Studien und die Medikamentenentwicklung einzusetzen? Gibt es heute eine zuverlässige Methode zur Identifizierung von Peptidmotiven für Peptid-präsentierende MHC-Klasse-I-Allotypen, dem Grundstein für Epitopvorhersagen und Wirkstoffidentifizierungen? Ist es möglich, Peptide zur Klassifizierung von HLA-Allotypen oder zur Unterscheidung zwischen gesundem und bösartigem Gewebe zu verwenden? Können tumorspezifische Peptide mit dieser Omik-Technologie zuverlässig charakterisiert werden?

In dieser Doktorarbeit wurde die immunpeptidomische Methode validiert, um die Zuverlässigkeit der LC-MS/MS-Peptid-Identifizierung zu gewährleisten, und es wurden alle erforderlichen Parameter der Europäischen Arzneimittel-Agentur und U. S. Food and Drug Administration untersucht. Darüber hinaus wurde ein aktualisiertes Protokoll für die Identifizierung von MHC-Liganden, die Entschlüsselung von Peptidmotiven und die Generierung von Matrizen für die Epitopvorhersage erstellt, das sowohl für monoallele Zellen als auch für multiallele Gewebe verwendet werden kann. Schließlich wurde eine Methode entwickelt, um allotypische Peptide zu identifizieren, die eine HLA-Typisierung ermöglichen. Diese Peptide können auch als interner Standard für die semi-quantitative Untersuchung der Tumorspezifität von Peptiden verwendet werden. Diese Methode wurde erfolgreich implementiert, um gewebe- und dignitätsspezifische Muster im Immunpeptidom zu identifizieren und die Dignität von immunpeptidomischen Proben zu bestimmen.

2 Introduction

2.1 A short trip through the immune system

There are numerous different cells and molecular structures in the innate and the adaptive immune system, protecting the organism against its own abnormal cells, foreign structures, and pathogens [1]. The first defense is provided by the cellular components of the innate immune system such as dendritic cells (DCs) and natural killer cells (NK cells), but also humoral components such as the complement system play a role [2,3]. The macrophages and DCs phagocyte foreign bodies by binding with their pattern recognition receptors (PRRs) to pathogen-associated molecular patterns (PAMPs), common foreign structures, or damage-associated molecular patterns (DAMPs), cell compounds released during cell damage and death [4,5]. In addition, these phagocyting cells also function as antigen-presenting cells (APCs) presenting protein fragments, termed peptides, from phagocyted and digested foreign bodies on their cell surface using human leukocyte antigen (HLA) molecules. This mechanism enables a connection of the innate and the adaptive immune system [6].

In contrast to the innate immune system, which ensures a rapid reaction and identification of harmful antigens [7], the adaptive immune system is highly specific [8]. It enables effective and specific control of pathogens by its targeted response against certain antigens [9]. The cellular components of the adaptive immune system are the lymphocytes, which can mainly be divided into T and B cells. The main groups of T cells are the cluster of differentiation (CD)8^+ cytotoxic T lymphocytes (CTLs), the $CD4^+$ T helper cells (T_H cells), as well as the regulatory T cells (T_{reg} cells). A characteristic of T cells is the pre-selection of antigen-specific T cells in the thymus. The avoidance of self-antigen specific T cells prevents T cell reactivity against self-antigens and provides central tolerance [10].

Naïve T cells having survived thymus selection can be activated by APCs which present their specific antigen [11]. As illustrated in Figure 1 three signals ensure T cell activation. In the first signal, the antigen-specific T cell receptor (TCR) binds the corresponding HLA-peptide complex, where the peptide is the specific antigen [12]. This binding is additionally stabilized by co-receptors. In case of CTLs, CD8 binds the α_3-domain of the HLA class I molecules or in case of T_H cells, CD4 binds the β_2-domain of HLA class II molecules [13]. The second signal is a costimulatory signal generated by interaction of the protein CD28 (T cell) and the CD80/CD86 complex (APC). A missing costimulatory signal might result in T cell anergy or peripheral tolerance, which, in addition to thymic selection, acts as a second protective measure to prevent T cell reactions against self-antigens [14]. The third signal is mediated by cytokines which are secreted from APCs and T cells and polarizes the T cell to an effector phenotype based on the cytokine milieu [11,15–17].

The immune response of the adaptive immune system is diverse and antigen-specific, which is amongst others achieved by recombination and specification of TCRs [18] against antigens, but also delays the immune response by several days. Another peculiarity of the adaptive immune system is the memory of already opposed antigens against which quickly T memory cells are reactivatable, permitting an earlier immune response [19,20]. The interaction of the innate and adaptive immune system enables an effective defense against pathogens.

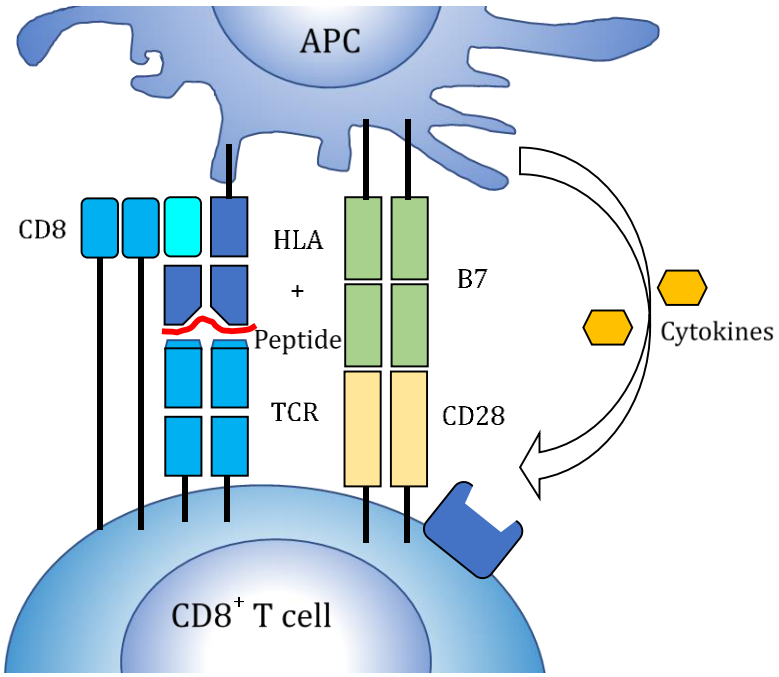


Figure 1: Activation of T cells via APCs. *First, a naïve T cell recognizes a foreign peptide-HLA (pHLA) complex from an APC. Second, the T cell receives the costimulatory signal resulting in the survival and proliferation of the T cell (*interaction of CD28 and the CD80/CD86 complex (B7)*. Third, the cytokine environment secreted by the APC polarizes the T cell to an effector phenotype based on the cytokine milieu* [21]*.*

2.2 The human leukocyte antigen molecule

A central element to enable T cells to recognize antigens and activate the highly specific adaptive immune system is the HLA molecule, a glycosylated membrane protein [22]. HLA molecules are divided into HLA class I and HLA class II and differ in structure, protein source and expression [23]. As this doctoral thesis is concerned with HLA class I presented peptides, this chapter will mainly focus on HLA class I.

There are three classical molecules for HLA class I, HLA-A, B and C, whose genes are expressed in a highly polymorphic and codominant way. Besides the classical HLA alleles, HLA-E, F and G are coded for class I.

The classical HLA molecules, which are examined in Chapter 4 and 5, are located on the cell surface of nucleated cells presenting peptides from intracellular proteins to CTLs. The compilation of different HLA presented peptides is described as the immunopeptidome [24,25]. The HLA class I peptide complex consists of a presented peptide, a heavy α chain (~ 43 kDa) composed of the α_1-α_3-domains and the non-covalently bound invariant light β_2-microglobulin (β_2m) [26]. The peptide is embedded in the peptide binding groove formed by two α-helices and an antiparallel β-strand between the α1- and α2-domains. The peptide binding groove is enclosed by the curved α-helices and large aromatic residues and limits the peptide length. The majority of HLA class I peptides consists of 8-12 amino acids (aa) [27]. The pHLA bond is generated due to interactions with the peptide terminus and with aa residues as illustrated in Figure 2 [28–30]. There are six pockets A-F which interact with the peptide, wherein pocket B and F, with their interactions with peptide residues located at position 2 and the C-terminus, most constrain the peptide bond. These positions in the peptide that are most restricted are called anchor positions [31,32]. The aa positions in the binding groove are highly polymorphic and thus lead to different presented peptides by the HLA allotypes [33,34].

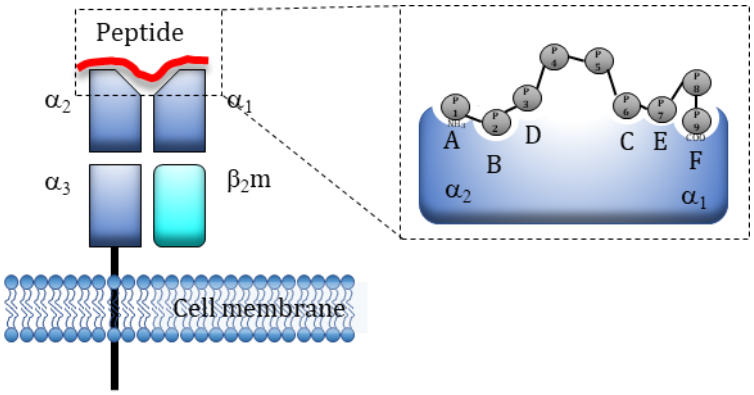


Figure 2: HLA class I peptide complex. *An exemplary nonameric peptide is shown in the groove with the pockets A-F binding to the N-terminus, position 2, 3, 6, 7 and the C-terminus* [35].

HLA class II molecules are mainly found on APCs where they present $CD4^+$ T cells peptides of majorly extracellular proteins. Compared to HLA class I, the anchor positions are less defined, resulting in a high promiscuity between HLA class II allotypes. The current state of known HLA alleles is 19.031 for class I (ebi.ac.uk/ipd/imgt/hla/stats.html, date: February, 2020).

2.3 Antigen processing

The immunopeptidome presented via HLA class I molecules is mainly formed by the large protein diversity in the cell's proteome, which is not only formed by the different genes in the genome, but also by different transcription and translation by various messenger ribonucleic acid (mRNA) and protein isoforms [36–40]. During protein synthesis in the cell, in addition to the correctly folded

and functional proteins a large number of misfolded proteins, so-called defective ribosomal products are formed and account for about 10% of all proteasomally degraded proteins [41–43].

The emergence of the HLA class I presented peptide begins with protein degradation (Figure 3). Proteins are labelled by addition of ubiquitin molecules [44] and are proteolytically degraded in the 26S proteasome, a macromolecule consisting of a 20S subunit (nucleus) and two 19S subunits (caps), in a cavity formed by four rings containing seven subunits ($\alpha_7\beta_7\beta_7\alpha_7$). In the ring structures of the core, the subunits β_1, β_2 and β_5 have proteolytic activity. The β_1 subunit prefers C-terminal acidic residues, the β_2 subunit has a tryptic-like specificity and prefers basic residues and the β_5 subunit has chymotryptic-like specificity and cleaves after C-terminal hydrophobic residues. The linear peptides resulting from protein degradation have a length of 4-25 aa, with an average length of 7-9 aa [45]. In the caps of the 26S proteasome there are two protein groups, the Rpn proteins, which recognize ubiquitin-labeled proteins, and the Rpt proteins, which unfold these proteins in an ATP-dependent manner [46].

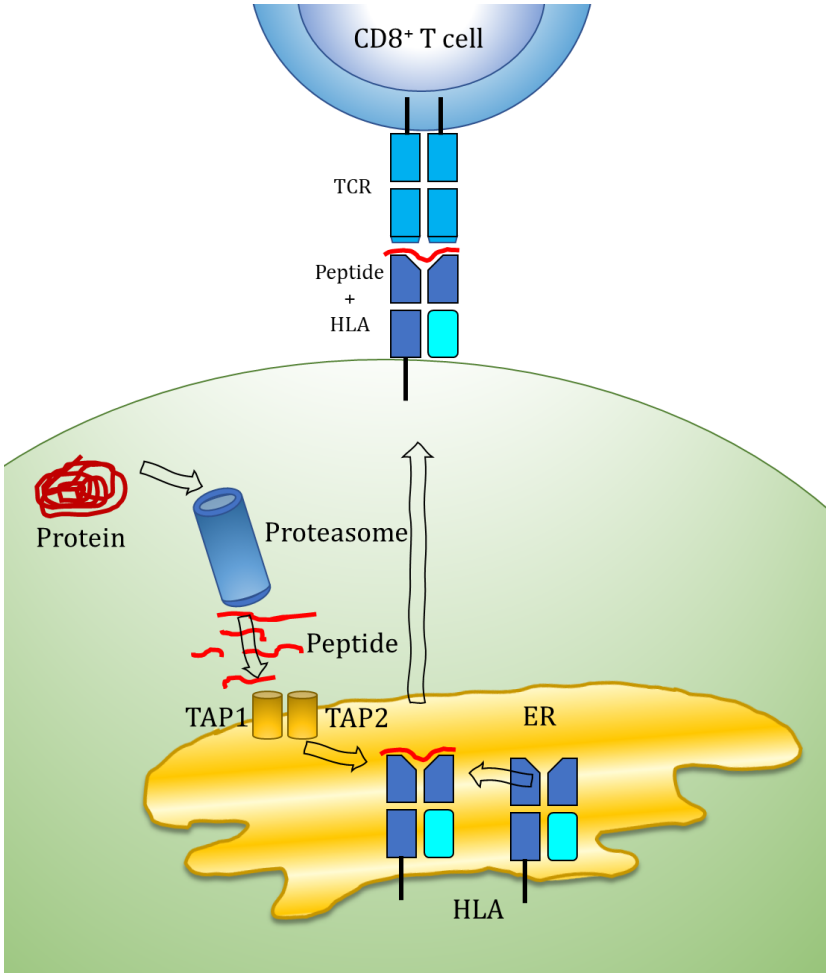


Figure 3: HLA class I antigen processing pathway*. Antigen processing commences with the degradation of intracellular proteins (antigens), via the proteasome and other proteases to peptides. The peptides are transferred to the ER via TAP and then loaded on chaperone stabilized HLA molecules. Finally, the complexes get transported on the cell membrane by the secretory pathway [47].*

Interferon-γ (IFN-γ) secretion by NK cells or CTLs in an immune response leads to the exchange of the proteolytic subunits β_1, β_2 and β_5 by β_1i, β_2i and β_5i and thus to the formation of the immunoproteasome [46,48]. This exchange results in altered proteolytic activity, whereby increased tryptic- and chymotryptic-like specificity facilitates the formation of peptides having hydrophobic and basic residues at the C-terminus [49,50]. Furthermore, the IFN-γ secretion stimulates the peptide generation by the PA28 subunit [51,52]. The peptides formed in the proteasome are also truncated by cytosolic peptidases or in the endoplasmic reticulum (ER), where N-terminal trimming of the peptides takes place [23]. Overall, there is a modulation of the immunopeptidome in both HLA class I and II after IFNγ stimulation [53,54].

The peptides are loaded onto the HLA molecules in the ER. Preferably peptides having hydrophobic or basic residues at the C-terminus and 9 to 12 aa are transported ATP-dependently into the ER by the heterodimeric TAP1/TAP2 complex (transporter associated with antigen processing). The peptide loading complex (PLC) transfers peptides onto HLA molecules. It consists of the HLA class I stabilizing proteins calreticulin, tapasin and Erp57. The finished pHLA is transported to the plasma membrane via the secretory pathway [55]. Ultimately, only one peptide in a thousand is likely to be presented on a HLA molecule after processing by the proteasome, proteases in the cytosol and ER, the restrictions of the TAP complex and the peptide motif of HLA allotypes on peptide length and C-terminal aa [56–58].

As described above, HLA class I molecules mainly present peptides derived from intracellular cytosolic proteins and HLA class II molecules peptides of extracellular proteins from the cell's environment. An exception of this division by peptide origin is presented by cross presentation. In this special case peptides from extracellular proteins can also bind to HLA class I molecules [59,60]. Peptides from intracellular proteins can be presented on HLA class II molecules by the processing pathway called autophagy [61].

2.4 The immunopeptidome and peptide motifs

The immunopeptidome is the entirety of peptides presenting intra- and extracellular processes to T cells. Besides the diversity of degraded proteins, which is influenced by intrinsic, extrinsic, physiological and pathological factors, the expressed HLA allotypes also determine the peptide repertoire [62]. Depending on the cell type, there are about 100,000 HLA molecules on the cell surface, which present one to 10,000 peptide copies per cell [56]. The immunopeptidome is considerably modulated after tumor transformation or virus infection by altered cellular transcription, metabolic pathways or metabolism [62–64].

HLA molecules influence peptide diversity through their high polymorphism. The molecules have distinctive peptide binding specificities, in particular through their different aa residues within the binding groove. This creates distinctive peptide repertoires, which can be represented in

peptide motifs. The motifs summarize the meaning of a position (in bits), the aa preferences in a particular aa position in the peptide (by size of the aa) [65] and non-covalent forces in the pHLA complex, which is described in detail in Chapter 4. There are three well established approaches to determine binding preferences of HLA molecules. First, *in vitro* binding experiments using synthetic peptides from mostly publicly available databases such as SYFPEITHI [66] or IEDB [67], second, direct peptide elution of HLA molecules on cell surfaces by HLA immunoprecipitation and subsequent peptide identification by liquid chromatographic tandem mass spectrometry (LC-MS/MS), also known as immunopeptidomics, and third, the *in silico* prediction of possible peptides based on already known peptides and the structural surrounding of the binding groove [68].

The second approach is still the only approach to investigate the entirety of HLA-presented peptides and, compared to the first approach, not only considers which peptides have an affinity to the HLA binding pocket, but also all previous antigen processing influences. In this thesis, the third approach, *in silico* prediction, will be addressed in Chapter 5, whereas in Chapters 3 and 4 the focus is mainly on the second approach, immunoprecipitation with subsequent LC-MS/MS analysis.

The actual knowledge on binding specificities of HLA class I molecules has been integrated into various prediction tools. Besides easy-to-use widely-known tools such as SYFPEITHI [66] and NetMHCpan/II [69,70] the constant emerge of novel binding prediction tools such as mixMHCpred [71], mixMHC2pred [72] and NNAlign_MA [73], underlines the necessity of reliable *in silico* prediction. The peptide prediction for HLA class I alleles is well established and in case of HLA class II there have recently been groundbreaking advances [72,73].

2.5 Key points in carcinogenesis and immune defense

Chapter 5 of the doctoral thesis deals with cancer in general and therefore this chapter provides a brief insight into the development and treatment of cancer but does not discuss specific types of cancer in detail.

The interplay between the carcinogenesis and the immune defense has been known for half a century [74,75] and is described by the cancer immunoediting hypothesis. It is a dynamic process between tumor cells and the immune system, which is divided into three phases, elimination, equilibrium and escape [76]. Through the mechanism of immunoselection, tumor cells try to circumvent clearance by immune cells by generation of tumor cell variants, which are less immunogenic. This leads to a state of equilibrium between tumor cells and the immune system, so the latter must continuously adapt its defense to the new tumor cell variants, which can extend over many years [77]. As soon as the tumor cells break through the immune defense (immunosubversion), the last phase of immunoediting occurs, in which the tumor cells escape.

A transformation from a normal cell to a tumor cell occurs step by step through successive genetic changes obtained in time [78]. Major steps are certain critical mutations in oncogenes, suppressor genes, and DNA repair genes, so-called hallmarks of cancer. These hallmarks must take place before unregulated cell division occurs and the cell ceases to be able to respond to signals like normal cells [79]. The tumors can remain localized as benign tumors or spread further as malignant tumors and metastasize to distant tissues [80,81]. Usually all these cellular changes do not occur unobserved. Tumor transformation is manifested in the immunopeptidome and leads to the presentation of tumor-specific antigens: antigens overexpressed in the tumor, tissue-specific antigens, differential post-translationally modified antigens, tumor-exclusive antigens such as cancer testis antigens, oncofetal antigens, oncoviral antigens or mutated and cryptic antigens, also known as neoantigens [63,82–88]. Ideally, a presentation of tumor specific antigens and T cell recognition leads to the recruitment of the immune system and subsequent elimination of the transformed cells as illustrated in Figure 4. According to the cancer-immunity cycle, tumor specific antigens are released by killed or dying tumor cells. Next, APCs take up the antigens, prime and activate T cells in the proximate lymph nodes and subsequently, these T cells migrate to the tumor cells. They recognize and kill the tumor cells and in turn lead to further release of tumor specific antigens [89].

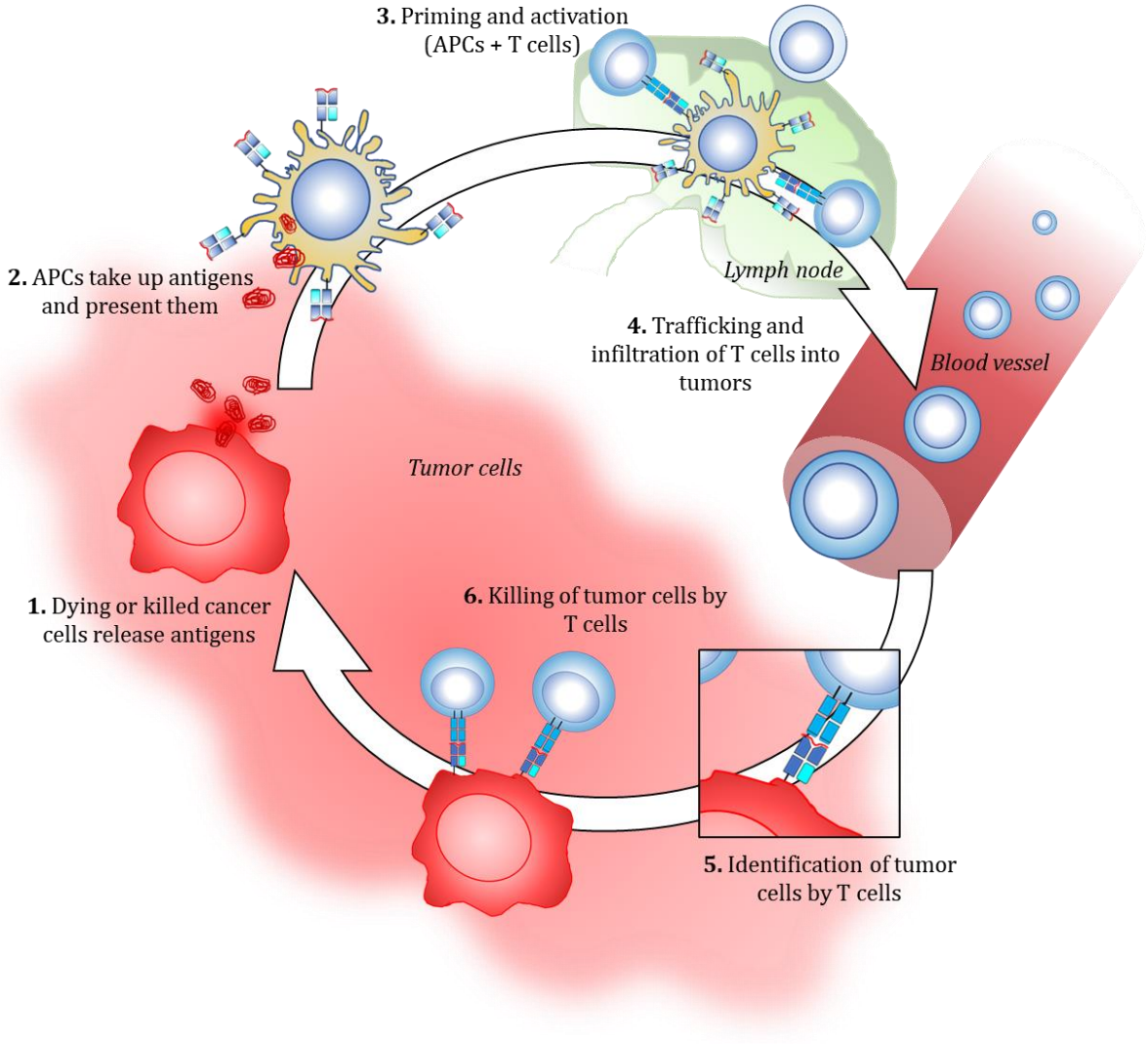


Figure 4: Overview of the cancer-immunity cycle. *The dying or killing of cancer cells leads to antigen release. An APC takes up the antigens, processes them and presents the peptides to T cells. T cells are primed, activated and finally migrate to the tumor, penetrate and recognize the tumor cells by the presented antigens. The recognized tumor cells are killed and antigens are released again* [89]*.*

The escape of tumor cells occurs through the mechanism of loss of immunogenicity or the acquisition of resistance to suppressive or cytotoxic mechanisms of the immune system [76,90,91]. Examples are reduced immune recognition through loss or downregulation of HLA or dysregulation of antigen processing [92]. A variety of immunosuppressive mechanisms are known, such as the production of transforming growth factor-β, indoleamine-2,3-dioxygenase, vascular endothelial growth factor or galectin, or the recruitment of regulatory immune cells, such as myeloid suppressor cells and T_{reg} cells [76,93].

2.6 A short outline of immunotherapy against cancer

The therapy of diseases by using the immune system is referred to as immunotherapy. In immunotherapy, an immune reaction is induced or enhanced, such as in tumor treatment, or the immune system is suppressed, as in the attenuation of the immune response in autoimmune diseases. Cancer immunotherapy began before the 19th century with William Coley's discovery of spontaneous tumor regression after treatment with bacteria, and with his development of Coley's toxin from dead bacteria, streptococcus pyogenes and serratia marcescens [94,95]. Today there are numerous immunotherapies, which can be divided into active and passive immunotherapies. Active immunotherapies activate the body's own immune system and passive immunotherapies are based on the activity of the drugs administered. In addition, both approaches can be classified into specific, such as direct targeting of cancer, or non-specific, general activation of the immune system. Specific immunotherapies are dependent on antigens, such as active immunotherapy with prophylactic virus vaccinations or with therapeutic anti-tumor vaccinations or passive immunotherapies such as tumor-specific antibodies or adoptive cell administration. Examples of non-specific immunotherapies that are independent of antigens are the active therapy with immune checkpoint inhibitors or the passive therapy with cytokines [96].

Based on the immunopeptidome specific immunotherapies can be developed. An advantage of the immunopeptidome is – in addition to surface antigens, which can be detected by antibodies - also antigens of intracellular proteins presented via pHLA, which are detected by T lymphocytes, can be targeted. This allows a wealth of new targets for immunotherapies, especially for neoantigens that are found in large numbers on intracellular proteins. Further advantages of specific immunotherapies, especially in comparison to current chemotherapies, are the induction of a memory function of the adaptive immune system and the low side effects due to immune tolerance mechanisms that maintain the integrity of the body in the presence of B and T lymphocytes with their antigen receptor specificities [97]. However, it must be mentioned that there have been individual cases of immunotherapy with significant toxicities [98,99]. Today, these advantages led to the development of numerous promising specific active immunotherapies based on pHLA epitopes including peptides, peptide loaded DCs, DNA, RNA, T cells for adoptive therapy (ACT) and virus-based systems. Multiple approaches progressed into clinical studies and are on the way to approval [100–103].

2.6.1 Peptide vaccination

The most important approach of the various immunotherapy platforms that were worked on in this doctorate at the Department of Immunology, Tübingen, is peptide vaccination, for which the Wirkstoffpeptidlabor was specially created and 2014 granted with the manufacturing authorization of peptide vaccination cocktails. Solid phase peptide synthesis enables the

production of peptide vaccines based on the epitope prediction described in Chapter 4, the tumor antigens characterized in Chapter 5 or tumor antigens that have already been discovered using the pipeline validated in Chapter 3.

The peptide vaccines in the Wirkstoffpeptidlabor are used as active immunotherapy to stimulate T cells by *in vivo* stimulation with subcutaneous or intradermal injected peptide formulations. An effective peptide vaccine is based on a strong activation of APCs leading to T cell activation [104] (Figure 5).

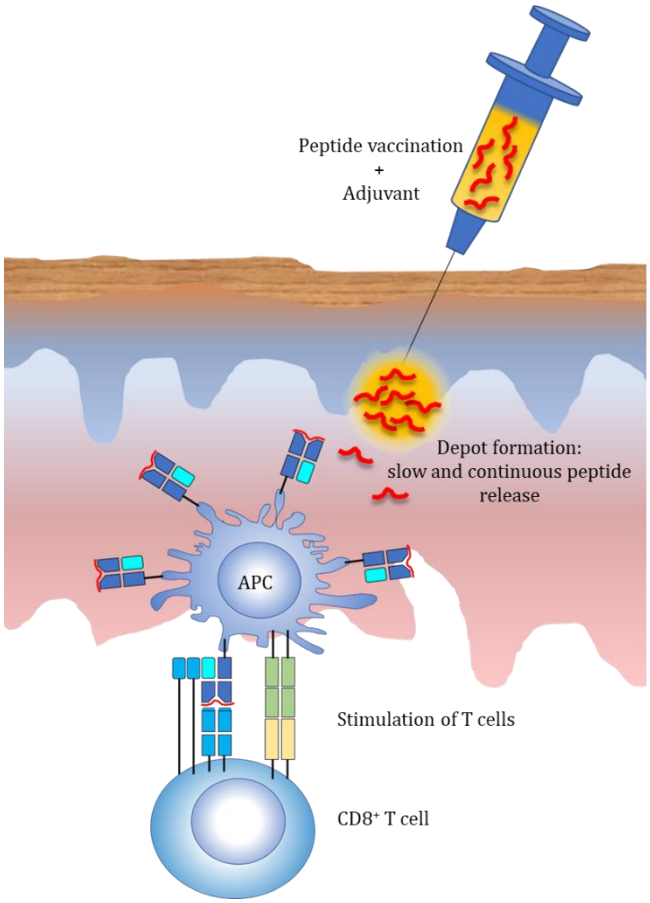


***Figure 5: Vaccination with peptides and adjuvant**. First, the peptide adjuvant formulation is injected subcutaneously or intradermally. Next, a depot is formed, and peptides are slowly and continuously released. This induces an activation of APCs by adjuvant binding to PRRs and T cells are stimulated* [105].

Normally, tumor-specific antigens are identified by immunopeptidomics of cancer samples (Chapter 3) or epitope prediction of HLA-presented peptides from the aa sequence (Chapter 4). Subsequently the immunogenicity, T cell recognition ability and ability to elicit an immune response of an epitope, are tested by *in vitro* immunogenicity screening [106]. Peptide vaccines contain one to more peptides consisting of short (CD8+ T cell epitope) to longer (T_H cell epitope) peptides to prevent immunological tolerance of tumor cells to these peptides [107]. Peptide cocktail formulations with multiple peptides allow a simultaneous attack of different presented antigens compared to single-peptide approaches, thereby reducing the likelihood of immune escape or adaptation of the tumor. Peptides per se have a weak immunogenicity and must therefore be applied in combination with an immunostimulatory adjuvant [104]. The combination with adjuvants enables the formation of depots for a gradual and steady peptide release and additionally an activation of APCs upon binding to PRRs. By mixing the peptides with the adjuvant a stable emulsion is formed, which generates a depot effect [108]. Examples of adjuvants with depot formation ability are the incomplete Freund's adjuvant, a

combination of water-in-oil, MF59, a combination of oil-in-water, or Montanide ISA™51/720, a water-in-oil emulsion [109–111].

Modern vaccine adjuvants activate specific APCs via different channels by specific induction of the Toll-like receptor (TLR)1/2 with XS15, TLR3 with poly(I:C), TLR4 with monophosphoryl lipid A (MPLA), TLR7 with imiquimod and TLR9 with CpG-containing oligonucleotides [112,113]. Strong immune responses can also be induced by combinations of adjuvants, such as the combination of XS15 and Montanide ISA™51, or the adjuvant system AS01, a liposome-based vaccine adjuvant system consisting of MPL and saponin QS-21, which enables vaccine development against malaria and herpes zoster through its synergistic effect [113,114]. Meanwhile, there are 266 clinical trials with peptide vaccines (25 active studies) and 4441 with immunotherapies for various diseases (clinicaltrials.gov, search term: "peptide vaccine"/"immunotherapy", date: March, 2020), which clearly shows the relevance and the many new insights expected from immunotherapeutic strategies in disease control.

2.7 Good manufacturing practice (GMP): Compliance with quality standards and method validation

Almost worldwide, all drugs must be manufactured, controlled, and distributed according to Good Manufacturing Practice (GMP) to assure pharmaceutical patient safety. This also applies to clinical trials where vaccine peptide treatments are currently being applied. The manufacturing process of the vaccine peptide cocktail from peptides to a drug is controlled and divided into peptide production and sterile filling, always monitored by quality control. Thus, GMP can guarantee a continual product quality in the manufacturing process. Production and analysis must follow standard operating procedures (SOPs), be continuously documented and allow complete traceability (Arzneimittel- und Wirkstoffherstellungsverordnung - AMWHV, 2006, last amended 09.08.2019). The active ingredients, the peptides, must be produced synthetically in the required quality by solid phase peptide synthesis and examined by high performance liquid chromatography and mass spectrometric analysis for identity, purity and content with which the product is declared.

Within the scope of this dissertation, following the recommendation of the Paul Ehrlich Institute, Langen, we intended to validate the reliability of the peptide identification pipeline, which provides the peptide sequences for future active ingredients, according to GMP quality standards. To ensure that the analytical method is suitable and reliable for identification, a method validation was performed (Chapter 3). According to the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), different characteristics are evaluated depending on whether the analytical procedure is an identification test, limit or quantitative test for impurities' content or quantitative measurement of the major component/s (Table 1).

Table 1: Validation characteristics*. Characteristics regarded as most important for validation of different analytical procedures. Robustness is not listed; however, it should be considered at an appropriate stage in analytical procedure development. Modified from validation of analytical procedures: text and methodology Q2(R1)* [115]*. Abbreviations: - normally not evaluated, + normally evaluated.*

Type of analytical procedure	**Identification**	**Testing for impurities**		**Assay**
Characteristics		Limit tests	Quantitative tests	
Accuracy	-	-	+	+
Precision				
- Repeatability	-	-	+	+
- Intermediate precision	-		+	+
Specificity	+	+	+	+
Detection limit	-	+	-	-
Quantitation limit	-	-	+	-
Linearity	-	-	+	+
Range	-	-	+	+

Characteristics are the accuracy, closeness of a true value to the detected value, precision, closeness between a series of measurements from multiple sampling (repeatability: over a short interval of time; intermediate precision: within-laboratories variation such as different measurement days), specificity, the unambiguous determinability of the analyte in the presence of components, limit of detection (LOD), the lowest detectable amount of the analyte in a sample, limit of quantification (LOQ), the lowest quantifiable amount of the analyte in a sample, linearity, the range that gives results directly proportional to the concentration of the analyte in the sample, range, section in which the analytical method has an appropriate degree of precision, accuracy and linearity and robustness, capability of the analytical procedure to remain unaffected by small variations of the method parameters [115]. It should be noted that the requirements are dependent on the approving authorities and may differ from the ICH guidelines.

2.8 Objectives

For almost three decades now, the immunopeptidome has been analyzed by eluting peptides from MHC molecules. Immunopeptidomics has already been established in several institutes and companies worldwide. The method is now used for various investigations from the simple identification of MHC peptide motifs for different organisms to the detection of cryptic disease-specific peptides. The range of applications is already widespread, yet the immunopeptidome still contains a great variety of information still waiting to be deciphered (Chapter 2).

Despite the successes, the method is still not ideal. The immunopeptidome contains a large number of peptides with different affinities. It can only be analyzed in its entirety using LC-MS/MS based immunopeptidomics, which is limited in its sensitivity and therefore has shortcomings such as a lower recovery rate. Consequently, immunopeptidomic data are not directly comparable between different mass spectrometers.

Is it still possible to validate immunopeptidomics and use it reliably for clinical studies and drug development? Is there nowadays a reliable method to identify the peptide motif for each MHC allotype, the cornerstone for epitope predictions or reliable active substance identification? What further information besides epitope identification for therapies can be derived from the individual peptides? Is it possible to use peptides to classify HLA allotypes or differentiate between healthy and malignant tissue? Can tumor-specific peptides be reliably identified with this omic technology?

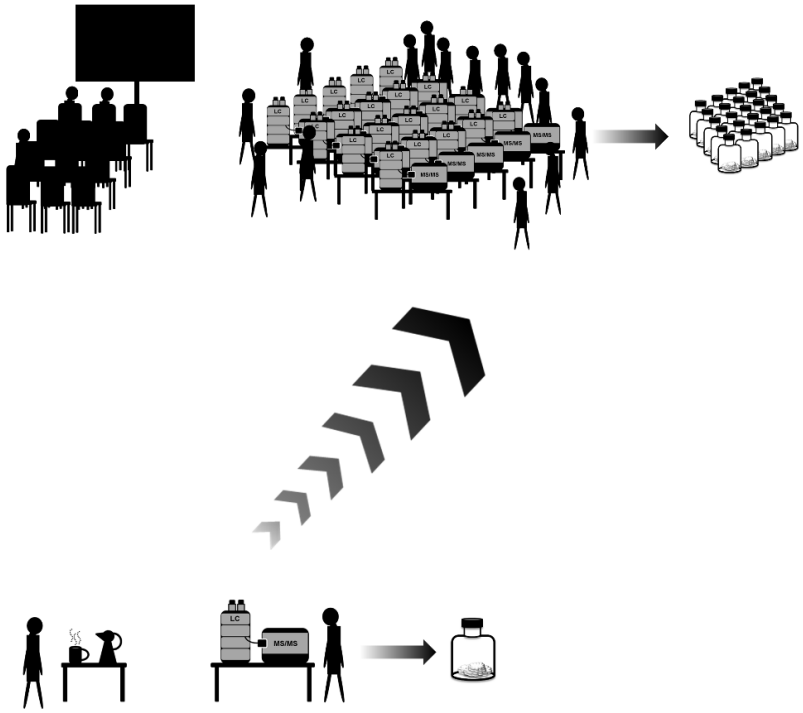
Within the scope of this dissertation the immunopeptidomic method should be validated to ensure the reliability of LC-MS/MS peptide identification. All required parameters of the EMA and FDA should be investigated to verify GMP suitability (Chapter 3). Furthermore, an updated protocol for the identification of HLA ligands, deconvolution of peptide motifs and generation of matrices for epitope prediction should be established, which can be used for monoallelic cells as well as multiallelic tissue (Chapter 4). Finally, a method should be developed to identify allotypic peptides that allow HLA typing. In addition, the peptides could also be used as an internal standard for semi-quantitative investigation of the tumor specificity of peptides. Further possibilities of this method were investigated in order to eventually be able to determine the tissue origin or even the dignity of samples (Chapter 5).

3 Guidance document: validation of a high-performance liquid chromatography-tandem mass spectrometry immunopeptidomics assay for the identification of HLA class I ligands suitable for pharmaceutical therapies

3.1 Publication and author contributions

The chapter was accepted for publication in Molecular & Cellular Proteomics titled "Guidance Document: Validation of a High-Performance Liquid Chromatography-Tandem Mass Spectrometry Immunopeptidomics Assay for the Identification of HLA Class I Ligands Suitable for Pharmaceutical Therapies" [116], was selected as issue highlight and covered MCP Vol. 19, Issue 3, 1 Mar 2020. Authors contributing to this work are listed below. All experiments, data analysis and manuscript writing were performed by the author of this thesis, except of the described contributions in following lines. Marion Gauger supported the data analysis and writing of the paper. Ana Marcu and Annika Nelde performed experiments (of the robustness section). Monika Denk supported the data analysis. Heiko Schuster, Hans-Georg Rammensee, Stefan Stevanović supported the design of research and data evaluation. All authors proofread the manuscript.

Cover



About the cover *(MCP Vol. 19, Issue 3, 1 Mar 2020):* *Validation of the LC-MS/MS immunopeptidomics pipeline according to good manufacturing practice (GMP) to accelerate the development from experimental laboratory scale to industrial high throughput. Validation of the accuracy, precision, specificity, detection limit and robustness of the pipeline is feasible and ensures reliable performance. We hope that this validation template will accelerate vaccine development and translation into clinical practice.*

Author information

Michael Ghosh[I,II], Marion Gauger[I], Ana Marcu[I], Annika Nelde[I], Monika Denk[I,III], Heiko Schuster[I,IV], Hans-Georg Rammensee[I,III], Stefan Stevanović[I,III]

Affiliations

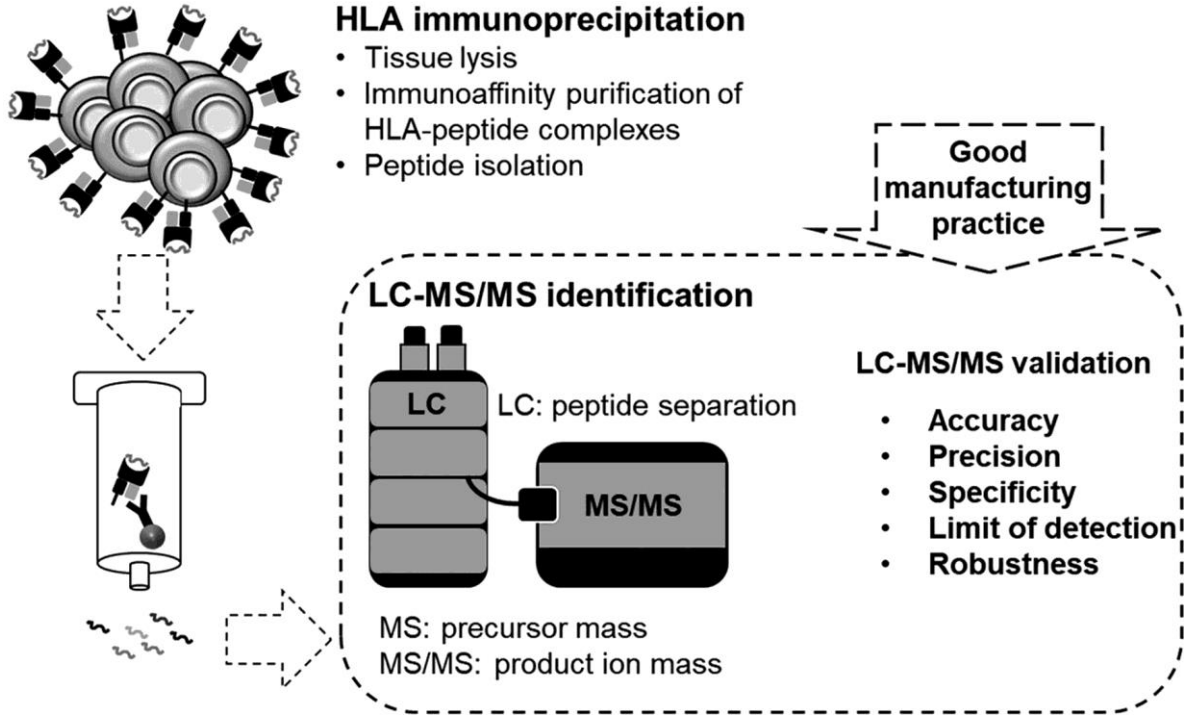
I Department of Immunology, Institute for Cell Biology, University of Tübingen, Tübingen, Germany

II Natural and Medical Science Institute at the University of Tübingen (NMI), Reutlingen, Germany

III German Cancer Research Center (DKFZ) partner site and German Cancer Consortium (DKTK) Tübingen, Tübingen, Germany

IV Immatics Biotechnologies GmbH, Tübingen, Germany

3.2 Graphical abstract and highlights



Highlights

- Validation of an omic method for antigen identification using LC-MS/MS.
- Validation of accuracy, precision, specificity, limit of detection and robustness.
- Validation according to the current FDA and EMA guidelines.

Abbreviations

- AcN, Acetonitrile
- BC, Bladder cancer
- CLL, Chronic lymphocytic leukemia
- EMA, European Medicines Agency
- FDA, Food and Drug Administration
- FDR, False discovery rate
- GLP, Good laboratory practice
- GMP, Good manufacturing practice
- HLA, Human leukocyte antigen
- LOD, Limit of detection
- OECD, Organisation for Economic Co-operation and Development
- PBMC, Peripheral blood mononuclear cells
- PPM, Parts per million
- SD, Standard deviation

3.3 Abstract

For more than two decades naturally presented, human leukocyte antigen (HLA)-restricted peptides (immunopeptidome) have been eluted and sequenced using liquid chromatography-tandem mass spectrometry (LC-MS/MS). Since, identified disease-associated HLA ligands have been characterized and evaluated as potential active substances. Treatments based on HLA-presented peptides have shown promising results in clinical application as personalized T cell-based immunotherapy. Peptide vaccination cocktails are produced as investigational medicinal products under GMP conditions. To support clinical trials based on HLA-presented tumor-associated antigens, in this study the sensitive LC-MS/MS HLA class I antigen identification pipeline was fully validated for our technical equipment according to the current US Food and Drug Administration (FDA) and European Medicines Agency (EMA) guidelines.

The immunopeptidomes of JY cells with or without spiked-in, isotope labeled peptides, of peripheral blood mononuclear cells of healthy volunteers as well as a chronic lymphocytic leukemia and a bladder cancer sample were reliably identified using a data-dependent acquisition method. As the LC-MS/MS pipeline is used for identification purposes, the validation parameters include accuracy, precision, specificity, limit of detection and robustness.

3.4 Introduction

The immunopeptidome is a vast and diverse compilation of HLA-presented peptides (HLA ligands), which serve as a showcase of inter- and intracellular processes. T cells recognize presented peptides in the immunopeptidome, which is constantly modulated by gene expression, transcription, translation, posttranslational modification and antigen processing and presentation[117–120]. Especially in tumor immunology, HLA ligands are used in many ways. They are suited as biomarkers, presenting intracellular abnormalities like malignant transformation and as active pharmaceuticals, activating cancer specific T cells[121].

Natural HLA ligands have been isolated and sequenced using LC-MS/MS for almost three decades[122–128]. So far, the LC-MS/MS analysis is the only method to investigate the entirety of HLA-presented peptides. However, based on these peptide data *in silico* prediction tools have been developed, which allow the prediction of possibly presented peptides from exome, RNA or whole genome sequencing data and have extended the toolbox even further[129–134].

Developing from such identifications, peptide vaccination cocktails have been produced as active pharmaceuticals under GMP conditions[135]. The acceptance, safety and efficacy of peptide vaccinations have been investigated[135] and several clinical studies testing peptide vaccinations have been performed with our contribution (GAPVAC[121], NCT02149225 and NOA-16, NCT02454634) or are ongoing (iVAC-CLL01, NCT02802943). The procedures of active substance production, analysis, and batch release have been validated and reliably lead to reproducible

products of desired quality. However, the initial antigen identification procedure using mass spectrometry-based immunopeptidomics has not been validated yet. In this study, the LC-MS/MS antigen identification procedure was fully validated for our technical equipment according to current FDA and EMA guidelines to support further clinical trials based on HLA-presented tumor-associated antigens. This validation should serve as a guidance that can be adapted to other LC-MS/MS platforms and samples.

Protocols for large-scale immunopeptidomics using LC-MS/MS and the identification of HLA ligands are established and have been published[64,136,137]. To our knowledge, no validation of an omics method using LC-MS/MS-has been published so far[138,139]. This article presents an immunopeptidomics assay using LC-MS/MS, which is fully validated according to the latest US Food and Drug Administration (FDA) and European Medicines Agency (EMA) guidelines[138–142]. We have to emphasize that such validations are specific only for dedicated equipment of one distinct laboratory. We provide a first protocol and template to enhance the validation of other laboratories with similar equipment and other omics fields using LC-MS/MS such as proteomics, metabolomics, and lipidomics.

3.5 Experimental procedures

3.5.1 Peptide synthesis

The automated peptide synthesizer Liberty Blue (CEM) was used to synthesize peptides following the 9-fluorenylmethyl-oxycarbonyl/tert-butyl (Fmoc/tBu) strategy. The identity and purity of the peptides were confirmed using a reversed-phase liquid chromatography (Alliance e2965, Waters) and an uHPLC system (nanoUHPLC, UltiMate 3000 RSLCnano, Dionex) on-line coupled LTQ Orbitrap XL hybrid mass spectrometer (ThermoFisher) system. Synthesized peptides were employed in the validation of LC-MS/MS identifications. Peptide sequences used for the validation are listed in supplemental Table S1.

3.5.2 Tissue samples

The EBV-transformed human B-cell line JY (ECACC 94022533) was cultured in RPMI1640 with 10% heat-inactivated fetal bovine serum (FBS) and 1% penicillin/streptomycin to a total number of $1*10^{11}$ cells, centrifuged at 1,500 rpm for 15 min at 4°C, washed two times with cold PBS and aliquots containing $75*10^{6}$ cells were frozen and stored at -80°C until use. The cells were tested negative for mycoplasma contamination via PCR.

The peripheral blood mononuclear cells (PBMC), chronic lymphocytic leukemia (CLL) and bladder cancer (BC) tissue samples were collected at the University Hospital of Tübingen with the informed consent of patients according to the principles of the Declaration of Helsinki. The local institutional review board (Ethics Committee at the Medical Faculty and the University Hospital of Tübingen) has approved the use of the patient samples.

3.5.3 Immunoaffinity purification of HLA ligands

HLA class I molecules were isolated using standard immunoaffinity purification as described[125,136,143,144] using the HLA class I-specific monoclonal antibody W6/32[145]. First, the cell pellets were lysed in 10 mM CHAPS (Applichem)/PBS (Lonza) containing protease inhibitors (Complete, Roche) and subsequently HLA molecules were purified using the pan-HLA class I–specific monoclonal W6/32 Ab covalently linked to CNBr-activated Sepharose (GE Healthcare). Repeated addition of 0.2% trifluoroacetic acid (Merck) eluted HLA molecules and peptides. The peptides were isolated employing ultrafiltration with centrifugal filter units (Amicon, Merck Millipore), extracted and desalted using ZipTip C18 pipette tips (Merck Milli-pore), eluted in 35 µl acetonitrile (Merck)/0.1% trifluoroacetic acid, vacuum centrifuged to 5 µl, and resuspended in 25 µl of 1% acetonitrile/0.05% trifluoroacetic acid. Finally, the peptide solutions were stored at -20°C until analysis by LC-MS/MS.

3.5.4 Analysis of HLA ligands by LC-MS/MS

Peptides were separated by nanoflow high-performance liquid chromatography (nanoUHPLC, UltiMate 3000 RSLCnano, Dionex) and subsequently analyzed in an on-line coupled Orbitrap Fusion Lumos or LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific). Volumes of 5 µl peptide solution were injected onto a 75 µm × 2 cm trapping column (Acclaim PepMap RSLC, Dionex) at 4 µl/min for 5.75 min in five technical replicates. Subsequently, peptide separation was performed at 50°C at a flow rate of 300 nl/min on a 50 µm × 25 cm separation column (Acclaim PepMap RSLC, Dionex) applying a gradient ranging from 2.4 to 32.0% of AcN over the course of 90 min. Eluted peptides were ionized by nanospray ionization and analyzed in the Orbitrap Fusion Lumos implementing top speed collision-induced dissociation (CID) fragmentation. Survey scans were performed at 120,000 resolution and fragment detection at 60,000 resolution in the Orbitrap.

To demonstrate a method transfer, the immunopeptidomics pipeline was transferred from the Orbitrap Fusion Lumos to a LTQ Orbitrap XL. In the LTQ Orbitrap XL peptides were analyzed using a top five CID method with survey scans at 60,000 resolution and fragment ion detection in the ion trap operated at normal scan speed. On both instruments, the mass range was limited to 400–650 m/z with precursors of charge states 2+ and 3+ eligible for fragmentation.

Maintenance and OQ of the LC-MS/MS system are performed annually (Thermo Fisher Scientific). A positive ion calibration using a Pierce™ LTQ Velos or LTQ ESI positive ion calibration solution (Thermo Fisher Scientific) and a system suitability test using natural HLA class I-presented peptides of JY cells is performed weekly.

3.5.5 Database search and spectral annotation

Data was processed against the human proteome included in the Swiss-Prot database (http://www.uniprot.org, release September 27, 2013; containing 20,279 reviewed protein

sequences) applying the Sequest algorithm[146] in the Proteome Discoverer (version 1.3, Thermo Fisher) software.

Precursor mass tolerance was set to 5 ppm, product ions mass tolerance was set to 0.02 Da for Orbitrap Fusion Lumos data and 0.5 Da for LTQ Orbitrap XL Data and oxidized methionine was allowed as the only dynamic modification with no restriction by enzymatic specificity. Percolator[147]-assisted false discovery rate (FDR) calculation was set at a target value of $q \leq 0.05$ (1% FDR). Peptide-spectrum matches with $q \leq 0.05$ were filtered according to additional orthogonal parameters to ensure spectral quality and validity. Peptide lengths were limited to 8–12 amino acids.

3.5.6 Validation procedures

The validation of the immunopeptidomics procedure was done according to the OECD principles of Good Laboratory Practice (GLP)[148] and accuracy, precision, specificity, limit of detection and robustness were validated according to the FDA and EMA guidelines[140,141]. Clear definitions can be found in[140,141,148].

As the immunopeptidome LC-MS/MS system separates the peptides using the LC and subsequently identifies the peptide ions in MS mode and product ions in the MS/MS mode, we tried to consider these three parts for every validation parameter summarized in Table 1 and Figure 1.

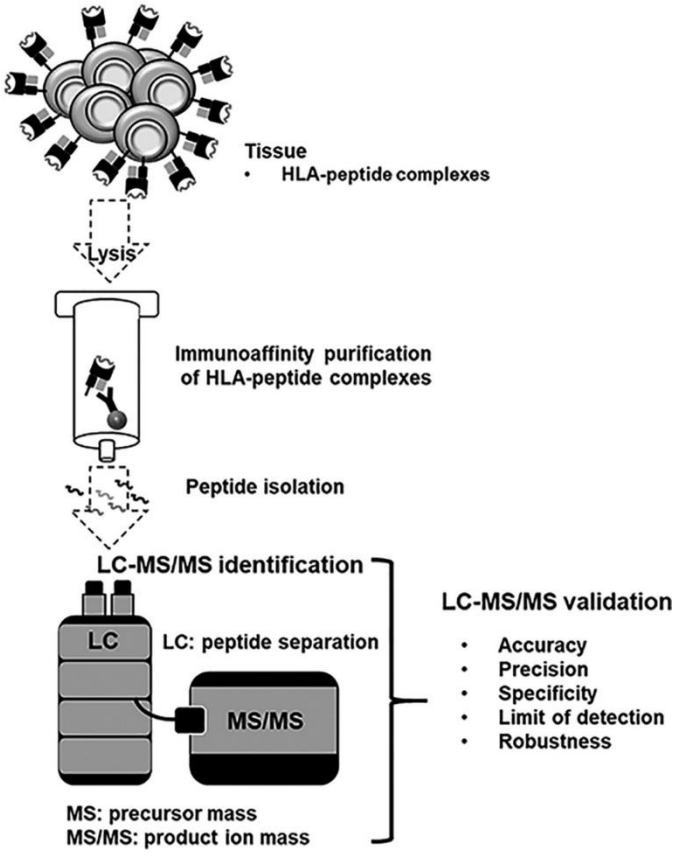


Figure 1: Schematic overview of the validation of the LC-MS/MS immunopeptidomics assay for the identification of HLA ligands suitable for pharmaceutical therapies. *The LC-MS/MS pipeline is used for identification purposes, consequently the validation parameters accuracy, precision, specificity, limit of detection and robustness were validated according to current FDA and EMA guidelines.*

3.5.7 Experimental design and statistical rationale

A summary of the performed experiments, samples, technical replicates, and MS RAW files is given in supplemental Table S2. Results were analyzed using GraphPad Prism (GraphPad software Inc).

The recovery rate was obtained by taking the average of the percentual overlapping peptides between the technical replicates normalized to the total number of peptides. The LC peptide retention times (RTs) were compared calculating the average of the Pearson correlations of the technical replicates.

3.6 Results

3.6.1 Accuracy

To investigate the accuracy and specificity, the purified HLA-eluted peptides from one JY batch was spiked with 100 fmol isotope labeled synthetic peptides (Table S1) and analyzed in three separate analytical replicates (for identified peptides, see supplemental Table S3). The accuracy of the mass spectrometer did fulfill the acceptance criteria (Table 1) with a deviation below 2 ppm between the median mass deviation from the theoretical mass of all identified natural (median ΔM: 0.05 ppm) and synthetic peptides (median ΔM: 0.19 ppm) in Figure 2A. The peptide RTs between the replicates of all natural and all synthetic peptides do have a mean Pearson correlation above 95%, verifying the accuracy of the LC (Figure 2B).

Table 1: Acceptance criteria for the different parameters selected for LC-MS/MS validation.

The acceptance criteria for the selected parameters are indicated for the mass spectrometer and the LC. Abbreviations: PC, Pearson correlation; SD, standard deviation.

	MS and MS/MS	**LC**
Characteristics	**Specification/ acceptance criteria**	
Accuracy (comparison to theoretical masses and synthetic peptides)	The median of the deviation of the theoretical masses (ΔM ppm) of • the entirety of peptides • the selected peptides between natural and synthetic peptides should be ≤ 2 ppm	PC ≥ 95%
Precision (natural peptides) - repeatability - intermediate precision	SD of peptide number: ≤ 10% Recovery rate: 80% ± 20%	PC ≥ 95%
Specificity (natural and synthetic peptides)	The SD between the precursor ion and five top fragment masses of selected peptides: ≤ 0.001 Da. The peptide must be identified in two of three replicates. Selectivity of all identified peptides based on precursor mass combined with top five fragments	PC ≥ 95%
Limit of detection (synthetic peptides)	50% (n = 31) of the peptides have to be identified Recovery rate: 80% ± 20%	PC ≥ 95%
Robustness (natural peptides from three primary samples isolated by three different persons)	**Accuracy:** as mentioned above **Precision:** • repeatability: as mentioned above **Specificity:** as mentioned above	PC ≥ 95%

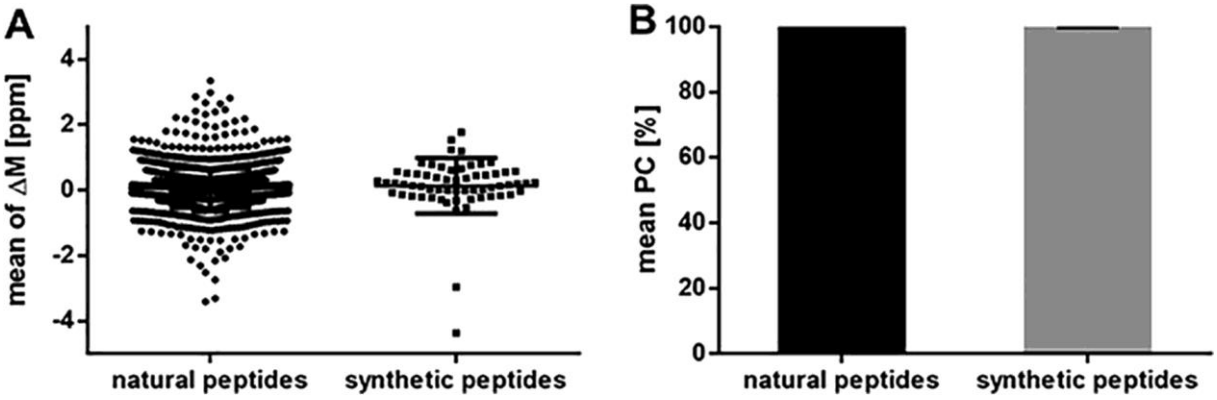


Figure 2: Validation of the accuracy using immunopeptidomes from JY cells and spiked isotope labeled synthetic peptides. *Three replicates were analyzed. A, Mass deviation of the detected precursor mass from the theoretical mass (ΔM ppm) of all identified natural (n = 1648) and synthetic peptides (n = 62). B, Mean Pearson correlation of the peptide retention times. Abbreviations: PC, Pearson correlation; ppm, parts per million; ΔM, mass deviation.*

3.6.2 Specificity

Based on our experience from the first series of analyses, the five peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL are expected as natural HLA class I-presented peptides of JY cells. In order to prove the specificity, the mass spectrometer must fulfill the MS mode acceptance criteria for precursor ions and the MS/MS mode acceptance criteria for five selected top product ions of the expected five peptides (Table 1). Here, we use two ways to select the five top product ions, we simply choose the top five most intensive fragments (last paragraph of specificity) or the most intensive fragments such as b and y fragments with the highest intensity and relevance (penultimate paragraph of specificity). Furthermore, in the LC separation, the correlation of the retention times of the natural and synthetic counterparts of the five peptides have to fulfill the acceptance criteria.

The difference of the median of the mass deviation from the theoretical mass (ΔM ppm) of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL, which were identified as natural (median ΔM: 1.34 ppm) and synthetic peptides (median ΔM: 0.09 ppm), is below 2 ppm (Figure 3A) (for identified peptides and product ions, see supplemental Table S4).

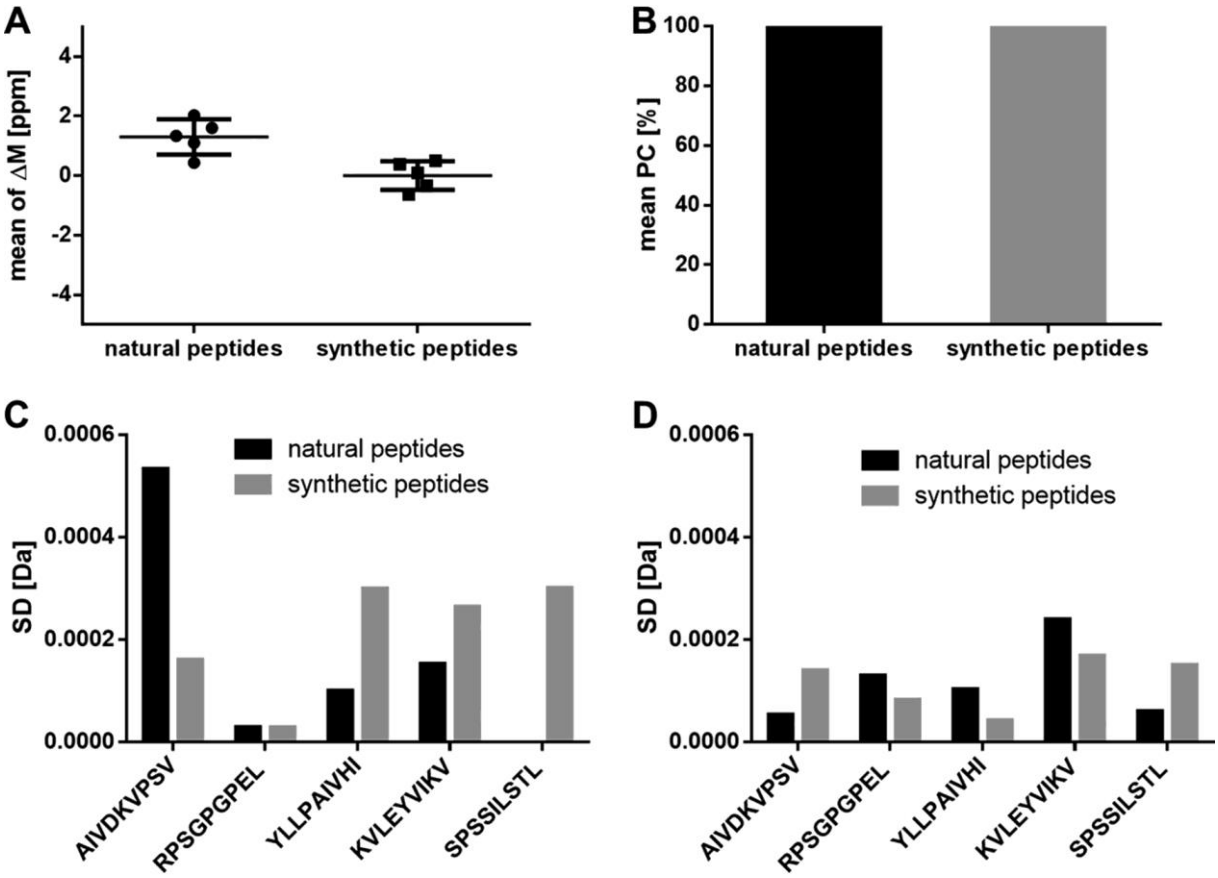


Figure 3: Validation of the specificity using immunopeptidomes from JY cells and spiked isotope labeled synthetic peptides. *Three replicates were analyzed. A, Mean of the mass deviation from the theoretical precursor mass (ΔM ppm) of the five identified natural and synthetic peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL in three technical replicates. B, Mean Pearson correlation of the RTs of the five identified natural and synthetic peptides. Mass deviation as S.D. of the (C) precursor ion masses in MS mode and the (D) resulting five selected top fragments in MS/MS modes of the five identified natural and synthetic peptides. Abbreviations: PC, Pearson correlation; ppm, parts per million; ΔM, mass deviation; S.D., standard deviation.*

The peptide RTs between the replicates of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL, which were identified as natural and synthetic peptides, do have a Pearson correlation above 95% (Figure 3B).

The standard deviation (SD) of the mass accuracy of the precursor masses in MS mode (Figure 3C) and of the five selected top product ions, selected based on intensity and relevance, in MS/MS mode (Figure 3D, for MS/MS spectra, see supplemental Figure S1) of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL is below 0.001 Da for both the natural and synthetic peptides.

As an additional step to prove the specificity of the LC-MS/MS system, we validated our manual quality control method, to distinguish all peptides based on the mass of a precursor ion combined

with the masses of the top five most intensive product ions. As the SD of the mass accuracy deviates at four decimals, our peptide identity criteria of the quality control should enable specificity at a four-decimal level (Table 1). Based on the precursor masses in MS mode at four-digit level there was an overlap of 7 to 12 natural peptide masses in the three replicates (Table 2A). These peptide masses could be separated using the top five masses with the highest intensity in MS/MS mode. In MS/MS mode the highest number of overlapping product ion masses were two duplicates in replicate two (Table 2B). All synthetic peptides could be separated based on the precursor masses at four decimals (Table 2C).

Table 2: Validation of the specificity and suitability of the top five product ion peptide quality control using immunopeptidomes from JY cells and spiked isotope labeled synthetic peptides. *A, Overlap of the detected peptide precursor masses in MS mode of the natural peptides at four decimals. B, Overlap of the measured top five product ion masses of the manifold peptide precursor masses from (A) in MS/MS mode of the natural peptides at four decimals. C, Overlap of the measured 62 synthetic peptide precursor masses in MS mode at four decimals.*

(A) MS: natural peptides, precursor masses			
replicates	**1**	**2**	**3**
unique	1181	1260	1362
duplicate	7	12	12
(B) MSMS: natural peptides, five top-fragment masses			
replicates	**1**	**2**	**3**
unique	70	116	120
duplicate	0	2	0
(C) MS: synthetic peptides			
replicates	**1**	**2**	**3**
unique	61	62	61
duplicate	0	0	0

3.6.3 Limit of Detection

To determine the limit of detection (LOD), four aliquots of purified HLA-eluted peptides from JY cells were spiked with 0.1 fmol, 1 fmol, 10 fmol, or 100 fmol isotope labeled synthetic peptides (Table S1) and analyzed in three replicates leading to 12 separate analytical replicates (for

identified peptides, see supplemental Table S3). Based on our experience with JY cells, in HLA ligandomic experiments an optimal setting enables a peptide recovery rate of 80 ± 20% between two replicates. Thus, there is no LOD where 100% of the peptides will be discovered, especially in data-dependent acquisition. Here, we set the LOD to the peptide concentration that enables an identification of at least 50% (n = 31) of the peptides per replicate, with a recovery rate of 80% ± 20% and a Pearson correlation of the peptide retention times above 95% between three replicates.

The JY sample spiked with 10 fmol synthetic peptides had the lowest peptide content enabling a reproducible identification of 50% of the 62 added isotope labeled peptides (Figure 4A). At the LOD the recovery rate of peptides in a replicate mass spectrometric measurement is in the range of 80% ± 20% (Figure 4B) and the mean Pearson correlation of the retention times of the synthetic peptides in the three technical replicates analyzed in the LC is above 95% (Figure 4C).

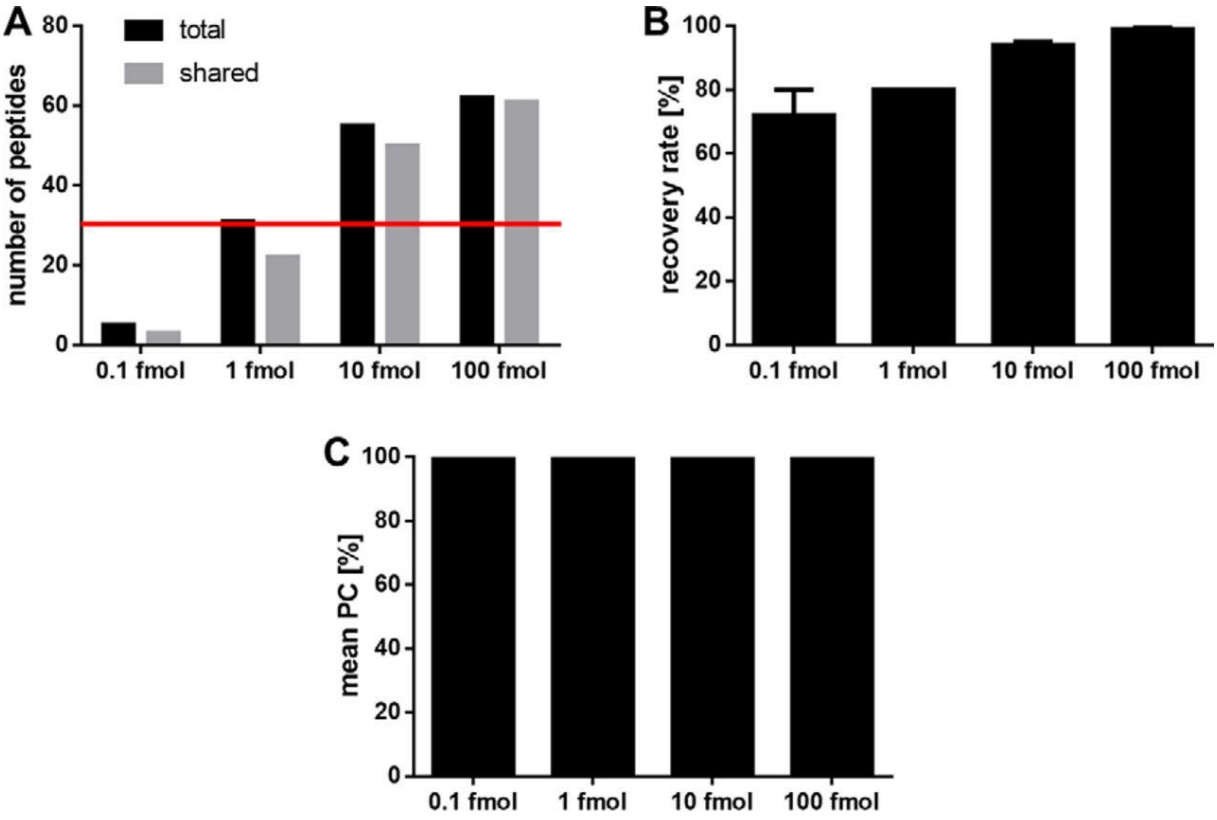


Figure 4: Validation of the limit of detection using spiked isotope labeled synthetic peptides. *Three replicates of JY samples spiked with 0.1 fmol, 1 fmol, 10 fmol, and 100 fmol isotope labeled synthetic peptides were analyzed. A, Number of total identified isotope labeled peptides and shared peptides identified in the three replicates. The LOD of 50% (n = 31) of the spiked synthetic peptides is indicated with a red line. B, Recovery rate of synthetic peptides recovered between the three replicates of each condition. C, Mean Pearson correlation of the peptide retention times between the replicates. Abbreviation: PC, Pearson correlation.*

3.6.4 Precision

The precision (also referred to as imprecision) was determined by assaying three aliquots of HLA-eluted peptides from JY samples in three technical replicates leading to nine separate analytical replicates. To prove the intermediate precision, the measurement series was repeated after seven days (for identified peptides, see supplemental Table S5). The number of identified peptides fulfilled the acceptance criteria of ± 10% SD of the repeatability on the initial day and after one week (Figure 5A, B). Furthermore, the acceptance criteria of the recovery rate with a recovery of 80 ± 20% of identified peptides in a repeated replicate were fulfilled on both measuring days (Figure 5C). A closer look at the LC demonstrates that the mean Pearson correlation of the peptide retention times between all nine replicates was above 95% and fulfilled the criteria of the repeatability and intermediate precision (Figure 5D).

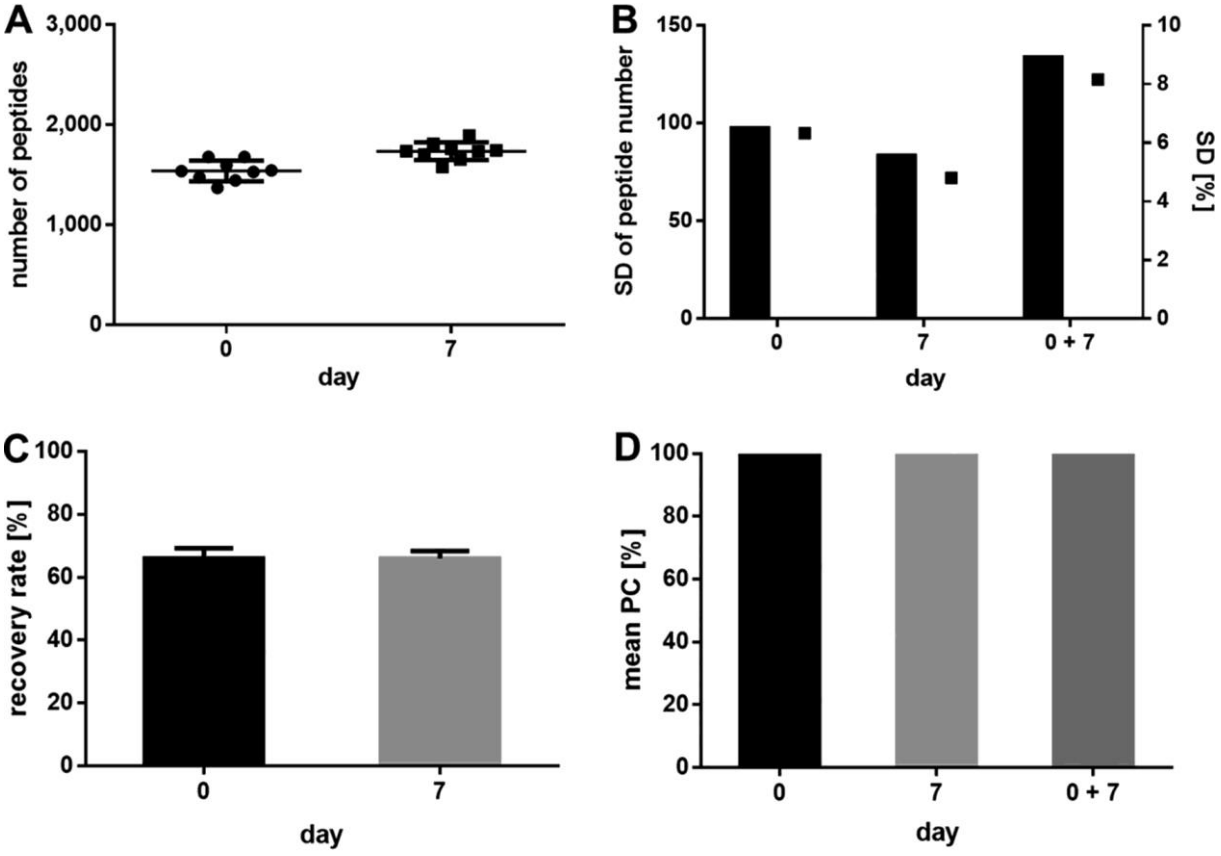


Figure 5: Validation of the precision using immunopeptidomes from JY cells. *A, Number of synthetic peptides identified at day 0 and day 7 in nine replicates, respectively. B, Standard deviation (S.D.) given in total peptide numbers and in percent. C, Recovery rate of peptides between the replicates. D, Mean Pearson correlation of the peptide retention times between the replicates. Abbreviation: S.D., standard deviation; PC, Pearson correlation.*

3.6.5 Robustness of the precision, accuracy and specificity

The robustness was investigated performing a retrospective analysis of the purified HLA-eluted peptides of three primary samples: peripheral blood mononuclear cells from a healthy donor, a chronic lymphocytic leukemia sample, as well as a bladder cancer sample. The immunopeptidomes were isolated and analyzed by three different persons. To validate the robustness the specifications indicated in Table 1 should be fulfilled with regard to accuracy, precision, and specificity.

The identified peptides of the three primary samples fulfill the acceptance criteria of the accuracy for the mass spectrometer and for the LC. The difference of the median of the mass deviation from the theoretical mass of all identified natural peptides (median ΔM: PBMC -0.06 ppm, CLL 0.05 ppm, BC 0.14 ppm) (Figure 6A) and synthetic peptides (median ΔM: 0.19 ppm) is below 2 ppm (Figure 6A). The peptide RTs between the replicates of all natural and all synthetic peptides have a Pearson correlation above 95% (Figure 6B).

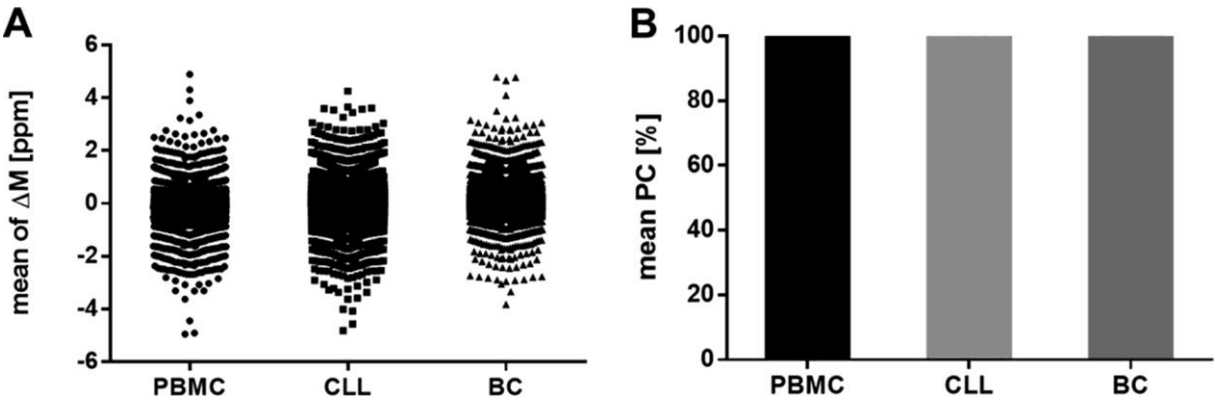


Figure 6: Validation of the accuracy using immunopeptidomes from primary PBMC, CLL, and BC samples. *Three replicates were analyzed. A, Mass deviation from the theoretical precursor mass (ΔM ppm) to the theoretical mass of all identified natural peptides. B, Mean Pearson correlation of the peptide retention times between the replicates. Abbreviations: PC, Pearson correlation; ppm, parts per million; ΔM, mass deviation.*

Regarding the precision, the three technical replicates of the three primary samples fulfill the acceptance criteria of the repeatability for both the mass spectrometer and the LC. The replicates have a percentage SD of identified peptide numbers below 10% (Figure 7A, B) and the recovery rate is in the range of 80% ± 20% (Figure 7C). The peptide RTs between the replicates have a Pearson correlation above 95% (Figure 6B).

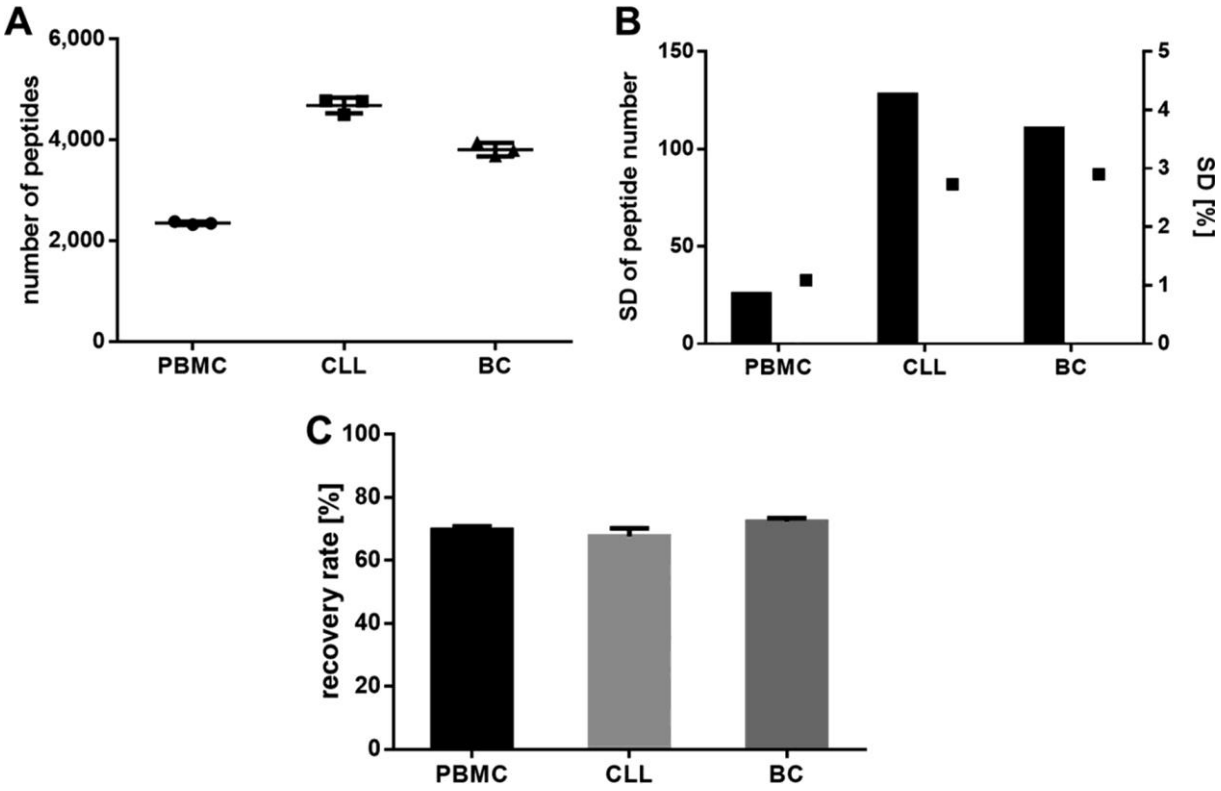


Figure 7: Validation of the repeatability using immunopeptidomes from primary PBMC, CLL, and BC samples. *A, Total number of natural peptides identified in three technical replicates, respectively. B, Standard deviation (S.D.) given in total peptide numbers and in percent. C, Recovery rate of peptide identifications between the replicates. Abbreviation: S.D., standard deviation.*

The specificity can be investigated, as all analyzed primary samples are expected to contain at least one of the previously used 62 synthetic peptides as natural HLA class I-presented peptide. The natural peptides should fulfill the acceptance criteria of the accuracy and specificity indicated in Table 1 for the mass spectrometer and the LC, respectively (for identified peptides and product ions, see supplemental Table S6).

The three replicates fulfill the acceptance criteria of the specificity for the mass spectrometer and for the LC. The difference of the median of the mass deviation from the theoretical mass (ΔM ppm) of the peptides AIVDKVPSV, YLLPAIVHI, GTYVSSVPR, RPSGPGPEL, SVINLVIVK and RVYGGITTK, which were identified as natural and 100 fmol spiked synthetic peptides, is below 2 ppm (Figure 8A).

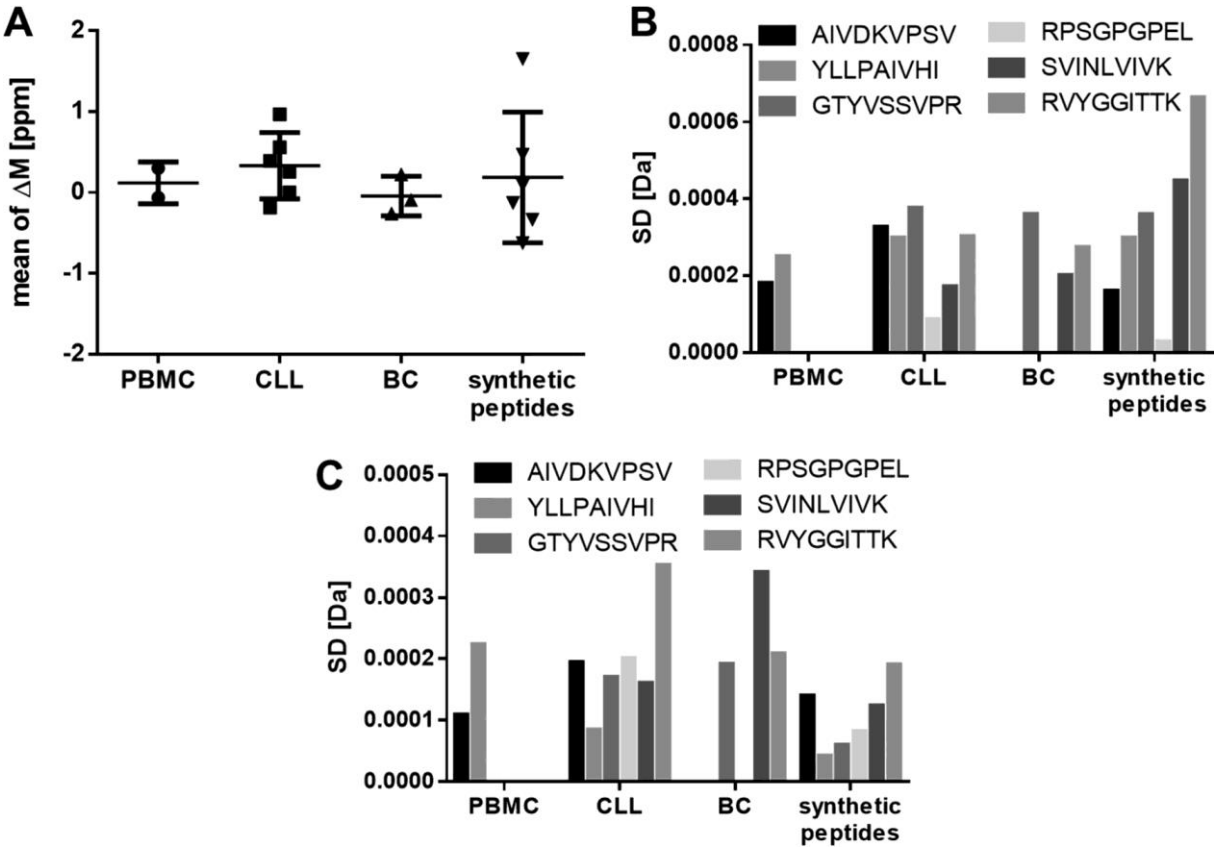


Figure 8: Validation of the specificity using immunopeptidomes from primary PBMC, CLL, and BC samples and isotope labeled synthetic peptides spiked into JY. *The samples were analyzed in three replicates. A, Mean mass deviation of the detected precursor masses from the theoretical masses (ΔM ppm) of the identified natural peptides AIVDKVPSV, YLLPAIVHI, GTYVSSVPR, RPSGPGPEL, SVINLVIVK and RVYGGITTK and 100 fmol synthetic peptides spiked into JY. B, Mass deviation as S.D. of the precursor masses detected in MS mode and (C) the resulting five selected top fragments in MS/MS modes of the identified natural and synthetic peptides. Abbreviations: ppm, parts per million; ΔM, mass deviation; S.D., standard deviation.*

The SD of the mass accuracy of the precursor masses in MS mode and of the selected five top product ions in MS/MS mode of the peptides AIVDKVPSV, YLLPAIVHI, GTYVSSVPR, RPSGPGPEL, SVINLVIVK and RVYGGITTK is below 0.001 Da for both the natural and synthetic peptides (Figure 8B and C, for MS/MS spectra, see supplemental Figure S1).

3.6.6 Transfer of the method to other LC-MS/MS systems

In addition to the previous robustness analyses, the method was transferred to an LC-MS/MS system with a less sensitive LTQ Orbitrap XL and the HLA-eluted peptides from JY cells were analyzed. To demonstrate the method transfer to another LC-MS/MS system, the robustness measurements of the method were investigated with regard to accuracy, precision, and specificity on the LTQ Orbitrap XL containing LC-MS/MS system regardless of the specifications indicated in Table 1 set for an Orbitrap Fusion Lumos containing LC-MS/MS system. In order to increase the

number of identified peptides, the MS/MS analysis is performed in the ion trap to enable a faster scanning throughput. Consequently, different peptide spectra are expected using adapted settings (described in experimental procedures) and therefore besides the JY samples also 500 fmol synthetic peptides of the five selected sequences AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL were spiked in JY matrix and measured on the LTQ Orbitrap XL for a spectral comparison. New five top product ions were selected based on intensity and relevance in MS/MS mode for the LTQ Orbitrap XL system. To investigate the accuracy, the purified HLA-eluted peptides from one JY batch were analyzed in three separate analytical replicates. Regarding the accuracy of the mass spectrometer in MS mode the median mass deviation from the theoretical mass of all identified natural peptides (median ΔM: -0.37 ppm) is below 2 ppm in Figure 9A similar to the Orbitrap Fusion Lumos system. The peptide RTs between the replicates of all natural peptides do have a mean Pearson correlation above 95%, demonstrating the accuracy of the LC (Figure 9B).

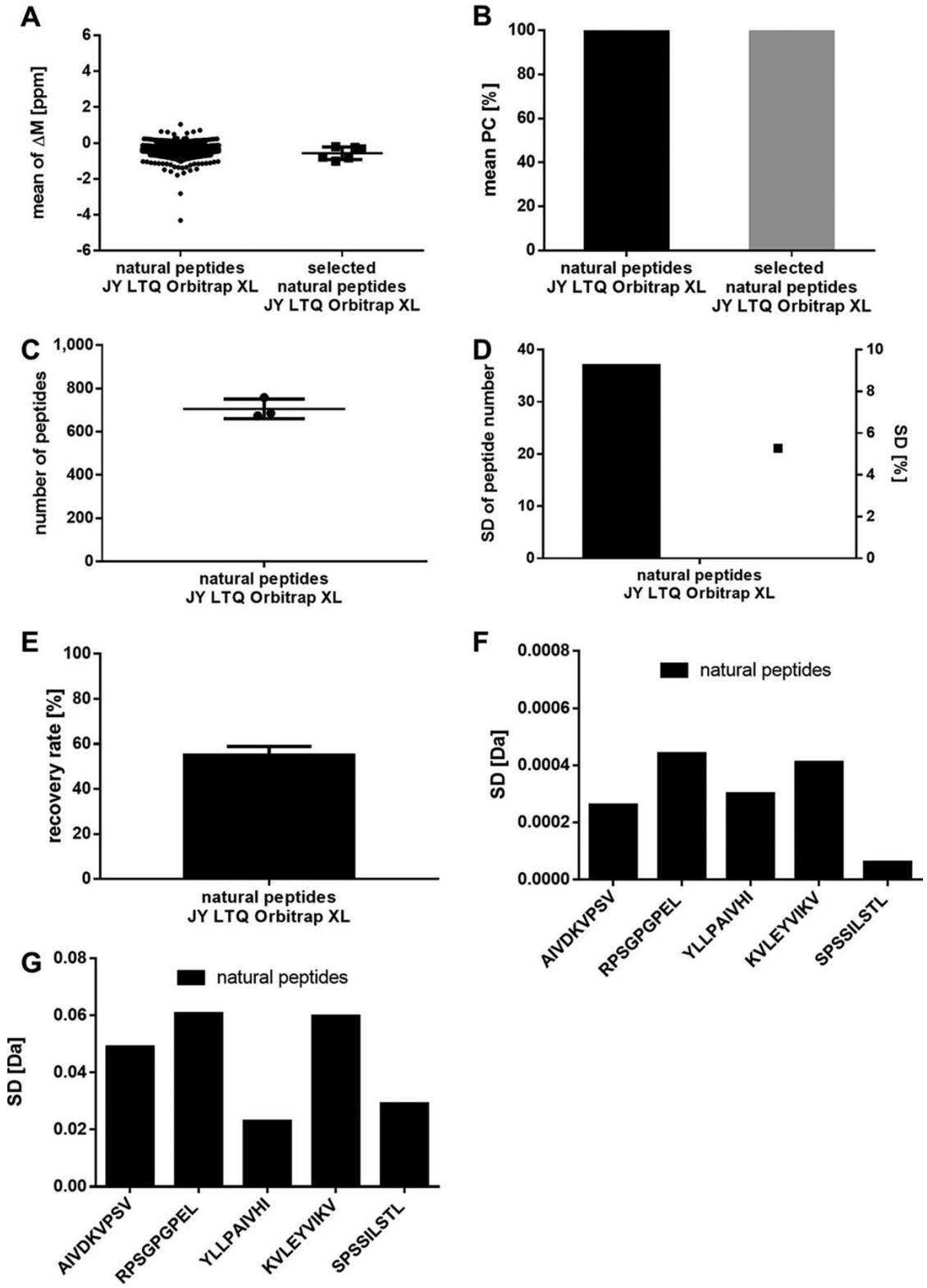


Figure 9: Investigation of the accuracy, repeatability and specificity using immunopeptidomes from JY cells analyzed after method transfer to a less sensitive LC-MS/MS system. *Three replicates were analyzed. A, Mass deviation from the theoretical precursor mass (ΔM ppm) to the theoretical mass of all identified natural peptides and of the five identified natural peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL. B, Mean Pearson correlation of the peptide retention times between the replicates for all identified natural peptides*

and the five selected peptides. C, Total number of natural peptides identified in three technical replicates, respectively. D, Standard deviation (S.D.) given in total peptide numbers and in percent. E, Recovery rate of peptide identifications between the replicates. Abbreviation: S.D., standard deviation. Mass deviation as S.D. of the (F) precursor ion masses in MS mode and the (G) resulting five selected top fragments in MS/MS modes of the five selected natural and synthetic peptides. Abbreviations: PC, Pearson correlation; ppm, parts per million; ΔM, mass deviation; S.D., standard deviation.

The precision was determined by assaying the previously mentioned three technical replicates. The number of identified peptides was similar to the validated LC-MS/MS system below 10% SD with 5% SD of the repeatability on the initial day (Figure 9C, D). However, the recovery rate with a recovery of 55 ± 4% of identified peptides in a repeated replicate is below the specifications of the Orbitrap Fusion Lumos system (Figure 9E).

The specificity was again investigated using the mass deviation from the theoretical mass of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL. The median of the identified natural peptides is ΔM: -0.57 ppm (Figure 9A) (for identified peptides and product ions, see supplemental Table S7).

The peptide RTs between the replicates of the five selected peptides identified as natural and synthetic peptides do have a Pearson correlation above 95% (Figure 9B).

The standard deviation (SD) of the mass accuracy of the precursor masses in MS mode (Figure 9F) and of the five selected top product ions in MS/MS mode (Figure 9G, for MS/MS spectra, see supplemental Figure S3), selected based on intensity and relevance, of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL is below 0.001 Da in MS Mode, analyzed in the Orbitrap, and below 0.1 in MS/MS mode, analyzed in the ion trap, for the natural peptides.

3.7 Conclusion/Discussion

To provide reliable biomarker and patient-individual tumor-associated target antigen identification for clinical studies, the fast and sensitive LC-MS/MS assay for the identification of natural and synthetic HLA-restricted peptides was validated for the technical equipment of our laboratory, consisting of a nanoUHPLC, UltiMate 3000 RSLCnano on-line coupled to an Orbitrap Fusion Lumos mass spectrometer. The immunopeptidomics pipeline is used for identification and impurity detection, thus according to FDA and EMA guidelines a validation of the specificity and LOD is required. Additionally, we validated the accuracy, precision, and robustness to demonstrate the reliability of the pipeline.

The results of the JY samples spiked with isotope labeled synthetic peptides enabled verification of the accuracy of the LC-MS/MS system in terms of similarity of the analyzed natural peptides to the theoretical and synthetic peptide masses. With the same dataset we were able to identify five specific peptides, expected as natural and synthetic peptides, which fulfil the acceptance criteria of the accuracy and could prove the specificity. Furthermore, we could show that with five selected MS/MS product ions all identified peptides within one replicate can be distinguished. Instead of picking simply the five most intensive product ions for therapeutic peptide candidates, our quality control selects the top five product ions also according to meaningfulness (expert review: b- or y-ions are preferred), thus further increasing specificity. The validation of the precision showed a reliable identification of peptides with a uniform recovery rate proving the repeatability and intermediate precision after one week.

A major limitation of MS-based data-dependent acquisition (DDA) discovery approaches is the low recovery rate. In our immunopeptidomics experiments a recovery rate of 80% ± 20% was achieved for cell lines and tissue samples, owing to the tissue heterogeneity and high dynamic range. Due to the recovery rate, in our lab routinely triplicate measurements are performed. At the LOD a reasonable reliability should be provided with a recovery rate of 50%, when triplicate measurements are performed. A peptide content of 10 fmol synthetic peptides in JY matrix enabled a reliable identification of 50% of the peptides. An improvement of the recovery rate might be obtained with a replacement of the DDA analysis with data-independent acquisition, which has demonstrated a superior reproducibility[149–153].

In order to prove the robustness of the immunopeptidomics assay, we synthesized a large variety of known HLA ligands, with different length, mass, grand average of hydropathicity (GRAVY), theoretical isoelectric point (pI) and HLA allotype restriction. In addition, we employed several primary, clinically relevant samples in addition to the JY cell line and further analyzed soluble peripheral blood mononuclear cells from a healthy donor, a soluble chronic lymphocytic leukemia and a solid bladder cancer sample. Lastly, for the three primary samples the HLA immunoaffinity chromatography and immunopeptidomics analysis was performed from three different persons. We could successfully verify the specifications of the accuracy, precision, and specificity for both the mass spectrometer and LC, respectively.

Besides the robustness of the method, we exemplarily further investigated for precision, accuracy and specificity after the method transfer to a LC-MS/MS system utilizing an LTQ Orbitrap XL. However, in order to obtain a high number of identified peptides, the MS/MS analysis is performed in the ion trap, instead of the Orbitrap, to enable a faster scanning throughput. Consequently, the mass accuracy of the product ions in the MS/MS analysis varies already at the second decimal and for the previously selected top five ions, two new ions had to be defined for

SPSSILSTL and one new ion for the other peptides, except of AIVDKVPSV. Furthermore, the recovery-rate is much lower using the LTQ Orbitrap XL.

The specifications of our validated Orbitrap Fusion Lumos based LC-MS/MS system are not met by the LTQ Orbitrap XL. In order to fully validate the method on the LTQ Orbitrap XL for peptide identification according to the current FDA and EMA guidelines, at least the analyses regarding specificity and LOD have to be performed or, ideally, all mentioned parameters should be investigated. The specifications of the precision, specificity and LOD have to be adapted to the capabilities of the less sensitive mass spectrometer. However, the specifications should be tight enough to recognize immediately any functional problems with the LC-MS/MS system.

To enable standardization between different MS platforms and laboratories, the same immunopeptidome batch from one cell line should be used for immunopeptidomics. Here we present how the parameters accuracy, specificity, LOD, precision and robustness can be investigated for the validation of LC-MS/MS systems. However, every LC-MS/MS system in every laboratory needs to be validated independently with its own specifications[142]. For the validation the specifications should be adapted as closely as possible to the optimal performance of the respective LC-MS/MS system, but should also consider the harmless device-related performance variations. Furthermore, it must be considered that more sensitive devices can detect peptides in lower quantities that less sensitive devices cannot identify. Here, the Orbitrap Fusion Lumos discovers twice the number of peptides.

The validation is performed in a new process to show that the previously specified requirements (acceptance criteria) are reproducibly met in practical use and that the analytical method is appropriate for its intended use. After the validation a system suitability test (SST) is used to continuously monitor the performance of the instrument in different fixed intervals, to verify that an analytical method is suitable for the intended purpose on the day of analysis. Control peptides known from experience to be reliably identified in the respective immunopeptidome of the cell line in higher quantity should be selected for this purpose. The identification and the retention time of these peptides in the immunopeptidomes could be routinely checked in the LC-MS/MS analysis. The selected peptides can be standardized between different suitable LC-MS/MS systems and laboratories. The scope of each LC-MS/MS validation and the SSTs can prove the suitability and comparability of different laboratories for a particular analysis. The immunopeptidomic pipeline is currently in use for the identification of tumor-associated target antigens of multiple patients in the peptide vaccination study iVAC-CLL01(NCT02802943) and for the preparation of further studies (e.g., PepIVAC01). Due to the peptide recovery rate of 80±20% per replicate, patient samples are routinely analyzed in triplicates to ensure a high recovery. Finally, candidate peptide antigens are always verified using synthetic peptides to exclude false positives and

artefacts. The identification procedure of tumor-specific HLA ligands using a comparison of ligand source proteins to established tumor antigens or ligand mapping on different malignant and benign tissues was carried out for cell lines and various tumor entities[124,143,154–158]. During the last decades and in ongoing studies the immunopeptidomics pipeline has demonstrated its reliability and applicability.

In addition, we have now validated the accuracy, precision, specificity, LOD, and robustness in line with the current FDA and EMA guidelines. This validated pipeline enables the reliable identification of tumor-associated HLA-presented target antigens to support current and future clinical studies. Furthermore, different validation approaches are presented, that can be translated to other laboratories with similar equipment, or to other MS-based discovery approaches, such as proteomics, metabolomics, and lipidomics.

3.8 Acknowledgements

This work was supported by the German Cancer Consortium (DKTK) and the Natural and Medical Sciences Institute at the University of Tübingen NMI. We thank the Wirkstoffpeptidlabor, especially Patricia Hrstić, Ulrich Wulle, Nicole Bauer, Camille Supper, and Mirijam Bohn for expert peptide synthesis and quality control.

3.9 Data availability

The mass spectrometry data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE[159] partner repository with the dataset identifier PXD012797.

3.10 Supplementary data

Supplementary data files of the final publication in Molecular & Cellular Proteomics can be accessed via:

https://www.mcponline.org/content/19/3/432/tab-figures-data

Supplementary data files of the previous version in bioRxiv can be freely accessed via:

https://www.biorxiv.org/content/10.1101/821249v1.supplementary-material

3.10.1 Supplementary Tables

Supplemental Table S1: *Selected individual synthetic peptides for the validation of the LC-MS/MS method. The 62 synthetic isotope labeled peptides consist of different amino acid sequences and physicochemical properties and depict a cross-section of peptides identified in a typical immunopeptidomics experiment. For the assessment of the differences between the peptides, the grand average of hydropathicity (GRAVY), the theoretical isoelectric point (pI) under physiological conditions, and the molar mass of the peptides were calculated using ProtParam (https://web.expasy.org/protparam). Heavy-labeled amino acids are indicated in bold (leucine +7 Da, proline and valine +6 Da). Abbreviations: pI, isoelectric point, GRAVY, grand average of hydropathicity.*

Batch	Sequence	Molecular weight	pI	GRAVY
174121	DG**P**SSAPATPTK	1,128.20	5.84	-1.00
174013	SLNSN**V**YDV	1,010.07	3.80	-0.13
174014	FPHLPGKTF**V**Y	1,305.54	8.60	0.08
174015	S**V**LTPLLLR	1,011.27	9.47	1.31
174085	RYQA**L**FHDF	1,196.33	6.74	-0.53
174122	G**P**SSAPATPTK	1,013.12	8.75	-0.77
174123	F**P**HLPGKTF	1,043.23	8.76	-0.22
174124	FLL**P**AGWIL	1,029.29	5.52	1.96
174125	LL**P**AGWIL	882.11	5.52	1.85
174126	T**P**LLLRGL	882.11	9.41	1.00
174017	AI**V**DKVPSV	927.11	5.88	1.01
174018	YLLPAI**V**HI	1,038.30	6.74	1.83
174019	**V**YVVGTAHF	992.14	6.71	1.29
174020	GTY**V**SSVPR	965.07	8.75	-0.19
174021	TYQE**V**AQKF	1,113.24	5.66	-0.84
174022	SPQGR**V**MTI	988.17	9.47	-0.10

Batch	Sequence	Molecular weight	pI	GRAVY
174086	ITDSAGHI**L**Y	1,089.21	5.08	0.47
174087	RVYGG**L**TTK	994.16	9.99	-0.43
174088	A**L**KTGIVAK	900.13	10.00	0.80
174089	SV**L**NLVIVK	984.25	8.47	1.83
174090	D**L**IIKGISV	957.18	5.84	1.43
174127	NTDS**P**LRY	965.03	5.84	-1.51
174128	R**P**SGPGPEL	909.01	6.00	-1.18
174023	DLKEKKE**V**V	1,087.28	6.18	-1.11
174024	TLHDQ**V**HLL	1,075.23	5.90	0.17
174061	N**V**GGLIGTPK	955.12	8.75	0.16
174062	FYFPTPT**V**L	1,084.28	5.52	0.86
174063	RSYHLQI**V**TK	1,244.46	9.99	-0.54
174064	NPKAFFS**V**L	1,022.21	8.75	0.62
174065	NPS**V**REFVL	1,060.22	6.00	0.12
174066	**V**LVDQSWVL	1,058.24	3.80	1.28
174068	APDAKSF**V**L	947.10	5.88	0.51
174069	K**V**LEYVIKV	1,090.37	8.50	0.92
174070	G**V**YDGREHTV	1,132.20	5.32	-0.91
174071	K**V**LEHVVRV	1,078.32	8.75	0.61
174072	ALDEK**V**AEL	987.12	4.14	0.11
174073	G**V**YDGEEHSV	1,091.10	4.13	-0.82
174074	F**V**YGEPREL	1,109.25	4.53	-0.44
174075	NA**V**GVYAGR	906.01	8.75	0.21
174076	**V**WSDVTPLTF	1,164.32	3.80	0.68

Batch	Sequence	Molecular weight	pI	GRAVY
174077	A**V**LPLTVAEVQK	1,267.53	6.05	0.88
174078	R**V**RELAVAL	1,026.25	9.60	0.79
174079	HLTE**V**YPEL	1,100.24	4.51	-0.22
174080	S**V**LADLVTTK	1,046.23	5.55	0.82
174081	IPFSNPR**V**L	1,042.25	9.75	0.37
174082	**V**LYGPAGLGK	974.17	8.56	0.56
174083	SPS**V**SQLSVL	1,016.16	5.24	0.77
174084	**V**LYPVPLESY	1,179.38	4.00	0.59
174092	T**L**LKALLEI	1,013.29	5.66	1.49
174093	A**L**REEEEGV	1,031.09	4.09	-1.01
174094	S**L**LKFLAKV	1,018.31	10.00	1.29
174095	**L**IHFLLLK	996.30	8.76	1.93
174096	G**L**YDGREHSV	1,132.20	5.32	-0.96
174129	LAQ**P**PSGQR	953.07	9.75	-1.14
174130	F**P**SLREAAL	1,003.17	6.00	0.40
174132	S**P**SKAFASL	907.03	8.47	0.26
174205	RLLDS**V**SRL	1,058.25	9.60	0.17
174206	TYSEKTT**L**F	1,089.21	5.66	-0.56
174207	S**L**LQHLIGL	993.21	6.46	1.31
174208	S**P**SSILSTL	904.03	5.24	0.73
174210	VLLAGFK**P**PL	1,054.34	8.72	1.27
174211	LYL**P**KSWTI	1,120.36	8.59	0.32

Supplemental Table S2: *Experimental design. Information about the sample, performed analysis, number of replicates, and raw file name of each sample is provided. Abbreviations: LOD, Limit of detection; PBMC, Peripheral blood mononuclear cell; CLL, Chronic lymphocytic leukemia; BC, Bladder cancer.*

Sample	Purpose	Condition/ number	Technical replicate	Raw file name
JY cell line	Precision	1 (day 0)	1	JY_repeatability_1_1
JY cell line	Precision	1 (day 0)	2	JY_repeatability_1_2
JY cell line	Precision	1 (day 0)	3	JY_repeatability_1_3
JY cell line	Precision	2 (day 0)	1	JY_repeatability_2_1
JY cell line	Precision	2 (day 0)	2	JY_repeatability_2_2
JY cell line	Precision	2 (day 0)	3	JY_repeatability_2_3
JY cell line	Precision	3 (day 0)	1	JY_repeatability_3_1
JY cell line	Precision	3 (day 0)	2	JY_repeatability_3_2
JY cell line	Precision	3 (day 0)	3	JY_repeatability_3_3
JY cell line	Precision	1 (day 7)	1	JY_intermediate_precision_1_1
JY cell line	Precision	1 (day 7)	2	JY_intermediate_precision_1_2
JY cell line	Precision	1 (day 7)	3	JY_intermediate_precision_1_3
JY cell line	Precision	2 (day 7)	1	JY_intermediate_precision_2_1
JY cell line	Precision	2 (day 7)	2	JY_intermediate_precision_2_2
JY cell line	Precision	2 (day 7)	3	JY_intermediate_precision_2_3
JY cell line	Precision	3 (day 7)	1	JY_intermediate_precision_3_1

Sample	Purpose	Condition/ number	Technical replicate	Raw file name
JY cell line	Precision	3 (day 7)	2	JY_intermediate_precision_3_2
JY cell line	Precision	3 (day 7)	3	JY_intermediate_precision_3_3
JY cell line	LOD, specificity, accuracy	spiked 0.1 fmol synthetic peptides	1	JY_100amol_1_1
JY cell line	LOD, specificity, accuracy	spiked 0.1 fmol synthetic peptides	2	JY_100amol_1_2
JY cell line	LOD, specificity, accuracy	spiked 0.1 fmol synthetic peptides	3	JY_100amol_1_3
JY cell line	LOD, specificity, accuracy	spiked 1 fmol synthetic peptides	1	JY_1fmol_1_1
JY cell line	LOD, specificity, accuracy	spiked 1 fmol synthetic peptides	2	JY_1fmol_1_2
JY cell line	LOD, specificity, accuracy	spiked 1 fmol synthetic peptides	3	JY_1fmol_1_3
JY cell line	LOD, specificity, accuracy	spiked 10 fmol synthetic peptides	1	JY_10fmol_1_1
JY cell line	LOD, specificity, accuracy	spiked 10 fmol synthetic peptides	2	JY_10fmol_1_2
JY cell line	LOD, specificity, accuracy	spiked 10 fmol synthetic peptides	3	JY_10fmol_1_3
JY cell line	LOD, specificity, accuracy	spiked 100 fmol synthetic peptides	1	JY_100fmol_1_1
JY cell line	LOD, specificity, accuracy	spiked 100 fmol synthetic peptides	2	JY_100fmol_1_2

Sample	Purpose	Condition/ number	Technical replicate	Raw file name
JY cell line	LOD, specificity, accuracy	spiked 100 fmol synthetic peptides	3	JY_100fmol_1_3
PBMC	Robustness	1	1	
PBMC	Robustness	1	2	
PBMC	Robustness	1	3	
CLL	Robustness	1	1	
CLL	Robustness	1	2	
CLL	Robustness	1	3	
BC	Robustness	1	1	
BC	Robustness	1	2	
BC	Robustness	1	3	
JY cell line	Precision, accuracy (LTQ Orbitrap XL)	1 (day 0)	1	
JY cell line	Precision, accuracy (LTQ Orbitrap XL)	1 (day 0)	2	
JY cell line	Precision, accuracy (LTQ Orbitrap XL)	1 (day 0)	3	
JY cell line	Specificity, accuracy (LTQ Orbitrap XL)	spiked 500 fmol synthetic peptides	1	

Supplemental Table S3*: Limit of detection. List of identified peptides in four aliquots of HLA eluted ligands from JY cells spiked with 0.1 fmol, 1 fmol, 10 fmol, and 100 fmol isotope labeled synthetic peptides. The lowercased amino acids indicate the isotope labeled amino acids. Abbreviations: ΔM, mass deviation; PPM, parts per million; RT, retention time.*

Direct link:

https://www.mcponline.org/highwire/filestream/54083/field_highwire_adjunct_files/6/154045_2_supp_455925_q3txy1.xlsx

Supplemental Table S4*: Specificity. Information about the precursor mass, the mass deviation from the theoretical mass, the retention time, and the selected top five product ions is provided. Abbreviations: ΔM, mass deviation; PPM, parts per million; RT, retention time.*

Direct link:

https://www.mcponline.org/highwire/filestream/54083/field_highwire_adjunct_files/7/154045_2_supp_455926_q3txy2.xlsx

Supplemental Table S5*: Precision. List of identified peptides in nine separate analytical runs and further nine runs after one week. Abbreviations: PSM, peptide-to-spectrum matches; PEP, posterior error probability; Xcorr, cross correlation score, ΔM, mass deviation; PPM, parts per million; RT, retention time.*

Direct link:

https://www.mcponline.org/highwire/filestream/54083/field_highwire_adjunct_files/8/154045_2_supp_455927_q3txy2.xlsx

Supplemental Table S6*: Specificity in primary patient samples. Information about the precursor mass, the mass deviation from the theoretical mass, the retention time, and the selected top five product ions is provided. Abbreviations: ΔM, mass deviation; PPM, parts per million; RT, retention time.*

Direct link:

https://www.mcponline.org/highwire/filestream/54083/field_highwire_adjunct_files/9/154045_2_supp_455928_q3txy2.xlsx

Supplemental Table S7*: Specificity of the LTQ Orbitrap XL containing LC-MS/MS system. Information about the precursor mass, the mass deviation from the theoretical mass, the retention time, and the selected top five product ions is provided. Abbreviations: ΔM, mass deviation; PPM, parts per million; RT, retention time.*

Direct link:

https://www.mcponline.org/highwire/filestream/54083/field_highwire_adjunct_files/10/154045_2_supp_455929_q3txy2.xlsx

3.10.2 Supplementary Figures

Supplemental Figure S1*: Validation of the specificity using the MS/MS spectra of the natural presented peptides with the corresponding synthetic peptides of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL.*

Direct link:

https://www.mcponline.org/content/mcprot/suppl/2020/01/14/C119.001652.DC1/154045_2_supp_455919_q3txy1.pdf

Supplemental Figure S2*: Validation of the robustness and specificity using the MS/MS spectra of natural presented peptides with the corresponding synthetic peptides of AIVDKVPSV, YLLPAIVHI, GTYVSSVPR, RPSGPGPEL, SVINLVIVK and RVYGGITTK. Abbreviations: PBMC, Peripheral blood mononuclear cell; CLL, Chronic lymphocytic leukemia; BC, Bladder cancer.*

Direct link:

https://www.mcponline.org/content/mcprot/suppl/2020/01/14/C119.001652.DC1/154045_2_supp_455920_q3txy1.pdf

Supplemental Figure S3*: Investigation of the specificity of the LTQ Orbitrap XL LC-MS/MS system using the MS/MS spectra of the natural presented peptides with the corresponding synthetic peptides of the five selected peptides AIVDKVPSV, RPSGPGPEL, YLLPAIVHI, KVLEYVIKV and SPSSILSTL.*

Direct link:

https://www.mcponline.org/content/mcprot/suppl/2020/01/14/C119.001652.DC1/154045_2_supp_455921_q3txy1.pdf

4 Identification of MHC Ligands and Establishing MHC Class I Peptide Motifs

4.1 Publication and author contributions

The chapter was accepted for publication as a protocol in the book "Methods in Molecular Biology" titled "Identification of MHC Ligands and Establishing MHC Class I Peptide Motifs" [65]. Authors contributing to this work are listed below. All experiments, data analysis and manuscript writing were performed by the author of this thesis, except of the described contributions in the following lines. Moreno Di Marco cultured and analyzed the monoallelic HLA-C*01:02 transfected C1R cells as previously described [154] and Stefan Stevanović supported the project draft. All authors proofread the manuscript.

Author information:

Ghosh M[I], Di Marco M[I], Stevanović S[I]

Affiliations:

I. Department of Immunology, Institute for Cell Biology, University of Tübingen, Tübingen, Germany.

4.2 Summary

MHC class I peptide motifs are used on a regular basis to identify and predict MHC class I ligands and CD8+ T cell epitopes. This approach is above all an invaluable tool for the identification of disease-associated epitopes ranging from pathogen associated epitopes, tumor associated natural and neoepitopes to autoimmune disease associated epitopes. As a matter of fact, the vast majority of T cell epitopes discovered during the past two decades was identified by means of epitope prediction and MHC ligand identification. Here we describe the steps which are necessary to identify MHC epitopes from monoallelic and multiallelic cells and establish MHC class I peptide motifs to compose a reliable scoring matrix for epitope prediction. As an example, the ligands of monoallelic C1R cells and multiallelic peripheral blood mononuclear cell tissue will be identified and a scoring matrix for the prediction of HLA-C*01:02-presented T cell epitopes will be developed.

4.3 Introduction

Major histocompatibility complex (MHC) class I molecules play an important role in cellular immunology with impact on the growing fields of individualized medicine and cancer immunotherapy [64,160,161]. The main purpose of these protein complexes is to present peptides of intracellular origin to T cells for immunosurveillance. Only a specific subset of peptides is

showcased in such a way. Peptide precursors are mainly produced by the proteasome and are transported into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP). Further trimming by ER proteases yields peptides which are transported to the plasma membrane after being bound to MHC complexes [162–164]. All these steps are dependent on the peptide sequence and are summarized in the selected ligands presented by MHC molecules. Hence peptides presented by a certain MHC allotype share common characteristics referred to as MHC peptide motif [125]. These motifs usually determine the preferred peptide length and specific amino acid residues in certain positions called anchors [165]. In most cases the peptides comprise nine amino acids and the anchors are the second and the C-terminal amino acid [166] (Figure 5). Here we show how the amino acid distribution of naturally processed and presented ligands of the MHC allotypes HLA-C*01:02 and HLA-B*56:01 can be exploited for the elucidation of the corresponding peptide motif. We further describe the procedural steps to confirm the accuracy and validity of such a motif. Nowadays, there are multiple ways to identify MHC ligands and generate a peptide motif. Figure 1 illustrates the consecutive steps to identify MHC ligands, to generate a peptide motif and to set up a scoring matrix for epitope prediction. Here we focus on the identification of MHC ligands from mono- and multiallelic cells, which can be used to identify unknown MHC motifs.

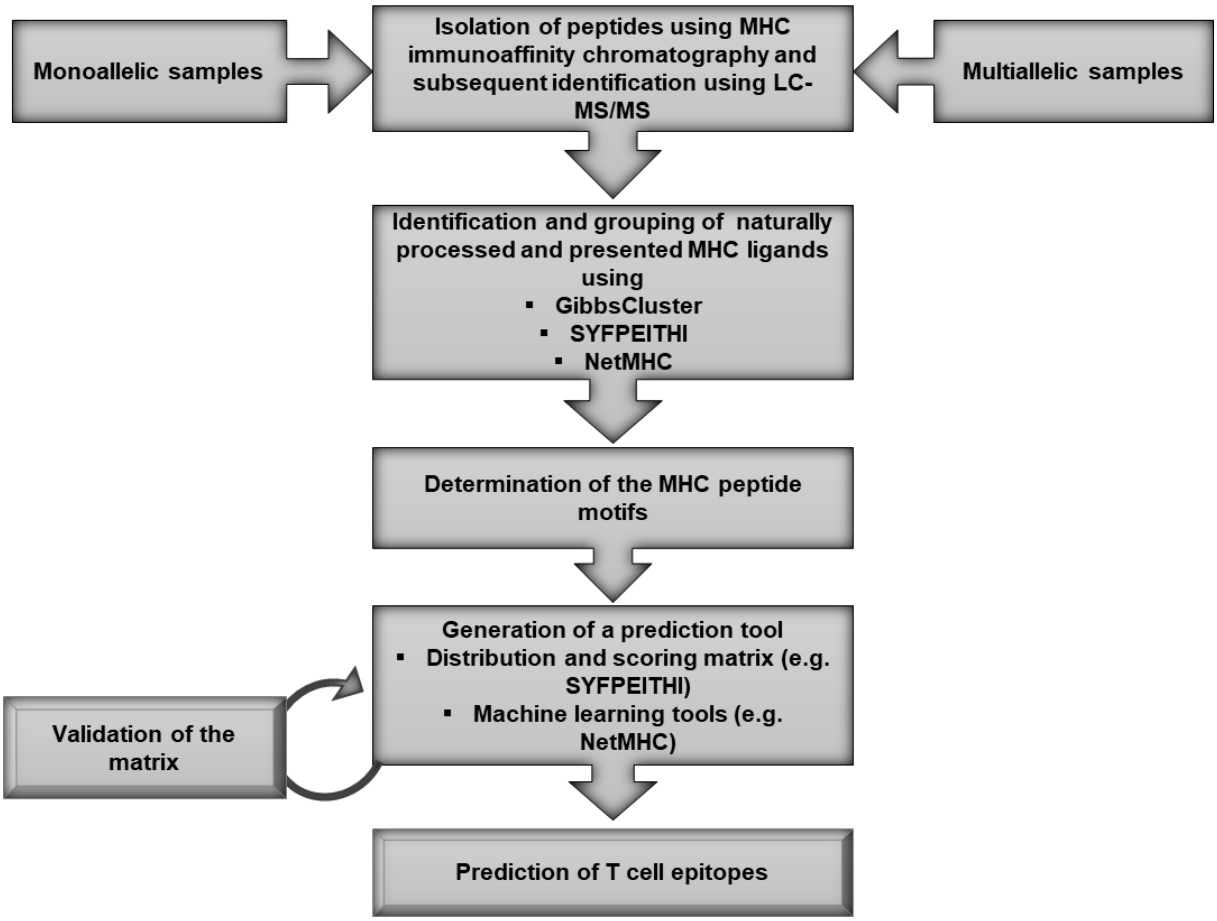


Figure 1: Workflow to identify MHC ligands and establish MHC class I peptide motifs. *Nowadays, there are several options to identify MHC ligands (see Note 5). A simple and efficient workflow using clustering is performed for the identification and successive scoring matrix generation of MHC ligands from mono- and multiallelic cells described in more detail below.*

4.4 Methods

1. Isolate at least 200 naturally presented MHC ligands of the desired MHC allotype (see Note 1).

2. Create a curated list of these ligands that contains only valid MHC ligands (see Note 2). Here GibbsCluster 2.0 is used to identify MHC class I ligands from mono- and multiallelic cells [167]. Use the recommended settings for MHC class I with one to six clusters enabled.

    a. To analyse monoallelic cell lines use GibbsCluster 2.0 with an appropriate number of clusters to remove contaminants. In case of the HLA-C*01:02 transfected C1R cell line (data provided from Di Marco et al. [154] used in our example a low expression of endogenous HLA-B*35:03 and C*04:01 is expected. At least two

clusters will be needed to separate MHC ligands and contaminants from the MHC ligands of the desired allotype (Figure 2).

b. For the analysis of MHC ligands isolated from multiallelic cells multiple clustering steps need to be performed (Note 3). In case of multiallelic cells it needs to be considered, that clustering might not separate similar motifs and MHC ligands of HLA-C, as there will be a lower number of ligands of every HLA-C compared to the HLA-A and -B clusters. Hence a repeated clustering using only ligands from the clusters consisting of multiple not deconvoluted motifs might lead to a motif separation. The MHC ligands extracted from peripheral blood mononuclear cells (PBMC) were clustered into five different motifs using two clustering approaches (Figure 2). As HLA-B*56:01 is rare, no reliable prediction motif based on eluted B*56:01 ligands is available yet. Thus, the motif can be identified excluding the other ligands of known MHC motifs.

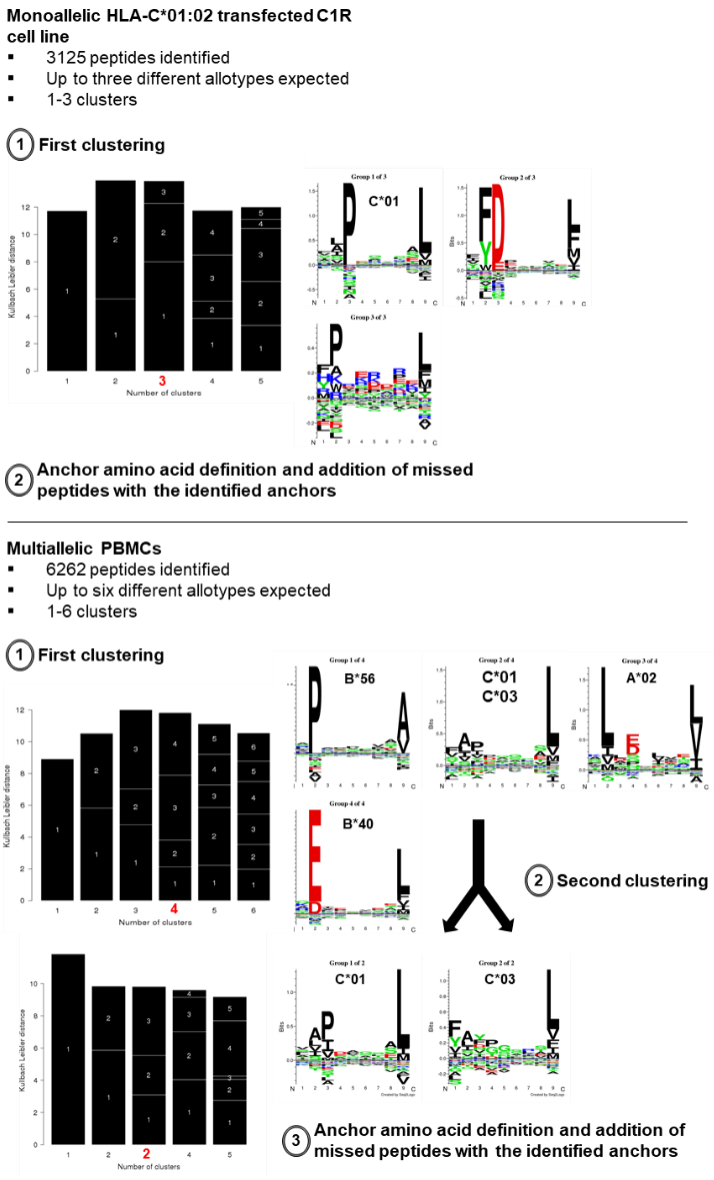


Figure 2: Identification of MHC class I ligands using GibbsCluster 2.0. *All identified peptides are uploaded into the Gibbs cluster web application (cbs.dtu.dk/services/GibbsCluster-2.0/). In case of the HLA-C*01:02 positive C1R cells, three clusters were created. Typically, the best number of clusters with the highest Kullbach Leibler distance will be suggested. However, often a lower or higher number of clusters is beneficial with an improved or no further unnecessary motif segregation. In comparison to known motifs (Figure 5), the desired C*01:02 motif can be identified. Here we generate a matrix for ligands comprising 9 amino acids (Figure 3). Subsequently, performing a filtering for the desired anchor positions (position 3: P and 9: L, M, V, I and F) will result in a list of peptides with the selected anchor positions independent of the peptide length. Using the web application Seq2motiv* [168] *the motifs can be visualized. In case of the multiallelic PBMC the first round of clustering could elucidate 4 motifs. We could identify the HLA-A*02, B*40 and HLA-C motifs based on Figure 5. Another round of clustering using the peptides of cluster 2 deconvolutes the two HLA-C motifs. The first motif is most similar with HLA-B*55. HLA allelotyping of the PBMCs resulted in HLA-A*02:01, B*40:02, B*56:01, C*01:02 and C*03:04 and confirmed the previous findings. Knowing that the donor is HLA-B*56:01 positive we can identify the first cluster as the missing B*56:01 motif. As the number of ligands of the HLA-A*02 cluster is the largest, and no further known HLA-A motif was identified, it is assumed that the patient is HLA-A*02 homozygous.*

3. Create a distribution matrix based on the relative frequency of amino acids in the different positions. Matrices should be generated for peptide lengths which make up more than

10% of the entire peptide repertoire, usually 8-10 amino acids regarding HLA class I. An example is depicted in Figure 3.

Generation of an amino acid distribution matrix

9-mer:	1	2	3	4	5	6	7	8	9
A	0.07	0.28	0	0.07	0.11	0.09	0.09	0.21	0
C	0	0.02	0	0	0	0.01	0.01	0.01	0
D	0	0	0	0.07	0.03	0.02	0.01	0.03	0
E	0.01	0	0	0.15	0.03	0.02	0.03	0.16	0
F	0.1	0.02	0	0.04	0.05	0.05	0.01	0.02	0.03
G	0.02	0.08	0	0.1	0.11	0.07	0.03	0.03	0
H	0.02	0	0	0.02	0.04	0.02	0.04	0.03	0
I	0.08	0.09	0	0.03	0.03	0.05	0.04	0	0.03
K	0.06	0	0	0.04	0.04	0.04	0.06	0.01	0
L	0.07	0.18	0	0.06	0.03	0.08	0.1	0.02	0.84
M	0.03	0.02	0	0.01	0.01	0.02	0.02	0.01	0.05
N	0.06	0	0	0.03	0.05	0.03	0.04	0.02	0
P	0	0	1	0.08	0.11	0.06	0.04	0.05	0
Q	0.02	0.01	0	0.04	0.02	0.04	0.09	0.03	0
R	0.05	0	0	0.02	0.07	0.07	0.1	0.01	0
S	0.14	0.18	0	0.12	0.12	0.09	0.17	0.21	0
T	0.08	0.05	0	0.05	0.07	0.1	0.09	0.08	0
V	0.12	0.06	0	0.06	0.04	0.09	0.05	0.05	0.05
W	0	0	0	0.01	0.01	0.01	0	0.01	0
Y	0.08	0	0	0.02	0.03	0.03	0	0.02	0

⇩

Generation of a scoring matrix

9-mere	1	2	3	4	5	6	7	8	9
A	0	3	0	0	1	1	1	2	0
C	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	1	0
F	1	0	0	0	0	0	0	0	4
G	0	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	4
K	0	0	0	0	0	0	0	0	0
L	0	2	0	0	0	0	1	0	10
M	0	0	0	0	0	0	0	0	4
N	0	0	0	0	0	0	0	0	0
P	0	0	10	0	1	0	0	0	0
Q	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	1	0	0
S	1	2	0	1	1	0	1	2	0
T	0	0	0	0	0	1	0	0	0
V	1	0	0	0	0	0	0	0	4
W	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0

⇩

Validation of the generated nonamer scoring matrix using ROC curves.

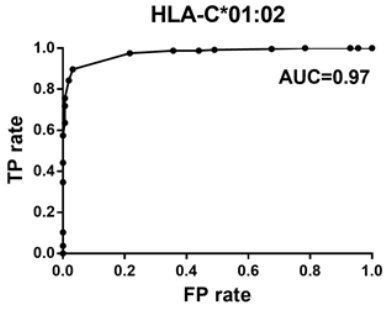


Figure 3: Amino acid distribution matrix, scoring matrix and ROC curve validation of HLA-C*01:02 ligands. *In the distribution matrix columns represent the respective position of amino acids (AA) within the peptide sequence (shown here for 9mer peptides). The numbers indicate the preference for amino acids of the respective positions indicated as percentage of occurrence. For example, 100% of the 9mer ligands carry proline (P) in position 3. Values above 10% are highlighted in grey. The distribution matrix can be used to generate a scoring matrix. A way to generate scoring matrices is presented in step 4 and Note 4. Finally, the matrix has to be validated. To visualise the performance of the matrix in step 8 it is described how to generate a receiver operating characteristic (ROC) curve. Each point of the curve represents the TP- and FP-predicted peptides using the generated scoring matrix applying thresholds from 0 to 100% of the maximal score of the matrix in 5% steps.*

4. Using the results of this amino acid distribution matrix a scoring matrix is assigned. A scoring matrix based on the frequency of amino acids in each position of the investigated MHC class I ligands represents the simplest example of such a matrix. There are numerous strategies to create a scoring matrix, including those based on automated systems (for example machine learning, support vector machines, artificial neural networks [169,170]. We will describe a procedure that can be followed without the need of bioinformatics. A method is to assign points for the different amino acids in the different positions [66]. Obviously, this can be done in various ways.

   The following approach is suggested as a guideline: Amino acids with similar chemical properties are combined in one group (see Note 4). If the relative frequency of one group in a certain position exceeds 80% it is declared an anchor. The most frequent amino acids are scored 10 points, amino acids with 1/3, 1/5, or lower frequency are awarded 8, 6 or 4 points. If a group constitutes 50-80% in one position it is declared an auxiliary anchor. The most frequent amino acids are scored 6 and amino acids below 1/2 of the most frequent amino acid are scored with 4 points. If a group constitutes 10-50% of amino acids in a specific position, 1 to 3 points can be awarded e.g. 30-49% (3 points), 20-29% (2 points) and 10-19% (1 point). These amino acids are called preferred residues. This is only one exemplary strategy to award points, which can be altered on a logical basis (see Note 4 and Figure 4). A typical scoring matrix is shown in Figure 3. Rare amino acids such as methionine can be awarded with points even if their respective frequencies do not reach the%-threshold in a specific position. The natural frequency of the amino acids in the Swiss-Prot database is depicted in Figure 4.

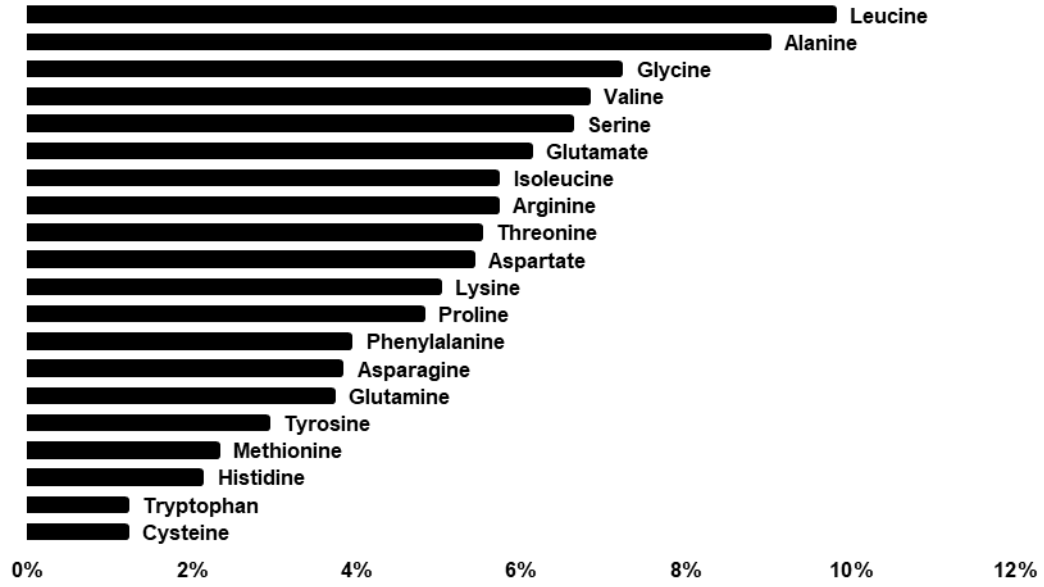


Figure 4: Amino acid frequency in the Swiss-Prot database (www.uniprot.org, release February, 2018). *Regarding the importance of every amino acid in the scoring matrix the frequency of the amino acid in the proteome should be considered.*

5. Defining anchor residues: Identify highly abundant residues in your distribution matrix. Investigate whether chemically closely related amino acids are also overrepresented. As shown in Figure 3 in position 9 (P9) leucine dominates, furthermore methionine, valine, phenylalanine and isoleucine are most present. Together these amino acids occur at 100% in P9 of the investigated MHC ligands. The amino acids share some characteristics and are therefore put in one group. Because leucine is dominant 10 points are awarded. The other amino acids in this group are less abundant and are awarded 4 points. The second anchor is proline in position 3 (P3). A vast majority of the investigated ligands (100%) contain this residue in P3. Thus 10 points are assigned to this amino acid in P3.

6. Defining auxiliary anchors: This is carried out analogously to the definition of anchor residues described in Step 5. A group of chemically related amino acids is defined as an auxiliary anchor, if its relative frequency in a specific position lies between 50% and 80% (see Note 4). In the present example (see Figure 3) no auxiliary anchors can be defined according to these criteria.

7. Defining preferred residues: All non-anchor positions are investigated for the occurrence of preferred residues. If the relative frequency of a specific amino acid exceeds 10% in one position, it is considered a preferred residue.

8. The next step is to evaluate the performance of the proposed scoring matrices. There are different approaches, here an approach from Di Marco et al. [154] is used for nonamer peptides. A 5-fold cross-validation was performed. All peptides identified in step 2a were randomly split into 5 same sized groups. Four folds are used to generate a scoring matrix (step 2a to 7), which is subsequently used to identify the ligands in the fifth fold. Clustered peptides matching to the HLA-C*01:02 motif were defined as true positives and the remaining peptides as true negatives. ROC curve analysis was performed using GraphPad Prism (version 6.00; La Jolla California USA) in 5% steps from 0 to 100% of the maximal possible score of the matrix. Figure 3 indicates the specificity and selectivity of the developed matrix. Based on the HLA motifs generated so far, a recommended AUC of 0.8 and higher should be achieved. Other methods to verify the reliability is using decoy peptides [171] or testing the ability to predict the identified peptides from their source protein sequence as shown in [172].

9. If the matrix performs poorly, the clusters have been to unspecific, and should be purified using a further clustering step. MHC class I molecules have clear anchors, which should enable a stringent epitope prediction.

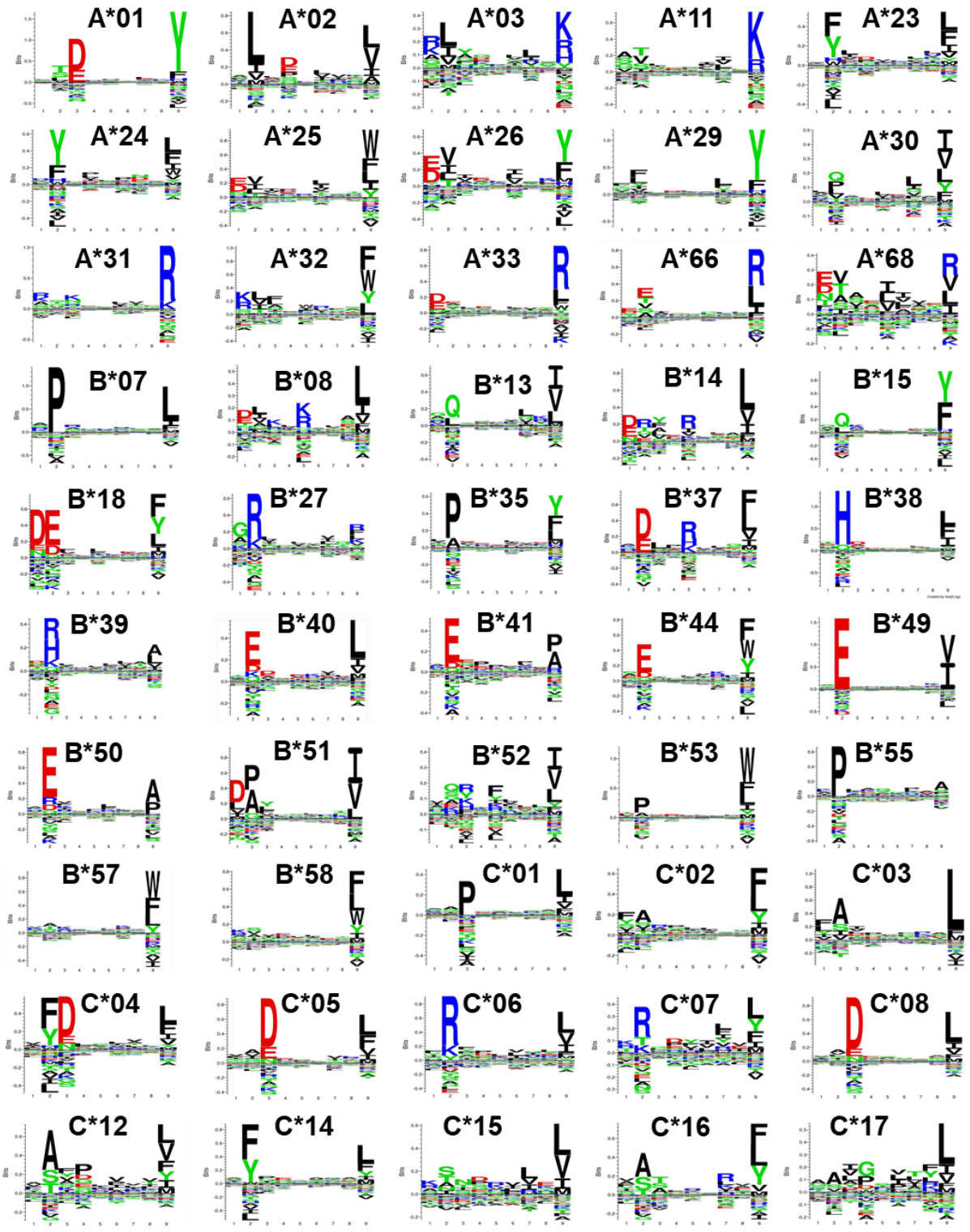


Figure 5: Summary of the most frequent known MHC motifs from primary tissue visualized using Seq2Logo 2.0 [168]***.*** *Black, aliphatic residues; green, hydrophilic residues; blue, basic residues; red, acidic residues.*

4.5 Notes

1. For a brief review of strategies for the isolation of MHC ligands and the protocol for immunoaffinity purification of MHC class I ligands used for the example at hand, see "Purification and identification of naturally presented MHC class I and II ligands" by Nelde et al. [173] in this issue. Usually the following step should be a filtering as described in Note 2 to identify the MHC ligands with a motif and exclude contaminants.

2. To validate MHC class I derived peptides they have to be verified for several criteria. The peptides should have an appropriate length for MHC ligands. It is suggested to use only peptides with 8-12 amino acids for HLA class I and 9-25 amino acids for HLA class II. Secondly, the occurrence of specific amino acids at the C-terminal position (which always serves as an anchor position in MHC class I restricted peptides) has to be evaluated. MHC class I ligands usually do not possess glycine, serine, or acidic amino acids in the C-terminal position. Furthermore, peptides derived from known contaminations or artefacts have to be excluded [174]. Depending on the tissue origin usually 60-100% of the isolated peptides are MHC ligands after a proper immunoaffinity purification of MHC class I derived peptides.

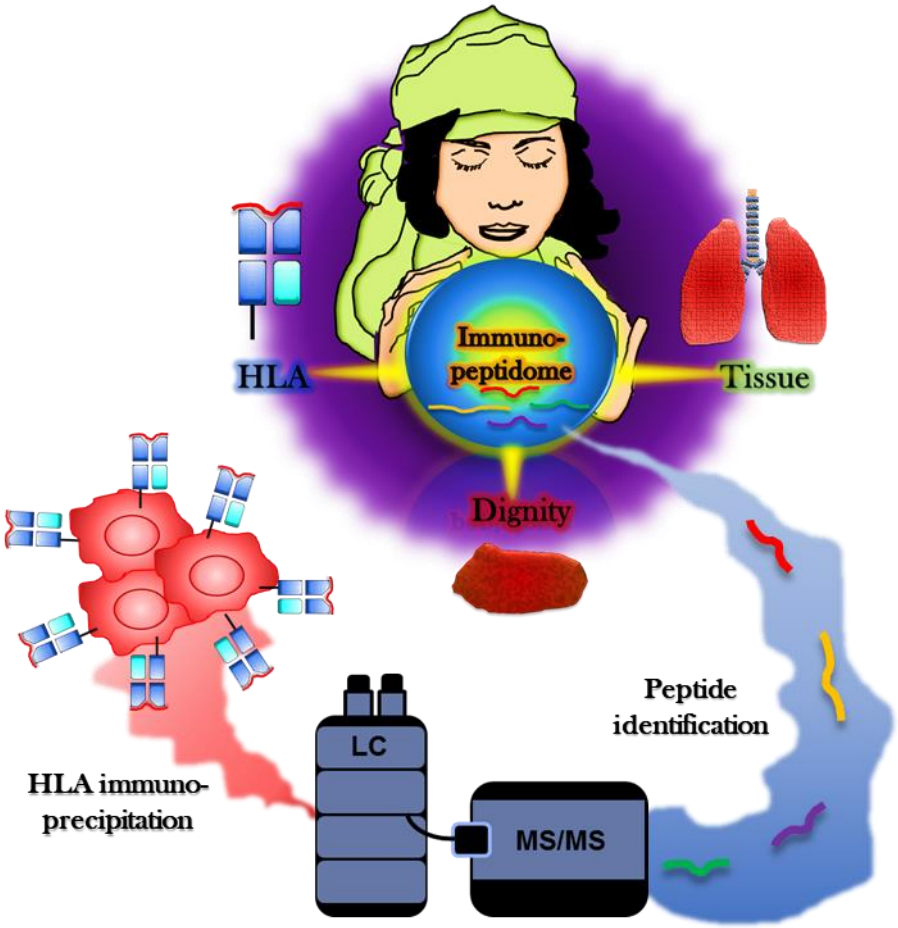
3. To identify the different motifs of the MHC allotypes, sufficient ligands of every allotype have to be eluted, which is mostly ensured only for HLA-A and B. Secondly, the clusters will be identified if the motifs have strong differences in the anchor positions. In case of missing motifs, the peptides identified in each cluster should be repeatedly clustered independently to identify, whether similar motifs have been merged and are separated in a second clustering. HLA-C molecules are lower expressed than HLA-A and -B and therefore lead to less isolated peptides. Hence, enough peptides have to be isolated first. After the first clustering the remaining peptides should be excluded from HLA-A and -B ligands and clustered independently again.

4. Amino acids are usually grouped according to their chemical properties and size. Such grouping may vary between MHC ligand pools of different allotypes. For example, aspartic acid and glutamic acid are often grouped together. However, there are several cases where only one of these amino acids is accepted in an anchor position (e.g. HLA-B*49 majorly allows glutamic acid in position 2). Sometimes even amino acids with considerably different properties may occupy the very same anchor position. For example, in HLA-B*13:02 glutamine and leucine constitute the P2 anchor position.

5. In this example GibbsCluster 2.0 was used to enable an unbiased clustering of MHC ligands and no knowledge about the HLA allotype of the sample is required. Knowing the MHC alleles of the samples, MHC ligands can be simply identified using tools such as SYFPEITHI (http://www.syfpeithi.de/index.html) and NetMHC (http://www.cbs.dtu.dk/services/NetMHCpan/). Here peptide lists can be entered directly into the web application after selection of epitope prediction in SYFPEITHI or peptide in type of input in NetMHC. Having more bioinformatical skills FRED 2 [175] and MixMHCpred [171] can be implemented, which provide additional benefits. FRED 2 enables a prediction of the best fitting MHC motif for a peptide based on e.g. all MHC provided in SYFPEITHI and NetMHC. Recently, MixMHCpred enabled a high deconvolution of 78% of HLA allotypes in 50 HLA ligandomes including rare HLA types like HLA-B*56:01 [171].

5 An innovative approach for HLA typing, molecular tumor testing and the validation of tumor exclusive antigens

5.1 Graphical abstract



5.2 Abstract

The immunopeptidome, representing all human leukocyte antigen (HLA) presented peptides, is the key for adaptive immunity. Each presented peptide holds an abundance of information not yet well understood. Up to now, the scientific focus has been on the definition of pathogenic or tumor-derived epitopes and the deconvolution of HLA peptide motifs of the entire immunopeptidome. Here we go one step further and assess the properties of individual peptides to identify defined HLA allotype-specific and frequently presented peptides. Such allotypic peptides represent a versatile tool to determine HLA allotypes or serve as internal standard for characterization of cancer antigens and differentially processed antigens. Finally, individual tissue- and dignity-specific antigens were defined, and the latter were successfully implemented for molecular tumor testing.

Using mass spectrometry-based immunopeptidomics a database was generated consisting of ~900 HLA-typed samples. The identified allotypic peptides enabled a HLA class I allotype determination, which was 95% correct in our in-house dataset and 98% in an external dataset. These abundant peptides were implemented as internal standard for a semi-quantitative investigation of established tumor antigens and antigens processed differentially in malignant and benign tissue. Defined dignity-specific antigens allowed 87% correct tumor detection across numerous tumor types.

In summary, we describe a machine learning approach for mining immunopeptidomic data in order to develop a classification method, allowing to differentiate HLA class I-allotypes of a sample or to distinguish between healthy and malignant state of tissues. Furthermore, based on this method, we developed a procedure for the validation of tumor exclusive antigens. Our results support the classification of immunopeptidomic data sets using machine learning and highlight their potential utility for biomarker development.

5.3 Introduction

The immunopeptidome comprises a vast number and diversity of human leucocyte antigen-presented peptides (ligands), providing T cells insights into inter- and intracellular processes. The ligands are constantly modulated by cellular metabolism, including gene expression, transcription, translation, posttranslational modifications, and antigen processing and presentation[117,119,176]. In the HLA class I pathway, intracellular proteins are enzymatically degraded by the proteasome and resulting peptides are transported into the endoplasmic reticulum (ER) via the transporter associated with antigen processing (TAP) protein complex. In the ER peptides can bind to HLA class I molecules and are transported to the cell surface, where CD8+ T cells can interact and recognize these HLA-peptide complexes. Such peptides are mainly

presented on the classical HLA molecules, HLA-A, HLA-B and HLA-C. HLA-A and HLA-B are most abundant, whereas HLA-C represents only about 10% of the classical HLA molecules [154].

Each HLA allotype has different peptide preferences which lead to a vast and diverse peptide presentation in each individual. HLA class I molecules present short peptides of 8-12 amino acids (AA) length. The peptides presented by an allotype can be summarized in a peptide motif [35]. Amino acids occurring with increased frequency at a defined position of the peptide sequence were identified and designated as anchors [65]. For the distinct HLA alleles various genetic associations with disease susceptibility have been established [154,177–179]. Besides genetic associations, also pathogen-derived T cell epitopes presented on infected cells have been discovered for various HLA allotypes and an array of immunogenic epitopes is available [180–182]. In cancer many endogenous tumor-specific peptides [143,157,183–187] and mutation-derived peptides have been identified and proven to be immunogenic [124,188–190]. Such immunogenic presented peptides may also be recognized by T cells in case of an autoimmune disease [191].

Today, immunopeptidomic approaches are increasingly used and the application is widely diversified. The knowledge of HLA peptide motifs has enabled epitope prediction with a high sensitivity and reliable specificity [65,69,154,167,192]. Still, current predictions are lacking in discrimination of peptides derived from HLA molecules with similar peptide motifs. Inferring the HLA allotype of a patient based on HLA-presented peptides has so far not been possible in an automated manner.

In this study we identified allotype-specific peptides for the most frequent HLA class I allotypes to enable peptide-based HLA class I allotyping. In addition, allotypic peptides were implemented as internal standard to clarify whether established tumor antigens are presented exclusively in tumors and whether certain antigens are differentially processed in tumors compared to benign tissue. Finally, we applied the machine learning strategy to investigate, whether it is possible to distinguish between tumor and benign tissue on the basis of the immunopeptidome.

5.4 Materials and Methods

5.4.1 Tissue samples and cell lines

Immediately after surgery, patient derived primary tissues from histologically confirmed non-malignant tissue samples, tumor samples and tissue from adjacent sites were snap-frozen in liquid nitrogen and stored at -80°C. In Chapter 5 these sample types will be referred to as benign, malignant, and adjacent benign and are summarized by the term dignities. From every patient, written informed consent was obtained in accordance with the Declaration of Helsinki as well as local laws and regulations. Polymerase chain reaction (PCR)-based HLA typings of cell lines and tumor samples were carried out at the Department of Hematology and Oncology, University of

Tübingen, Germany. The samples were collected by many different employees of the Department of Immunology, Tübingen, for about the last ten years.

5.4.2 Cell lines, transfection, and selection

Data from monoallelic cells were retrieved from Abelin *et al.* [126], Di Marco *et al.* [154] and in-house analyzed C1R cells. In-house generated C1R cells were transfected and selected as previously described [154], cultured up to an amount of at least 10^9 cells and harvested by centrifugation at 1500 rpm for 15 min at 4°C. After two washing steps with cold PBS, cells were collected and frozen at -80°C. The immunopeptidome analysis of the in-house generated C1R cells was carried out by many different employees of the Department of Immunology, Tübingen.

5.4.3 Isolation of HLA ligands by immunoaffinity purification

HLA class I molecules were isolated using standard immunoaffinity purification as previously described [137], using the pan-HLA class I–specific monoclonal antibody W6/32 [145] (produced in-house) to extract HLA ligands. The obtained peptide solutions were stored at -20°C until analysis by liquid chromatography–tandem mass spectrometry (LC-MS/MS).

5.4.4 Analysis of HLA ligands by LC-MS/MS

Peptides were separated by nanoflow high-performance liquid chromatography (nanoUHPLC, UltiMate 3000 RSLCnano, Dionex) and subsequently analyzed in one of two on-line coupled mass spectrometers, either an Orbitrap Fusion Lumos or LTQ Orbitrap XL (both Thermo Fisher Scientific) using data-dependent acquisition (DDA) as previously described [116].

5.4.5 Database search, spectral annotation

The acquired LC-MS/MS data was processed against the human proteome included in the Swiss-Prot database (http://www.uniprot.org, September, 2013; containing 20,279 reviewed protein sequences), applying the SequestHT algorithm [146] with Proteome Discoverer software (version 1.3 and 1.4, Thermo Fisher Scientific). Precursor mass tolerance was set to 5 ppm, product ions mass tolerance was set to 0.5 Da (LTQ Orbitrap XL)/0.02 Da (Orbitrap Fusion Lumos) and oxidized methionine was allowed as the only dynamic modification. Percolator [147]-assisted false discovery rate (FDR) calculation was set at a target value of $q \leq 0.05$ (5% FDR). Peptide length was limited to 8–12 AA.

5.4.6 Classification with random forest

We chose the random forest machine learning classification algorithm using the default parametrization (ntree = 100; Mtry = 10) of the R Statistical Computing software (v. 2.5.0) package randomForest (v. 4.6). In a 10-fold cross-validation this parametrization was determined to achieve reliable accuracy (AUC). Here, the principle of the random forest decision tree is that the leaves at the top represent class labels (HLA positive or negative) and branches represent conjunctions of features (HLA-specific peptide sequences). The path for a sample, which is drawn

by the chosen branches, decides the output of the classification (only allotypes on ≥ 10 samples were considered). The leaves between the root and the tree top represent the weighted predictors (peptides) that decide about the best split of branches [193]. The random forest analysis in Chapter 5 was carried out by Gizem Güler and Leon Bichmann.

5.4.7 Experimental design and statistical rationale

Peptide lists of every sample were exported from Proteome Discoverer and summarized with sample ID, dignity, tissue type and HLA allotype in the open-source database SQLite. Random Forest classification and AUC calculations were performed using R. In order to determine the parameters - number of samples, percentage of allotype positive samples and percentage of allotype restriction (Table 1) - peptides were exported as csv file from the SQLite and the parameters determined using Microsoft Excel. Pearson correlation and unpaired *t*-tests were as well calculated using Microsoft Excel.

5.4.8 Generation of an immunopeptidome tissue database

To obtain maximal biological diversity and robustness, a database was generated containing the immunopeptidomes of human primary tissues and cell lines covering various organs from benign, adjacent benign, and malignant tissues from various donors, tissues, and tumor entities (Supplemental Figure S1 and S2). The database contains peptide data of samples that have been analyzed from many different employees of the Department of Immunology, Tübingen, for about the last ten years. The database was primarily compiled and maintained by Ana Marcu. For HLA class I, this database comprised 1,237 different patient or cell line samples containing 892 different HLA-typed tissues and 103 HLA-typed human cell lines (all samples were at least two-digit HLA-typed (2-d-DB)) with a total of 333,431 peptides derived from 18,624 source proteins (only unique protein annotations considered). Supplemental Figure S1 depicts the distribution of samples across different tissues. The samples in the database include a total of 19 different HLA-A, 28 HLA-B and 14 HLA-C allotypes on two-digit level (Supplemental Figure S2). Several publications have emerged from this immunopeptidome database with partially deposited data in online repositories [143,157,183–187]

Additional 28 in-house analyzed samples, not included in the database, and ten additional samples from the literature [192] were used to test the HLA typing based on HLA-presented peptides.

5.4.9 Generation of four-digit peptide frequency tables for HLA-A, B and C

All four-digit typed samples from the generated database (4-d-DB), containing 498 samples, were used to generate tables for HLA-A, B and C (Supplemental Figure S3). These tables list each peptide frequency for each HLA-A, B and C allotype in the 4-d-DB and enable the determination of the parameters described in table 1. The tables were created in collaboration with Jonas Scheid.

Comparing the HLA allele frequency of the four-digit typed samples (allotype ≥ 5%) with the allele frequency net database [194] (04.2020; German pop 8; n = 39,689) the frequencies in both cohorts correlate to 97% (Supplemental Table S1).

5.4.10 Data availability

The mass spectrometry data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE [159] partner repository with the dataset identifier PXD009531 for the HLA-C and G positive monoallelic cell lines. A summary of the monoallelic samples, technical replicates and MS RAW files is provided in Supplemental Table S2.

5.5 Results

5.5.1 Higher peptide frequency increases reliability of peptide identification

First, we investigated whether a larger database with multiple peptide identifications increases the reliability of the identified peptide sequence and minimizes LC-MS/MS artefacts. In our generated immunopeptidome database, the majority of samples was analyzed in three technical LC-MS/MS replicates. Using HLA-B*15:01-transfected C1R cells we could verify that three technical replicates of ⅕ of the total eluted peptides suffice to characterize most identifiable peptides (Supplemental Figure S4; 8%±2% novel peptides in the third replicate). The analysis of 30 replicates (three biological x ten technical replicates) reveals that a large part of the peptides (26%) was only found in one replicate (Supplemental Figure S5). A comparison of the q-value and x-Corr of the peptides found in one or 30 replicates clearly shows that the minimal false discovery rate at which the identification is considered correct and goodness of fit of experimental peptide fragments to theoretical spectra is significantly better in the latter (Supplemental Figure S5). In order to avoid measurement and processing artefacts, only peptides sequenced in ≥5 samples were selected for the database analysis (except for differential antigen processing, Supplemental Figure S23).

5.5.2 Characteristics of HLA-presented peptides

In the next step, we investigated how many of the HLA-A, B and C presented peptides are exclusively restricted to the respective HLA allotype and how many are shared by several allotypes. We compared ligands eluted from monoallelic HLA-transfected C1R cells (internal and external data to cover more allotypes) with the 4-d-tables (see Materials and Methods). We determined all HLA ligands of the transfected allele (workflow in Supplemental Figure S6) and mapped these ligands to the 4-d-tables (Figure 1; peptide sequences in Supplemental Table S3).

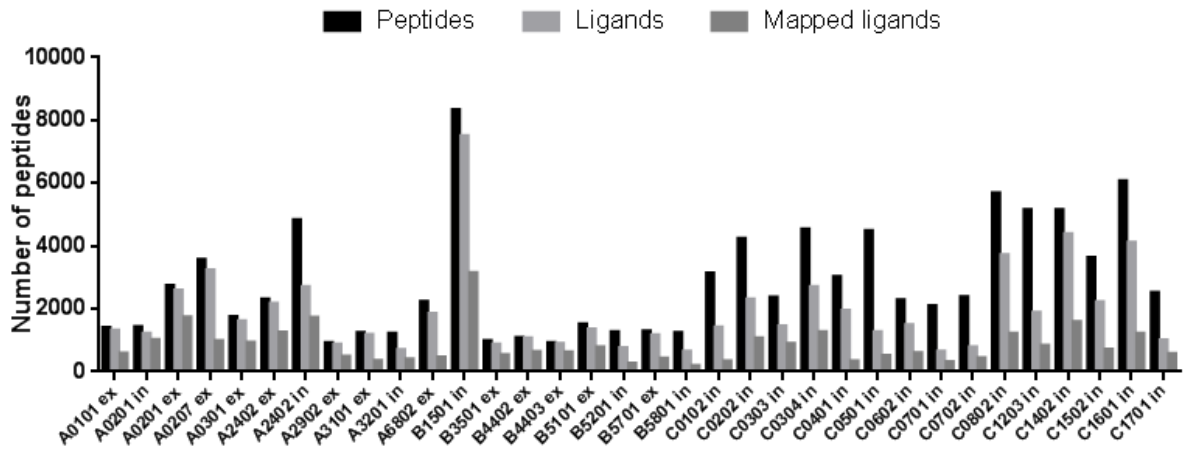


Figure 1: Peptides, HLA ligands and mapped ligands of monoallelic cell lines. *The total number of peptides identified, the peptides matching to each HLA peptide motif and the peptides mapped to the 4-d-DB of the internal (in) and external (ex) monoallelic cells are indicated.*

The peptide motifs of the mapped ligands matched the known peptide motifs of the respective HLA allotypes (Supplemental Figure S7). Most presented peptides of HLA-B*35:03 and C*04:01 endogenously expressed in C1R were removed. Regarding the peptide length distribution, nonamers were preferred for most allotypes. Only in the case of HLA-B*52:01 the octamers did predominate and also HLA-B*51:01 had a high proportion of 31% octamers (Supplemental Figure S8).

We were able to determine the parameters in Table 1 for the mapped ligands. The average number of samples on which ligands eluted from one allotype were found (mean number of samples, Figure 2), the average percentage of these samples that were positive for this allotype (mean %allotype positive samples, Figure 3) and the difference of this percentage between the allotype with highest percentage and the second highest (mean %allotype restriction, Figure 4).

Table 1: Parameters to assess the peptide relevance in the immunopeptidome. *Parameter and equation used for calculation.*

Parameter	Feature
Number of samples	• Number of samples presenting the peptide and positive for the allotype
%allotype positive samples	• Percent of allotype positive samples of all peptide presenting samples
%allotype restriction	• Difference of the specificity from the most specific to the second most specific allotype

When considering the parameters, we wanted to examine not only the average of all mapped ligands, but also explicitly the top ten peptides for each allotype and parameter. This allows to

determine whether low values for a parameter of the mapped ligands of an HLA allotype are valid for all peptides or whether individual peptides are presented that achieve higher values.

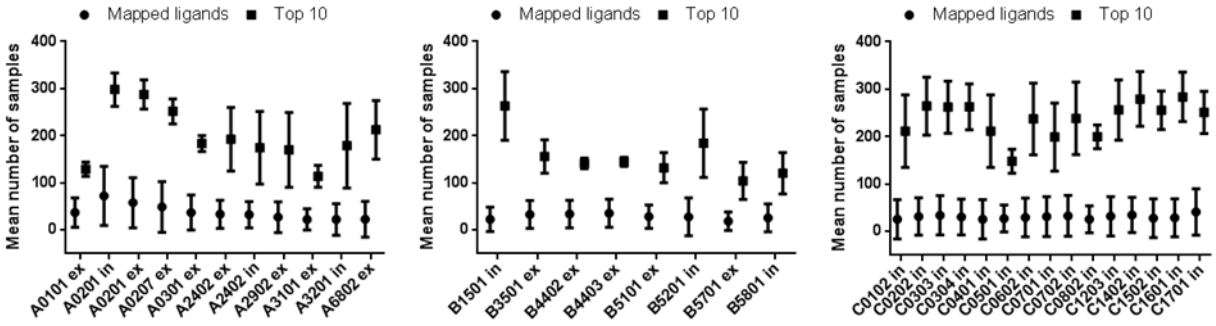


Figure 2: Average number of ligand positive samples. *The mean number of samples on which ligands eluted from one allotype of the internal (in) and external (ex) monoallelic cells were identified for HLA-A, B and C allotypes. The overall number of mapped ligands was considered as well as the ten peptides with the highest number of ligand positive samples per allotype (Top 10).*

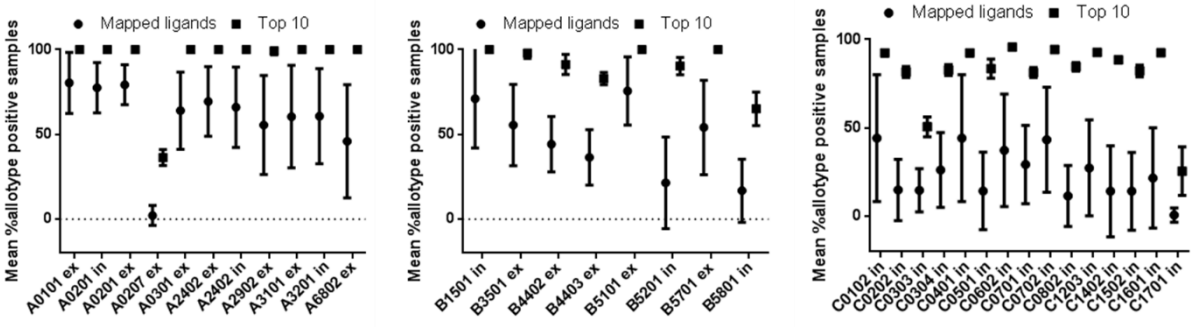


Figure 3: Average percentage of allotype positive samples*. The mean percentage of allotype positive samples on which ligands eluted from one allotype of the internal (in) and external (ex) monoallelic cells were found for HLA-A, B and C allotypes. The overall number of mapped ligands was considered as well as the ten peptides with the highest number of ligand positive samples per allotype (Top 10).*

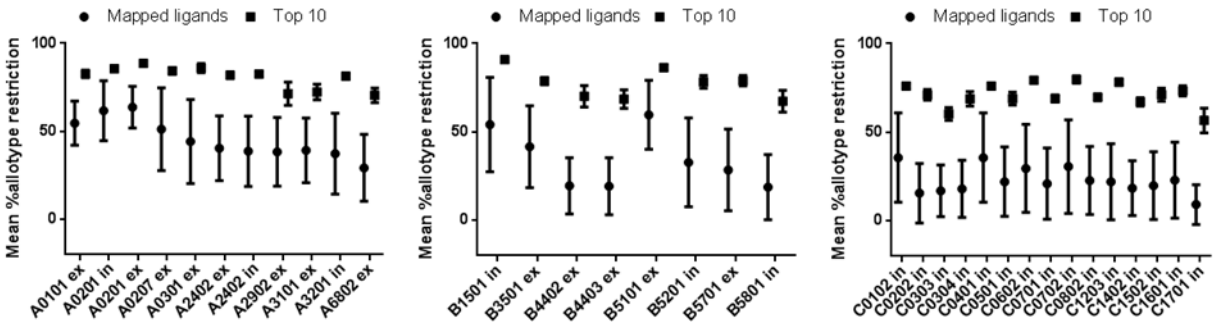


Figure 4: Average allotype restriction*. The mean allotype restriction is indicating the difference of the mean %allotype positive samples of the best fitting and the next highest allotype for the ligands eluted from one allotype of the internal (in) and external (ex) monoallelic cells for HLA-A, B and C. The overall number of mapped ligands was considered as well as the ten peptides with the highest number of ligand positive samples per allotype (Top 10).*

Despite the large disparities in the allotype frequency in the database, the differences between the mapped ligands of the allotypes appeared to be rather small regarding the three analyzed parameters. However, deviations between allotypes were observed particularly in the top ten peptides.

The average number of samples of the top ten peptides was consistent with the allotype frequency in the database. The top ten peptides of HLA-A*02 and -B*15 were identified in most samples. However, this was not the case with HLA-C. Against expectation HLA-C*07 did not stand out. This was probably due to fewer HLA-C molecules on the surface, resulting in a lower number of eluted peptides that can only be identified with sensitive LC-MS/MS, especially in larger samples or samples with high peptide presentation.

For the mean %allotype positive samples there were less differences between the top ten scores. An exception were the subtypes HLA-A*02:07, C*03:03 and C*17:01 which did not present highly specific peptides. Most of the HLA-A*02:07 presented peptides were also detected in A*02:01 positive samples. Many HLA-C*03:03 presented peptides were also identified on C*03:04. The HLA-C*17:01 presented peptides had a peptide motif that resembles several allotypes such as C*02:02, C*03:03/04, C*12:03 and C*16:01.

Concerning the %allotype restriction, HLA-A*02:07, C*03:03 and C*17:01 still reached high percentages. This is due to the calculation in Table 1, comparing the best allotype with highest %allotype-positive samples with the second best allotype. In these cases, the %allotype restriction is still high because the peptides of HLA-A*02:07 are also presented on A*02:01, the peptides of C*03:03 also presented on C*03:04 and the peptides of C*17:01 also presented on the allotypes C*02:02, C*03:04, C*12:03 and C*16:01. In contrast to the top ten, the mean %allotype restriction

of the total number of mapped ligands varied greatly. HLA-A*01, A*02, B*15 and B*51 mapped peptides were the most allotype-restricted. In general, the selected HLA-A allotypes, except HLA-A*02:07, seemed to have the most allotype-specific peptides. For HLA-B the overall peptide specificity was lower, especially for HLA-B*52:01 and B*58:01. In case of HLA-C the peptides seemed to be poorly specific.

5.5.2.1 Characteristics of allotype classification peptides

In order to identify allotypic peptides, which enable a classification of HLA allotypes, we applied the random forest (RF) algorithm to the 2-d-DB (as described in the method section). The calculated area under the curve (AUC) for each allotype was determined to estimate the performance of the trained RF model using the 2-d-DB and led to an AUC >90% (mean AUC: HLA-A = 95% ± 5%, HLA-B = 94% ± 8%, HLA-C = 93% ± 8%; maximal AUC: HLA-A*01 = 96% ± 3%; minimal AUC: HLA-C*17 = 93% ± 8%) indicating high specificity and sensitivity.

We determined the most weighted predictors for the allotypes - the 20 top ranking classification peptides (Top20_CP) - to investigate whether the top peptides for the classification were rationally reasonable (Top20_CP list provided in Supplemental Table S3) and, compared to the ligands eluted from monoallelic cells, have above average values near to the top ten values depicted in Figure 2-4. An alignment of the Top20_CP (generated from the 2-d-DB) with the 4-d-tables (generated from the 4-d-DB) revealed that the Top20_CP had a high consensus of sample frequency (number of peptide positive samples), specificity (%allotype positive samples) and restriction (%allotype positive samples) for the respective allotype as depictured for each individual peptide in Supplemental Figure S9-11. Similar to the monoallelic cells, the C*17 allotype was outstanding with poorly specific and restricted classification peptides.

A closer look at the peptide motifs of the Top20_CP demonstrated that the peptides of most HLA-A and B had homogeneous peptide motifs and match the known motifs of the allotypes (Supplemental Figure S12). In the case of all HLA-C, except for C*01, *05, *14 and *16, there was a relevant contamination of peptides with motifs matching other HLA allotypes. This was mainly related to the linkage disequilibrium. For an overview of alleles that could be inherited via a possible HLA linkage disequilibrium, a heat map for HLA with more than five positive samples in the 4-d-DB was created at two-digit level, showing how frequently the combination of two alleles occur in a sample of the 4-d-DB. When comparing the alleles, the allele frequencies of the German population (allele frequency net database [194], 04.2020; German pop 8; n = 39,689) were subtracted from the calculated frequency, resulting in the linked alleles (Supplemental Figure S13). Most HLA-C were coexpressed with other HLA-B alleles on the samples, for example in the form of peptides having a HLA-B*07 motif in the Top20_CP of HLA-C*07.

5.5.2.2 Characteristics of allotype classification peptides excluding linkage disequilibrium

To further characterize the top 20 allotypic peptides from each HLA, all peptide contaminants of linkage disequilibrium were removed. Among all RF predicted peptides, the 20 peptides with the highest RF score of all peptides matching the peptide motif of the allotype were determined using GibbsCluster-2.0 [167] or the peptides eluted from monoallelic transfected cell lines as described in Supplemental Figure S14. The top 20 allotype-associated peptides, devoid of peptide motifs characteristic for other allotypes, will be referred to as Top20 associated peptides (Top20_AP; Top20_AP list is provided in Supplemental Table S3). Especially in the case of HLA-C*07, C*08 and C*17 the Top20_AP, matching the peptide motifs, were distributed in the RF ranking from the first to the thousandth position. All linkage disequilibrium derived peptide contaminations in the Top20_CP peptide motifs were successfully removed in the filtering to the Top20_AP (Supplemental Figure S15).

The peptide length distribution of the Top20_AP was similar to that of C1R cells. In some allotypes a higher proportion of octamers as in HLA-B*52 or deca- and undecamers as in B*35 was found besides the typical nonamers, which matches the literature [71] (Supplemental Figure S16).

An analysis of protein abundance and protein turnover by overlapping Top20_CP and AP source proteins with proteome studies of Hela cells [195] demonstrates that the proteins cover the entire range of protein abundance and protein turnover of cytoplasmic proteins (Supplemental Figure S17). A closer look at the UniProt keyword annotations of the source proteins of the Top20_AP resulted in 69% cytoplasm annotations, but also 57% nucleus (Supplemental Figure S18; complete data in Supplemental Table S4). About 20% of the source proteins were involved in host-virus interaction.

5.5.3 Allotypic peptides implemented for HLA allotyping

In order to test whether a simple HLA typing based on the Top20 peptides is possible and can provide comparable results to the RF typing, HLA-classifications of 28 internally (17 malignant tumors, 5 adjacent benign and 6 benign samples) and ten externally analyzed malignant samples [196] (processed as described in the methods section) were performed using the RF algorithm, Top20_CP and AP. The cut-off was a minimum of three peptides of the Top20 peptides per allotype that have to be identified. Two internal samples contained allotypes not considered in the training data set (HLA-A*66 and B*47), as there were less than ten positive samples available (Supplemental Table S5).

HLA allotyping at peptide level using the RF algorithm enabled the correct typing of over 80% of the allotypes in the internal and external data set with a maximum of 3% mistyped alleles (Table 2). Typing of more than six allotypes was rare and the greatest limitation was the absence of almost 20% of the allotypes. Surprisingly, allotyping using the Top20_CP reduced the missing

alleles to a maximum of 5%. At the expense of false typing and overtyping, over 95% of the allotypes in the data sets could be typed correctly (98% in the external dataset). The Top20_AP could not keep up with the Top20_CP in typing.

Table 2: HLA allotyping performance. *HLA allotyping at the peptide level using the RF algorithm, Top20_CP and AP. A summary of all correct, false, missing, and overtyped allotypes of the 38 novel internal and external samples is given.*

Allotyping	Dataset	Correct	False	Missing	Overtyping
RF	Internal	83%	3%	17%	2%
	Internal#	81%	2%	18%	2%
	External	83%	2%	15%	0%
Top20_CP	Internal	96%	10%	3%	9%
	Internal#	95%	11%	5%	10%
	External	98%	13%	2%	13%
Top20_AP	Internal	94%	21%	5%	19%
	Internal#	93%	23%	7%	20%
	External	95%	17%	5%	17%

#containing HLA allotypes not considered in the RF training, Top20_CP or AP

When considering all allotypes per sample the overtyping of the Top20_CP seemed to result in no more correctly typed samples compared to the RF algorithm (Table 3).

Table 3: Sample typing performance*. HLA allotyping of each sample at the peptide level using the RF algorithm, Top20_CP and AP. A summary of all six correct and false typed HLA allotypes per sample of the 38 novel internal and external samples is given.*

		False allotype			
Allotyping	Dataset	0	1	2	>2
RF	Internal#	52%	31%	21%	17%
	External	60%	20%	0%	20%
Top20_CP	Internal#	59%	10%	17%	14%
	External	40%	30%	30%	0%
Top20_AP	Internal#	10%	55%	14%	21%
	External	30%	30%	20%	20%

#containing HLA allotypes not considered in the RF training, Top20_CP or AP

An advantage of using the Top20_CP was that the typing could be subjectively improved. If more than two allotypes per HLA locus were suggested and the allotype with the fewest peptides from the respective Top20_CP list was omitted, four samples of the internal (AML1, Mng1, MaCa1 and 2) and external (CD165, Mel12, 15 and 16) data set were typed without error. Thus 72% of the internal and 80% of the external samples would be typed without errors. The three

most challenging samples (CLL1; Mng1 and Sarc1) were tumor samples in which too many additional allotypes were suggested. A more precise examination of the allotyping of the alleles in the different dignities (benign, adjacent benign, malignant; exclusion of allotypes not considered in the RF training) showed that the overtyping occurred mainly in tumor samples (16% of the malignant samples; Table 4).

Table 4: HLA allotyping performance per dignity. *HLA allotyping at the peptide level using the RF algorithm, Top20_CP and AP for samples of each dignity (benign, adjacent benign and malignant tissue). A summary of all correct, false, missing, and overtyped allotypes of samples of each dignity of the 28 novel internal samples is given. The samples with infrequent HLA, not considered in the RF training, are excluded.*

Dignity	Dataset	Correct	False	Missing	Overtyping
Benign	Internal	89%	0%	8%	0%
Adjacent	Internal	100%	0%	0%	0%
Malignant	Internal	98%	17%	2%	16%

Overtyping occured mainly at higher peptide numbers (Figure 5D) and thus led mainly to incorrect annotations (Figure 5B). These false annotations involved mainly additional allotypes with similar motifs. In case of TIL1 the incorrect HLA-C*12 and C*16 were suggested, whose motifs have similarities to HLA-A*02 and are therefore likely to present Top20_CP-A*02 peptides in small quantities. This effect is expected to occur during in-depth measurements with large sample quantities and sensitive mass spectrometers resulting in a high peptide yield. The correct typing of allotypes per sample seemed to be less influenced by peptide numbers, as well as the proportion of missing allotypes (Figure 5A and C). Among the missing HLA, C*07 was most frequently missing. HLA-C was probably most challenging to type, due to the significantly lower expression and associated reduced number of HLA-C peptides [154]. HLA allotypes not considered during the training of the RF algorithm were not typed at all. In this case, the deciphering of the peptide motifs in the sample as shown in [65,192] can provide further indications whether an allele is missing or the sample is homozygous. There were no typing difficulties with the homozygous samples. If overtyping should occur, the excess HLA allotype most likely has significantly less Top20_CP than the correct HLA.

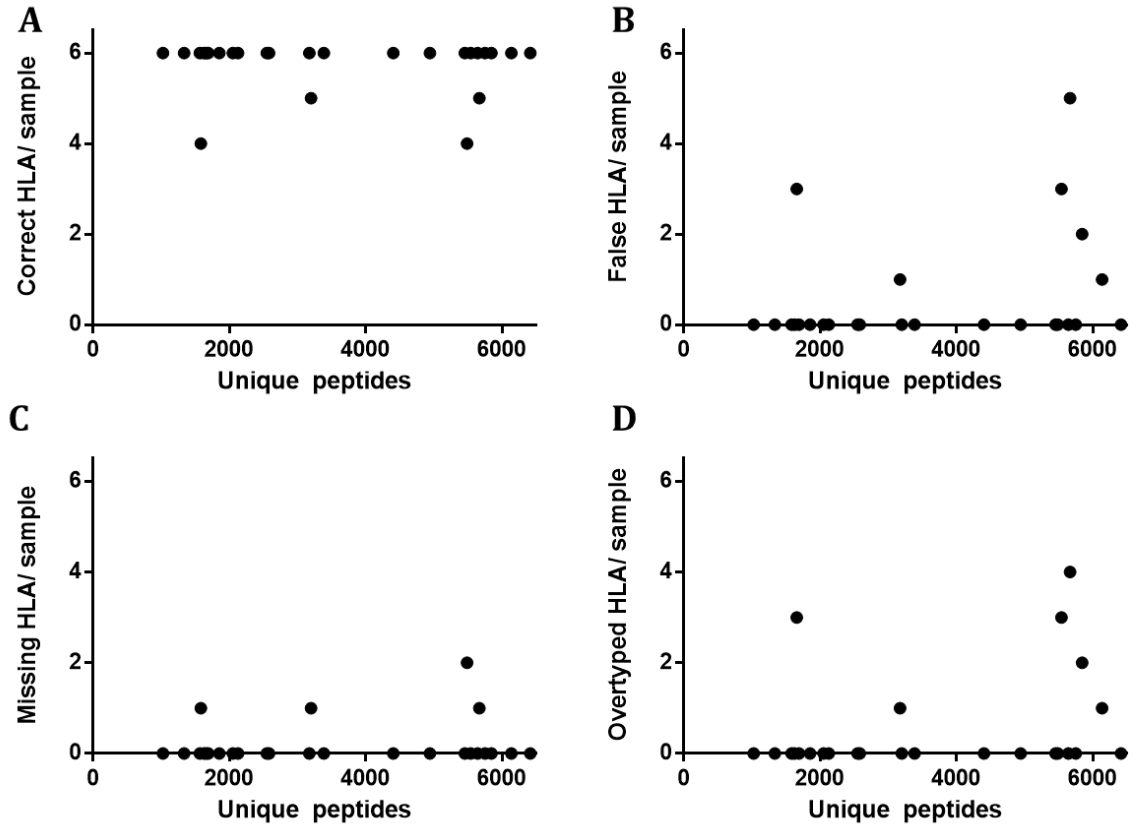


Figure 5: Visualized HLA allotyping performance. *HLA allotyping at the peptide level using the Top20_CP. A visualization of all correct (A), false (B), missing (C) and overtyped (D) allotypes of each sample of the 28 novel internal samples is given. The two samples with HLA not considered in the RF training are excluded.*

In case of the infrequent HLA allotypes that were not included in RF training, top lists were created from in-house data. For these allotypes at least a prioritization based on the peptide frequency on HLA positive samples and the motif was performed (Supplemental Figure S19; TopX lists in Supplemental Table S6). If an unknown allotype is suspected, a comparison with the TopX peptides of the infrequent HLA can be performed after typing using Top20_CP.

5.5.4 Allotypic peptides for application as internal standard

Due to the high immunopeptidome diversity and low abundance of HLA ligands, it is difficult to reliably identify peptides that can be used as internal standards to compare different samples. Here we evaluated the Top20_CP and AP as suitable unlabeled internal standards. Based on peptides eluted from a JY cell batch analyzed in 18 replicates over two weeks (as described in [116]), the technical influence on the Top20 peptide identification and unlabeled semi-quantification was investigated (Supplemental Figure S20A and B). In a comparison of two replicates each, there was a >90% overlap for the Top20_CP and AP (Supplemental Figure S20C). The area of the peptides overlapping in all 18 replicates, as compared to two replicates, had a low relative standard

deviation (%RSD) of 8% in the Top20_CP (n = 33) and 10% in the Top20_AP (n = 41) and across all 18 replicates the %RSD was still <20% (Supplemental Figure S20D).

Measuring a replicate of peptides eluted from a JY cell batch (batch production described in [116]) every two months, we were also able to determine a reliable identification of peptides from the top 20 pools over a period of one year (the JY cell batches were analyzed from Annika Nelde, Ana Marcu and Jens Bauer). The peptide abundance of the pool peptides fluctuated similarly to the total peptides (Supplemental Figure S21A and B). Many of the peptides from the Top20_CP and AP pool were highly abundant in the immunopeptidome (Supplemental Figure S21C and D). In comparison to the sum of peptide areas per replicate, which had a Pearson correlation of 100% between the total peptides and the Top20_CP and AP, the peptide overlap of the Top20_CP and AP between the replicates was significantly higher compared to the total peptides (Supplemental Figure S22A). Almost one third of the total peptides (28%) was only found on one replicate (Supplemental Figure S22C). For the peptides from the Top20 pools, however, this was only 7%. The majority of the pool peptides were found in all seven replicates (Top20_CP = 71%, AP = 67%) while only 22% of the total peptides were found in all replicates. In terms of retention time, the Top20 peptides showed less variation in RT compared to the total peptides (Supplemental Figure S22B). In contrast to the total peptides, the Top20 pool peptides seemed to be reliable candidates in the immunopeptidome.

5.5.4.1 Investigation of differential antigen processing and tumor exclusive peptides using allotypic peptides

Numerous factors influence protein metabolism from synthesis through antigen processing to antigen presentation, which modulate the pool of presented peptides. For these reasons it is difficult to determine whether peptides are actually not presented at all or only in low numbers. In order to investigate whether peptides are presented exclusively or modulated in tumors, malignant and adjacent benign tissues were examined (the peptide data of all malignant and adjacent benign tissues described in chapter 5.5.4.1 and the mentioned supplemental figures were derived from the internal database described in chapter 5.4). To identify peptides modulated in tumors, resulting from differential antigen processing (DAP), the top five single transcript proteins (STP) were selected, which were most frequently identified via a tumor-specific peptide and an additional peptide identified on healthy samples (Supplemental Figure S23). In addition, we selected the top STP with the reverse case, which was most frequently identified via a benign-specific peptide and additional peptides found on malignant samples. By using STPs (downloaded from ensembl.org, November, 2016), influences prior to antigen processing should be almost precluded with few remaining influences such as proteasomal splicing [197]. The allotypic peptides will be used to determine the respective tumor and adjacent benign sample size so that the differentially processed peptides in both samples could be compared. In addition, we searched for

established tumor-associated peptides (TUMAPs) in our samples. The allotypic peptides should be analyzed to ensure the TUMAPs would have been identified in adjacent benign tissue if the peptides had been presented in the same proportions.

We investigated exemplarily the samples of five donors and examined the DAP in four samples (colorectal cancer, non-small-cell lung carcinoma and renal cell carcinoma I and II; Figure 6; Supplemental Figure S24 and S26, 27) and the TUMAP presentation in three samples (colorectal cancer, non-small-cell lung carcinoma, and gastric cancer; Figure 6, Supplemental Figure S24 and S25).

In the colorectal cancer sample and the respective adjacent benign sample, the TUMAP VWSDVTPLTF and the peptide YGGLIFNSY, probably generated by DAP, were presented tumor-exclusively (Figure 6A). In comparison to the areas of the Top20_CP, the two peptides were presented poorly (Figure 6B). Nevertheless, there were peptides of the Top20_CP in the tumor sample which were presented in lower quantities and also found in the adjacent benign sample. Therefore, it could be assumed that the peptides would have been detected if they were presented to the same extent or slightly less in the adjacent benign sample. In case of the STP osteoclast-stimulating factor 1 (OSTF1) the peptides seemed to have a similar ratio of the mean area in the adjacent benign and tumor tissue which might indicate a similar processing (Figure 6C). In contrast, the STP cytochrome c oxidase subunit 2 (MT-CO2) seems to be processed differentially in the colorectal cancer sample. Despite the larger tumor sample according to Top20_CP, the peptide LPIEAPIRM was found in smaller amounts in the tumor and the peptide YGGLIFNSY was exclusively identified in the tumor. There seemed to be no downregulation of HLA class I molecules on the tumor (Figure 6D).

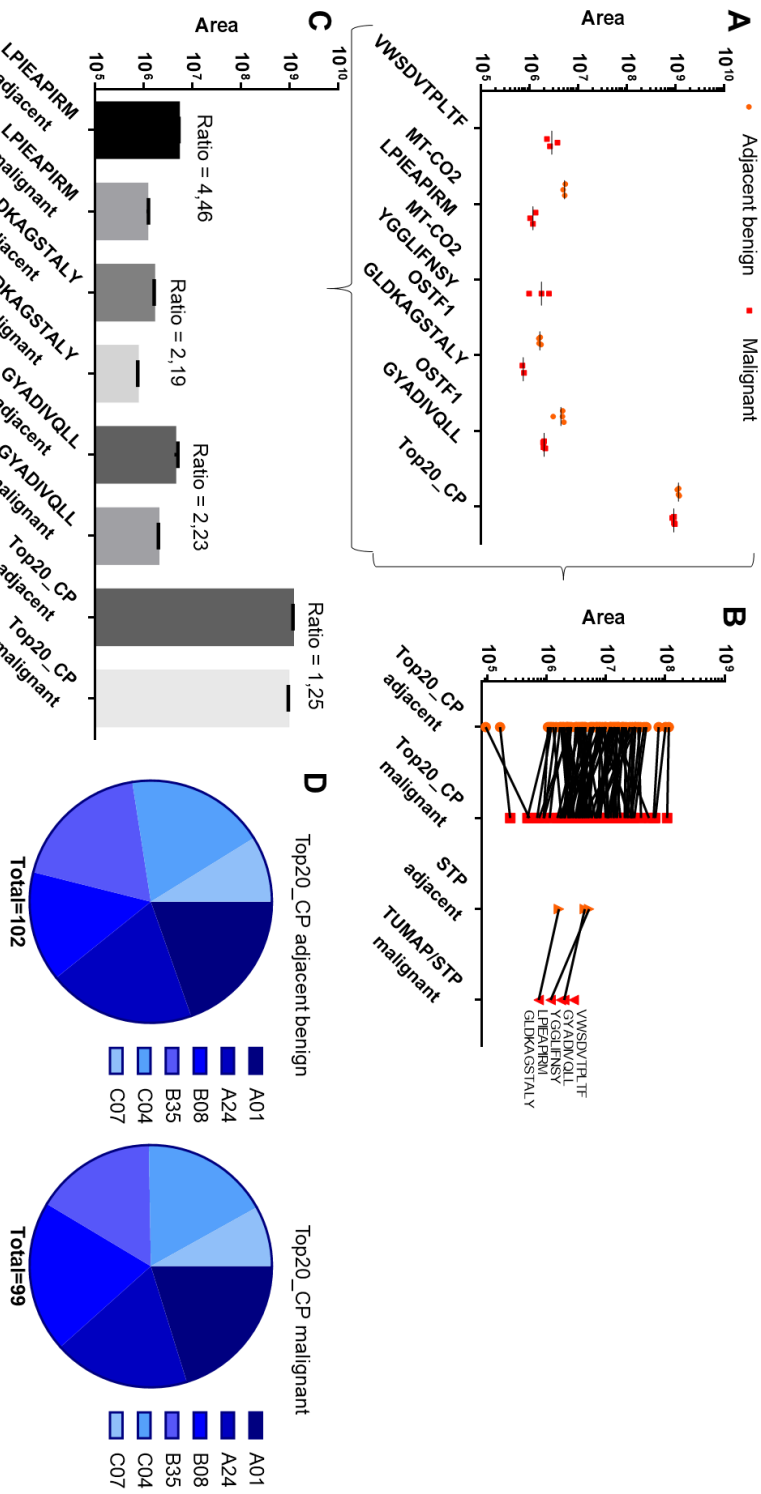


Figure 6: TUMAPs and DAP in colorectal cancer and adjacent benign tissue. *Area of the identified TUMAP (VWSDVTPLTF), STP (MT-CO2 and OSTF1) peptides and the Top20_CP of each replicate (total replicates: n = 4) in the adjacent benign and malignant sample (A). Individual peptide area of the TUMAP, STP peptides and Top20_CP (B). Ratio of the mean area of the peptides in the adjacent benign and malignant sample (C). The proportion of the allotype-specific Top20_CP in the malignant and adjacent benign sample (D).*

Similar to the case of colorectal cancer, also for the non-small-cell lung carcinoma and gastric cancer sample, it could be shown that the TUMAPs are most likely presented tumor-exclusively. Compared to allotypic peptides, TUMAPs are presented less abundantly (Supplemental Figure S24 and S25).

With regard to differential antigen processing, a similar processing of the STP of ferritin light chain (FTL) in both dignities seems to take place in the case of the non-small-cell lung carcinoma sample (Supplemental Figure S24). In the renal cell carcinoma I and II samples, however, a differential processing of the proteins STPs MT-CO2 and armadillo repeat-containing X-linked protein 1 (ARMCX1) seems to take place (Supplemental Figure S26 and S27).

In cases where a peptide is only found in a few replicates such as KVLEYVIKV (Supplementary Figure S25) or only a few replicates have been measured, no reliable assumption is possible. For example, in the case of the non-small-cell lung carcinoma (Supplemental Figure S24) the sensitivity of the mass spectrometric DDA method does not allow a statement whether the peptide is presented at all on adjacent benign tissue. Since only two replicates of this sample were analyzed and our LTQ Orbitrap XL has a recovery rate of about 60% with a 5% FDR [198], there still remains a small probability of about 20% that the TUMAP was existing in the adjacent benign sample but not detected.

The inhomogeneous Top20_CP ratios between the adjacent benign and tumor tissues in Figure 6B and Supplemental Figure S24-27B display the multiple factors impacting the immunopeptidome and influencing the peptide presentation of each peptide differently. Hence, the Top20_CP should be used summarized (as in Figure 6C and Supplemental Figure S24, 26 and 27C) to assess the size of the immunopeptidome of each tissue. For the STP derived peptides, however, there should be less difference in presentation compared to the Top20_CP as they derive from the same protein, which originates from one single transcript. There seemed to be no downregulation of HLA class I on the tumor tissues (Figure 6C, Supplemental Figure S24, 26, 27D and 25C). Besides the biological influences, additional mass spectrometer and data processing derived effects were possible as assessed in Supplemental Figure S20A. The Top20_CP serve as internal standard to estimate the size of the immunopeptidome and cover the detectable area range. Nevertheless, a spike-in of synthetically labelled peptides and targeted methods are the gold standard to investigate peptide quantities via LC-MS/MS.

5.5.5 Identification of dignity and tissue classification antigens

In addition to HLA classification based on peptides, RF has also been successfully used to classify the dignity (benign or malignant) of the samples in the database and additionally identify tissue-specific proteins presented in the immunopeptidome. Since tumor transformation is a multifaceted process that proceeds differently in each tumorigenesis, benign rather than

malignant was used for classification and dignity distinction. To identify tissue-specific proteins, only benign samples from tumor-free donors were used.

The dignity classification was carried out at both peptide and source protein level and the AUCs of the classification of the 2-d-DB were slightly better for the latter (Supplemental Figure S28A). The AUC was probably higher because the peptide restriction had less influence on proteins than on the presented peptides, and thus there was an improved classification of samples with different HLA-presented peptides but shared source proteins. For a closer look into the individual source proteins of the Top_20 benign classification peptides (Top20_BCPep), as well as Top20 benign classification proteins (Top20_BCPro), the number of benign samples positive for the peptides/proteins and the percentage of benign samples among the peptide/protein positive samples were analyzed (Supplemental Figure S28B and C). The Top20 indicated high sample coverage and the peptide and protein positive samples were on average 69% and 64% benign samples.

To provide simple Top20 lists which can be used for simple dignity-determination and to partially bypass the HLA allotype restriction, the criteria of the Top20_BCPro protein lists for benignity were used to create Top20 lists that were at least 65% positive for benign (Top20_BPro) and malignant (Top20_MPro) samples and have a high sample frequency (Supplemental Figure S28B and C). The Top20_Lists are provided in Supplemental Table S3. An analysis of the Top20_BCPep source proteins, Top20_BCPro, Top20_BPro and Top20_MPro annotated UniProt keywords and disease annotations from the GAD dataset [199] did not present conspicuous tumor annotations for the benign Top20, but the Top20_MPro did have tumor-associated annotations for both categories (Supplemental Figure S29).

To identify tissue specific proteins, only benign samples from Marcu *et al.* [183] were used. The HLA restriction was partially circumvented by classification using the peptide source proteins instead of the peptides. High AUCs could be achieved using RF that were comparable to the previously obtained AUCs for HLA and dignity classification (mean AUC: tissues = 92% ± 11%; maximal AUC: adrenal gland = 94% ± 9%; minimal AUC: Mamma = 76% ± 23%). A closer look at the UniProt keyword and disease annotations of the Top20 source proteins demonstrated that many of the annotations matched the respective tissue functions as depictured for the most annotated tissues liver, muscle, adrenal gland and kidney (Supplemental Figure S30, disease and keyword annotations in Supplemental Table S4).

5.5.5.1 Molecular tumor testing based on dignity classification antigens

After the promising HLA-classifications, we also aimed to investigate a dignity classification based on the immunopeptidome. Due to the idea of typing the "benignity" of the tissue, the typing should be applicable to all tumor types. Using the RF algorithm based on the peptides of our database it

was possible to perform a more than 83% correct dignity testing on the 28 internal test samples, covering more than ten tumor types (Supplemental Table S5). 94% of the tumor samples were correctly classified and one tumor sample was falsely assigned benign. In contrast, only 50% of the six benign samples were correctly typed as benign. A larger database, with more different tumor types and benign tissues, should significantly improve testing.

A tumor testing using Top20_BCPro should theoretically allow a good testing with more than three identified proteins, as these were found on average in 65% of the healthy samples. In practice, however, up to seven of the proteins were found on the tumor samples and a cut-off of 7-10 proteins led to the best result of 87% correctly typed samples (Supplemental Table S7; Top20_BCPro cut-off = 10 proteins). The use of Top20_BPro and Top20_MPro provides an additional parameter for tumor testing. In most tumor samples almost all Top20_MPro were located and the lowest were 13 proteins. However, many of the Top20_MPro could also be found on the benign samples, which may be attributable to the disproportionately high number of blood and bone marrow tumors in the database and much less benign blood and bone marrow samples. A cut-off with a ratio Top20_BPro/Top20_MPro of 0.4-0.8 led to the best classification with 87% correct samples. When using the Top20_MPro, care should be taken, especially with haematological samples, as benign samples might be often classified as tumor samples (Supplemental Table S7; Top20_BPro/Top20_MPro cut-off = 0.8). Interestingly, the mean value of the ratios from Top20_BPro/Top20_MPro for the adjacent benign samples with 0.25 ± 0.18 was between the tumor samples with 0.11 ± 0.12 and the benign samples 0.56 ± 0.40 and with the cut-off 0.8 all would have been classified as malignant.

5.6 Discussion

The immunopeptidome has been studied for almost three decades. It harbors ample information and will enable versatile future applications [200]. Nevertheless, we have only understood a fraction of the information potentially contained in the immunopeptidome. Machine learning enables deciphering the great diversity of the immunopeptidome to identify individual antigens that enable assessments about the HLA allotypes, tissue dignity or tissue origin. The primary goal of this present study was the identification of allotypic peptides that allow HLA-classification. The approach also turned out to be suitable to assign tissue dignity.

First, we confirmed that the approach to determine peptides with the highest possible sample frequency leads to reliable identifications. Assessment of true ligands of monoallelic cell lines revealed that most of these ligands are not specific for a distinct HLA allotype but are present on several allotypes. This supports the concept of grouping HLA allotypes according to the anchor specificities of the peptide motifs in supertypes [154,201,202]. However, there are highly specific peptides for each allotype. In the considered monoallelic cell lines the HLA-A presented peptides

were the most allotype-restricted peptides. The HLA-C presented peptides, on the other hand, seemed to be poorly restricted to each HLA-C molecule. HLA-A*01, A*02, B*15 and B*51 mapped peptides were the most allotypically restricted, which might be explained by their relatively unique motifs and the high molecule count on the surface compared to HLA-C.

To reliably determine which peptides are highly specific and identified on most respective samples, a large database was generated covering many HLA allotypes across different tissues. The RF algorithm enabled us to identify those HLA-presented peptides with highest specificity and frequency for HLA allotype assignment. The top 20 peptides of each allotype, found on many samples, are very specific and restricted, except for HLA-C*17. The RF algorithm considered the linkage disequilibrium and thus achieved optimal HLA-classification at the peptide level. Assessing the top 20 proposed peptides of the algorithm which were unaffected by linkage disequilibrium and fit the motif of the respective HLA, we discovered that these peptides are derived from proteins distributed over the entire range of protein quantities and protein degradation rates in the cytoplasm. These proteins represent the most common protein functions in the cytosol, and some are also involved in host-virus interactions.

The HLA allotyping using individual peptides enables a good HLA-classification based on the RF algorithm, trained with the entire database. In addition, using merely the top 20 proposed peptides per allotype (Top20_CP) lead to successful classifications. A manual allotyping based on the Top20_CP enables an additional subjective assessment of the proposed allotypes and can even improve the typing. An interesting observation was a frequent overtyping of tumor samples with additional allotypes. One possible explanation could be the increased peptide presentation in the tumor samples. So far, there was no simple tool available to reliably identify the allotype after immunopeptidomic analysis, except of PCR typing of additional donor tissue or extensive dataset comparisons (e.g. MixMHCpred [192]), which can now easily be done using the Top20_CP.

In addition to allotyping, the Top20_CP were also suitable as an internal standard and outperformed the total peptides in semi-quantitative estimation of the "size" of the immunopeptidome. This allows a comparison of two separate samples with different presented peptide numbers. In this way, it is possible to determine whether a certain peptide is indeed tumor exclusive and to ensure that the peptide did not remain unidentified due to a smaller sample of surrounding benign tissue. In addition, it can be investigated whether the source protein was differentially processed in the tumor compared to the surrounding benign tissue. To date, semi-quantitative analyses have usually considered the total amount of peptides as a comparative value [53,186]. However, the allotypic peptides might be a better reference as demonstrated in this study. The search for a reliable internal standard is an important prerequisite to ensure the comparability and robustness of immunopeptidomic analysis and is especially relevant for

promising applications such as quantitative immunopeptidomics [24,203]. So far, protocols exist for the validation of LC-MS/MS based immunopeptidomics pipelines according to pharmaceutical good manufacturing practices (GMP) for peptide identification [116]. Nevertheless, there is currently no method validation available for (semi-)quantitative immunopeptidomics, in which allotypic peptides might become an important tool as internal standard.

Allotypic peptides are a tool for a variety of applications, many of which are sure to be found. In this paper, presented applications are HLA allotyping or the use as internal standard for unlabeled semi-quantitative experiments. Further implementations include quality control in HLA-peptide monomer refolding or, as previously used, negative control in immunogenicity assays [204], or to generate monomers for UV-mediated ligand exchange [186].

Besides the determination of allotypic peptides, our RF pipeline also proved successful in the identification of tissue- and dignity-specific antigens. Using the RF algorithm, trained with the entire database, effectively the dignity (benign or malignant) of samples could be determined on protein level as well as by typing based on solely the top 20 benign and malignant associated antigens. The extensive peptide diversity in various tissues and the high sensitivity of mass spectrometry highlight the potential use of immunopeptidomic for antigen classification and biomarker diagnostics for cancer testing and further diseases [205,206].

In this study, we have demonstrated which conclusions are possible by means of the immunopeptidome, based on a large database combined with artificial intelligence. With even larger databases, more reliable and robust estimations will be possible. In future it might become feasible to classify individual tumor entities or different diseases. We have compiled top 20 lists that anyone can perform HLA-classification and molecular tumor testing of immunopeptidomic data without a database. It was also shown how the combination of subjective evaluation with the proposed results of machine learning could even improve testing. However, with a larger database, we assume subjective evaluation will be dispensable soon. In future, machine learning might revolutionize and improve clinical diagnostics, precision treatments and health monitoring [207]. The immunopeptidome, with its substantial diversity and detailed immunological information, which can be structured and made comprehensible with the help of artificial intelligence, has the potential to become an important part of it.

5.7 Acknowledgements

This work was supported by the German Cancer Consortium (DKTK) and the Natural and Medical Sciences Institute at the University of Tübingen NMI. The authors thank Claudia Falkenburger for the in-house antibody production.

5.8 Supplementary data

5.8.1 Supplementary Tables

Supplemental Table S1: HLA allele frequency. *Comparison of the HLA allele frequency of the four-digit typed samples (allotype ≥ 5%) with allelefrequencies.net (04.2020; German pop 8), the frequencies of HLA-A, B and C correlate 97% each (Pearson correlation).*

	Allele frequency	
HLA	**4-d-DB**	**Germany**
A*01:01	13%	15%
A*02:01	28%	27%
A*03:01	16%	15%
A*11:01	7%	6%
A*24:02	11%	10%
B*07:02	12%	12%
B*08:01	9%	10%
B*15:01	7%	6%
B*18:01	7%	5%
B*35:01	8%	6%
B*40:01	6%	5%
B*44:02	9%	7%
B*51:01	9%	6%
C*01:02	5%	4%
C*02:02	6%	5%
C*03:03	5%	5%
C*03:04	9%	7%
C*04:01	15%	13%
C*05:01	7%	6%
C*06:02	9%	10%
C*07:01	14%	15%
C*07:02	14%	13%
C*12:03	7%	6%
Pearson correlation		97%

Supplemental Table S2: Monoallelic cells. *Summary of the monoallelic samples, technical replicates, and MS RAW files.*

Cell line	Transfected HLA	Technical replicates	Raw file name
C1R	A*02:01	8	C1R-A02_msms1-8
C1R	A*24:02	3	C1R-A2402_msms1-3
C1R	A*32:01	5	C1R-A3201_msms1-5
C1R	B*15:01	3x10	C1R-B1501_msms1-3_1-10
C1R	B*52:01	5	C1R-A5201_msms1-5
C1R	B*58:01	5	C1R-A5801_msms1-5
C1R	C*01:02	5	…C1R-C0102…msms13-17
C1R	C*02:02	5	…C1R-C0202…msms7-11
C1R	C*03:03	5	…C1R-C0303…msms1-6
C1R	C*03:04	5	…C1R-C0304…msms31-35
C1R	C*04:01	5	…C1R-C0401…msms7-11
C1R	C*05:01	5	…C1R-C0501…msms13-17
C1R	C*06:02	5	…C1R-C0602…msms18-22
C1R	C*07:01	5	…C1R-C0701…msms41-45
C1R	C*07:02	5	…C1R-C0702…msms47-51
C1R	C*08:02	5	…C1R-C0802…msms13-17
C1R	C*12:03	5	…C1R-C1203…msms23-28
C1R	C*14:02	5	…C1R-C1402…msms5-12
C1R	C*15:02	5	…C1R-C1502…msms25-29
C1R	C*16:01	5	…C1R-C1601…msms19-23
C1R	C*17:01	5	…C1R-C1701…msms31-35
C1R	G*01:01	5	…C1R-G0101…msms7-15

Supplemental Table S3: Mapped ligands from monoallelic cell lines and Top20 peptide/protein lists. *List of peptide sequences, which were isolated, clustered, and mapped to the 4-d-DB from the in-house analyzed and unpublished monoallelic cell lines. Summary of generated HLA specific Top20_CP and AP, tissue specific Top20_TissuePro and dignity specific Top20_BCPro, Bpro and MPro lists.*

Supplemental Table S3 is attached externally.

Supplemental Table S4: Disease GAD and UniProt keyword annotations. *Summary of all Top20 (Top20_AP, BCPro, BPro, MPro, TissuePro) disease GAD and UniProt keyword annotations (Top20_AP: only source proteins of peptides with n ≤ 3 protein annotations were considered). Disease annotations and functional UniProt keyword annotations were obtained using DAVID david.ncifcrf.gov.*

Supplemental Table S4 is attached externally.

Supplemental Table S5: Novel samples for HLA allotyping verification. *Summary of the novel 28 samples analyzed in-house and 10 samples from the literature, not included in the database, which were used to test the HLA typing based on peptides.*

Supplemental Table S5 is attached externally.

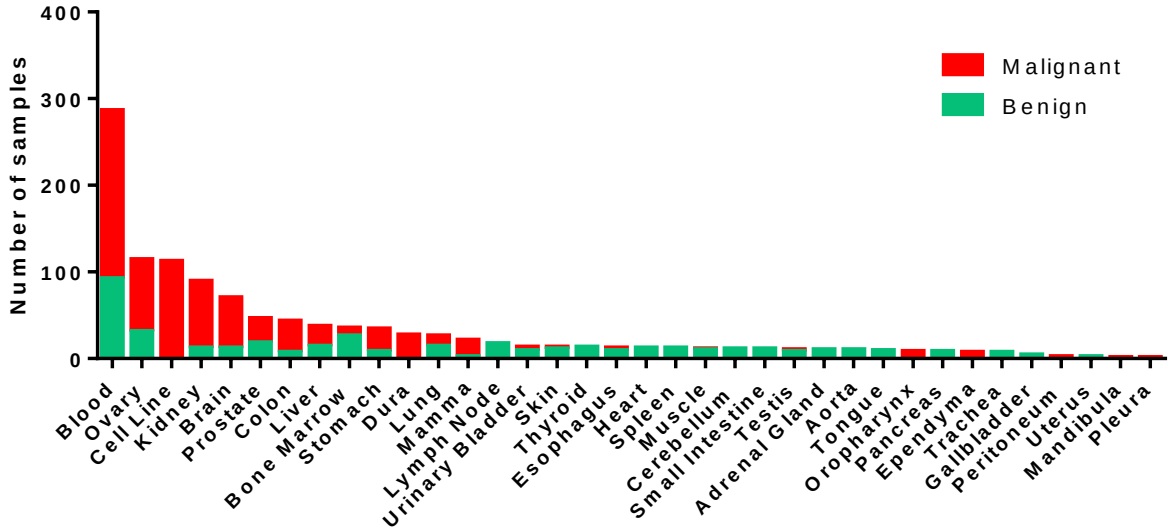
Supplemental Table S6: Infrequent allotypes. *Summary of generated HLA specific TopX peptide lists not included in RF. For infrequent HLA, a prioritization based on the frequency of the peptide on HLA positive samples and the peptide motif was performed. The number of overlaps was reduced until at least a Top20 list was reached (except for HLA-C*18).*

Supplemental Table S6 is attached externally.

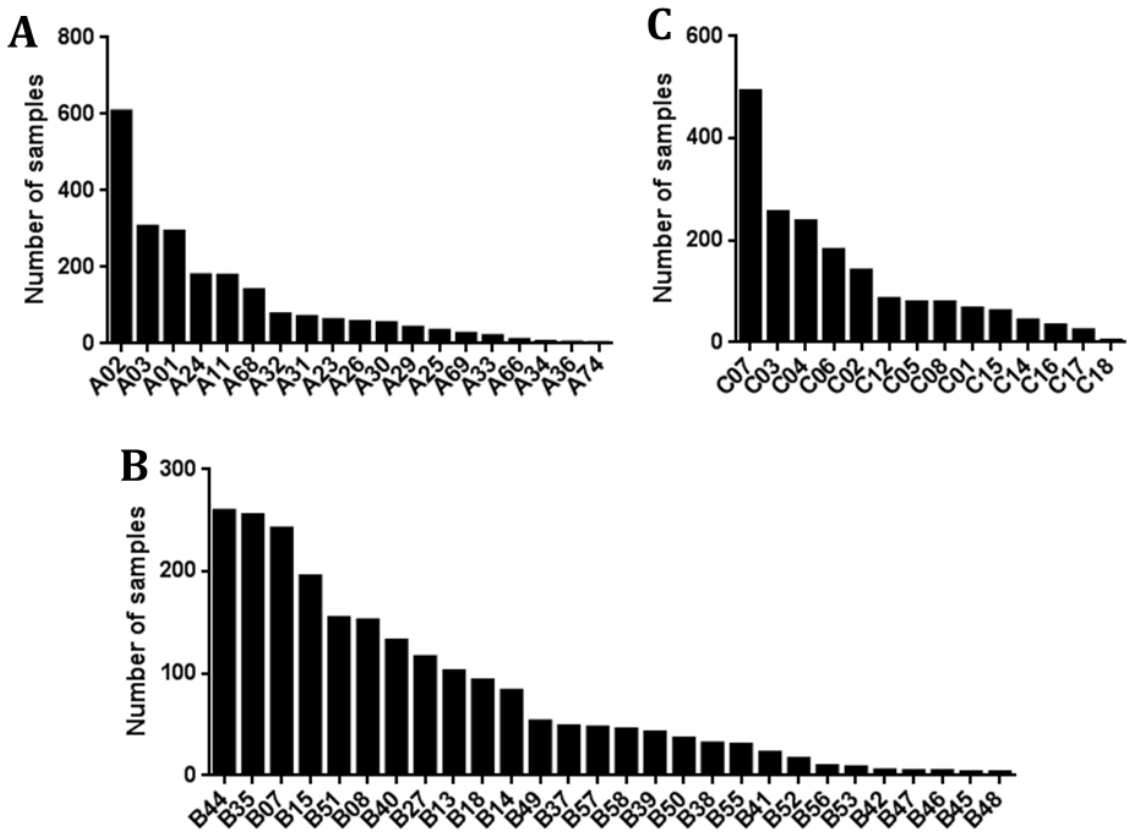
Supplemental Table S7: Novel samples for molecular tumor testing verification. *Summary of the novel 28 samples analyzed in-house, not included in the database, which were used to test the dignity-typing (malignant or benign) based on peptides.*

Supplemental Table S7 is attached externally.

5.8.2 Supplementary Figures



Supplemental Figure S1: Primary tissue and cell line database. *The immunopeptidome database contained benign and malignant primary samples from 36 different tissues from various donors and cell lines.*



Supplemental Figure S2: Number of HLA positive samples in the database. *The immunopeptidome database covered 19 different HLA-A, 28 HLA-B and 14 HLA-C allotypes on two-digit level.*

	A*01:01	A*02:01	A*03:01	…	Number of HLA-A typed peptide presenting samples
ITDSAGHILY	75%	26%	13%	…	121/498
RPSGPGPEL	15%	41%	30%	…	113/498
FAAGYNVKF	24%	64%	19%	…	42/498
…	…	…	…	…	…

	B*07:02	B*14:02	B*15:01	…	Number of HLA-B typed peptide presenting samples
ITDSAGHILY	18%	4%	5%	…	119/498
RPSGPGPEL	70%	4%	4%	…	113/498
FAAGYNVKF	9%	2%	4%	…	46/498
…	…	…	…	…	…

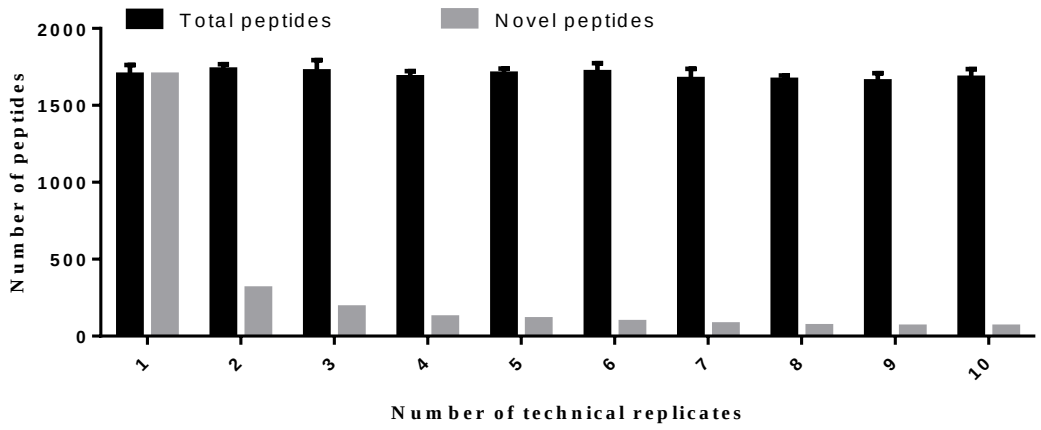
	C*01:02	C*02:02	C*03:03	…	Number of HLA-C typed peptide presenting samples
ITDSAGHILY	4%	13%	6%	…	112/498
RPSGPGPEL	5%	6%	2%	…	103/498
FAAGYNVKF	5%	90%	5%	…	40/498
…	…	…	…	…	…

Number of samples
- Number of peptide positive HLA-A, B or C typed samples
  - ITDSAGHILY: 121 (HLA-A)

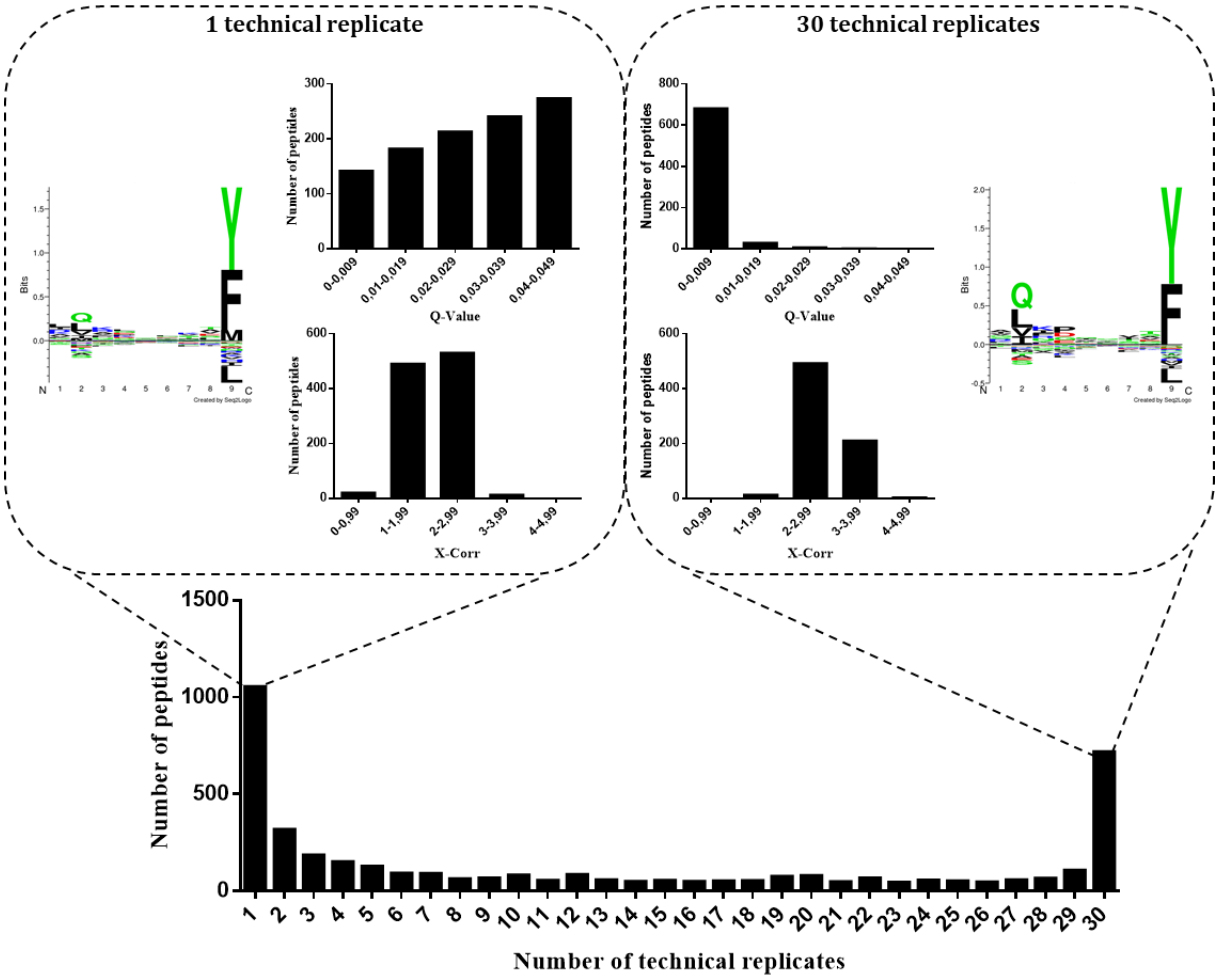
%allotype positive samples
- Percentage of peptide and allotype positive samples
  - ITDSAGHILY: 91 (HLA-A*01:01) ÷ 121 (HLA-A) = 75% (HLA-A*01:01)

%allotype restriction
- Difference between the %allotype positive samples of the first and second maximum allotypes (allotypes of similar HLA locus) for the peptide
  - ITDSAGHILY: 75% (HLA-A*01:01) – 26% (HLA-A*02:01) = 49%

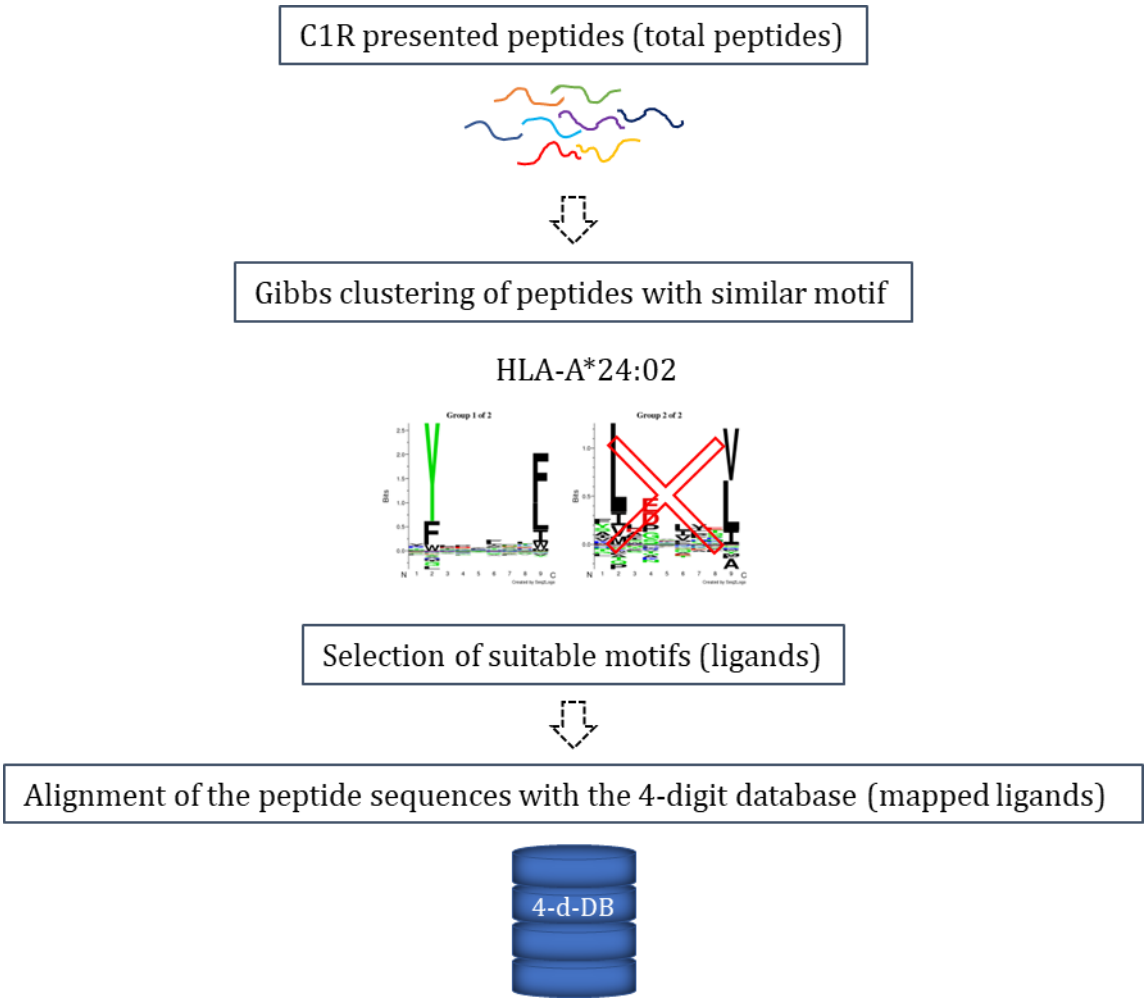
Supplemental Figure S3: 4-digit HLA-A, B and C tables. *The 4-digit typed samples were used to generate tables of the peptide frequencies for each HLA-A, B and C allotype per sample in the 4-d-DB and the total number of peptide positive samples in the DB (homozygous HLA allotypes were counted as one allotype per sample). The frequency tables were used to determine the parameters mean number of samples, mean %positive samples and mean %allotype restriction. Exemplarily, these three parameters were determined for the peptide ITDSAGHILY using the HLA-A table.*



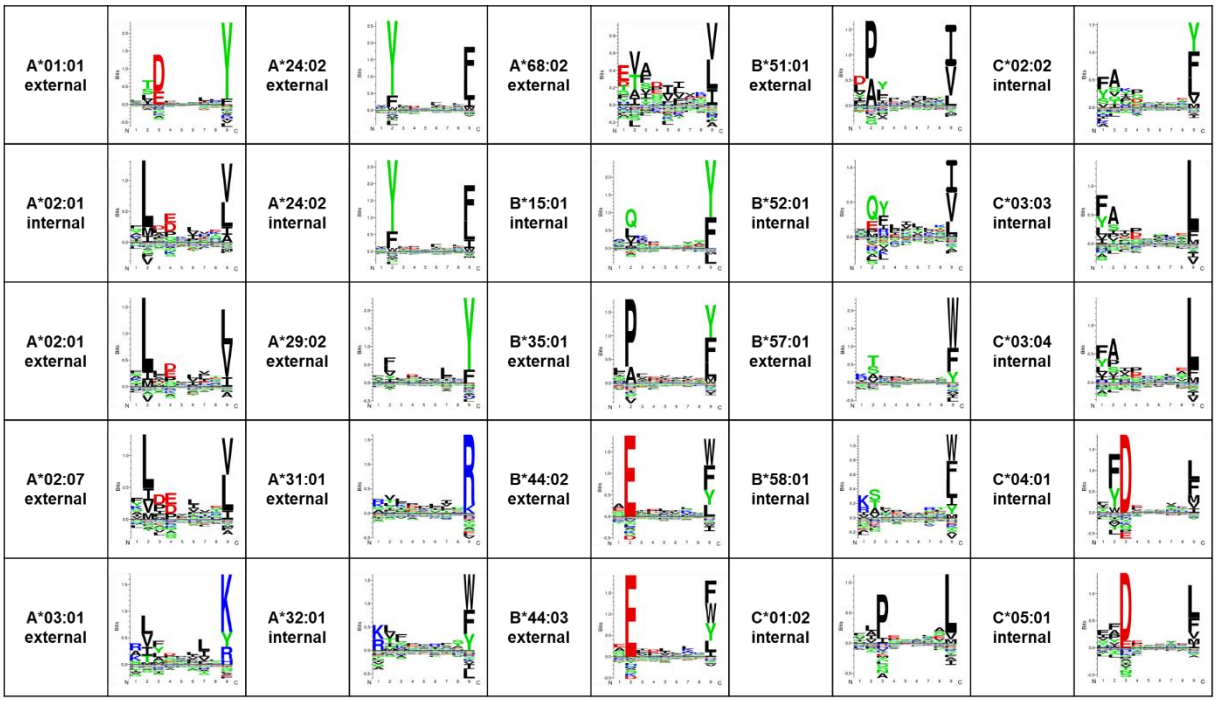
Supplemental Figure S4: Novel peptides per replicate. *Three HLA-B*15:01 transfected C1R samples were analyzed in ten replicates on the LTQ Orbitrap XL LC-MS/MS system. The total number of peptides and the novel peptides per replicate are indicated. All HLA-B*15:01 peptide extractions were carried out by Meret Beyer.*

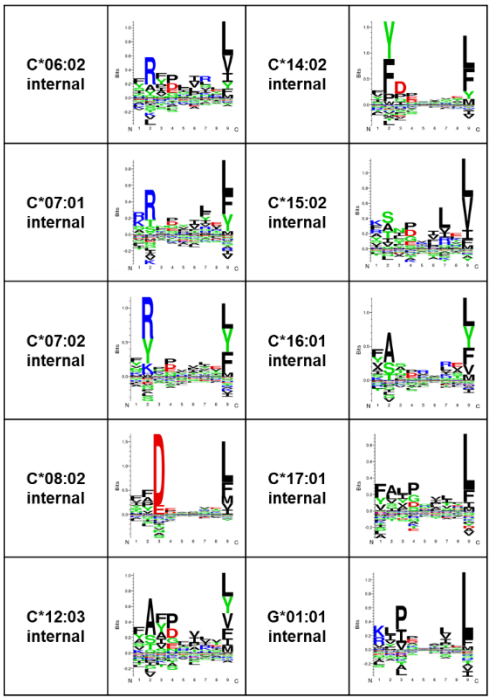


Supplemental Figure S5: Peptide overlap and quality per replicate. *Three HLA-B*15:01 transfected C1R samples were analyzed in ten replicates on the LTQ Orbitrap XL LC-MS/MS system. The peptide overlap after each replicate and the peptide motif, Q-Value and x-Corr of all peptides found in one or thirty replicates are indicated. All HLA-B*15:01 peptide extractions were carried out by Meret Beyer.*



Supplemental Figure S6: Isolation and mapping of HLA ligands. *Using Gibbs clustering, all peptides matching the motif of the transfected HLA were isolated and subsequently aligned to the peptides in the 4-d-DB.*

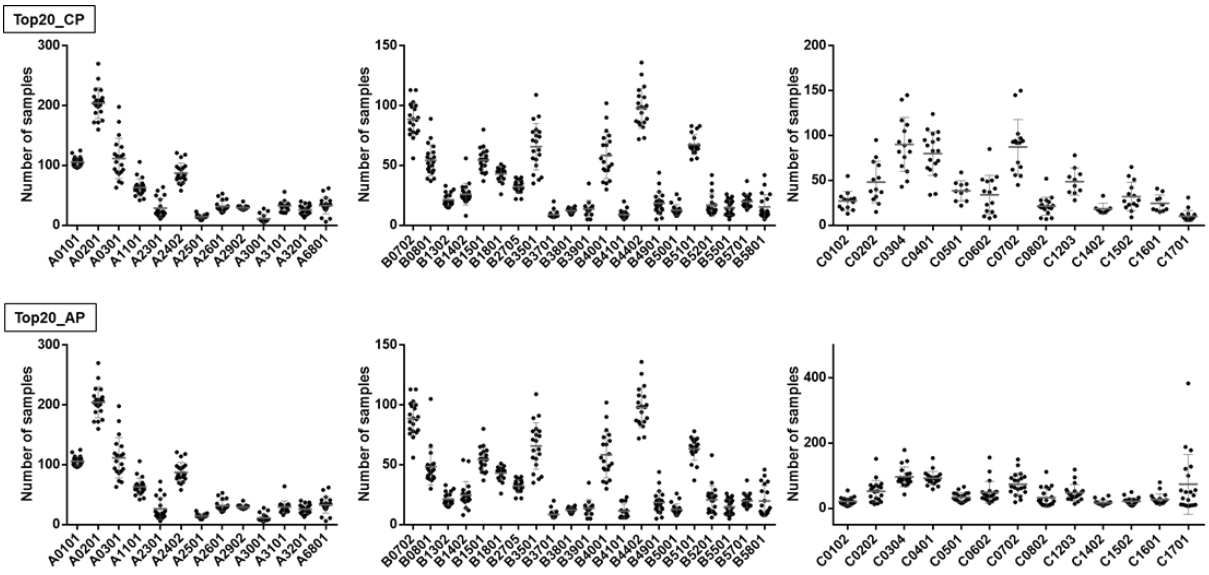




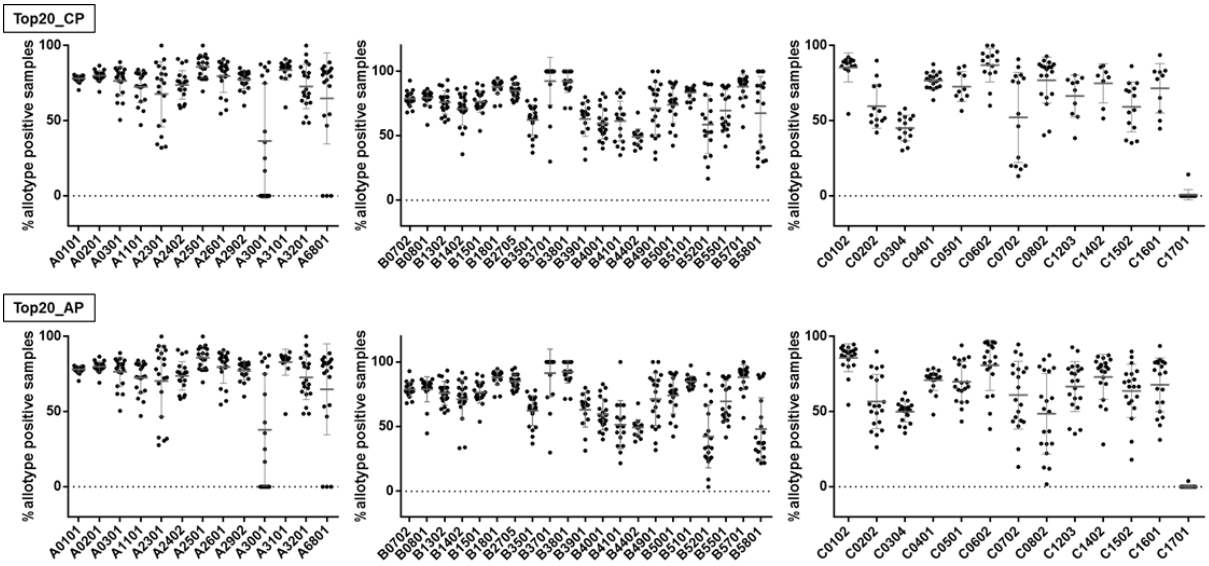
Supplemental Figure S7: Peptide motifs of mapped ligands from monoallelic cell lines. *Peptide motifs of peptides from the internal and external monoallelic cell lines, which were isolated, clustered, and mapped to the 4-d-DB.*

Allotype	Length	%	Allotype	Length	%	Allotype	Length	%	Allotype	Length	%
A0101 external	8	3	B4402 external	8	0	B5801 internal	8	6	C0702 internal	8	14
	9	54		9	52		9	73		9	68
	10	31		10	34		10	8		10	6
	11	9		11	11		11	7		11	5
	12	2		12	2		12	7		12	6
A0201 external	8	1	B4403 external	8	0	C0102 internal	8	8	C0802 internal	8	16
	9	75		9	53		9	72		9	73
	10	14		10	35		10	14		10	7
	11	10		11	9		11	4		11	3
	12	1		12	2		12	2		12	1
A0207 external	8	0	B5101 external	8	31	C0202 internal	8	5	C1203 internal	8	6
	9	89		9	62		9	77		9	83
	10	6		10	5		10	9		10	5
	11	5		11	2		11	5		11	4
	12	1		12	1		12	4		12	2
A0301 external	8	1	B5701 external	8	2	C0303 internal	8	7	C1402 internal	8	14
	9	55		9	57		9	81		9	68
	10	25		10	26		10	7		10	10
	11	17		11	13		11	3		11	5
	12	3		12	1		12	1		12	3
A2402 external	8	1	A0201 internal	8	2	C0304 internal	8	9	C1502 internal	8	7
	9	78		9	75		9	74		9	86
	10	15		10	15		10	9		10	2
	11	6		11	8		11	5		11	2
	12	1		12	1		12	3		12	2
A2902 external	8	0	A2402 internal	8	3	C0401 internal	8	20	C1601 internal	8	14
	9	82		9	80		9	71		9	72
	10	14		10	12		10	6		10	6
	11	3		11	4		11	2		11	5
	12	1		12	1		12	1		12	3
A3101 external	8	2	A3201 internal	8	0	C0501 internal	8	16	C1701 internal	8	14
	9	49		9	76		9	71		9	49
	10	19		10	11		10	8		10	14
	11	26		11	11		11	5		11	14
	12	3		12	2		12	1		12	10
A6802 external	8	0	B1501 internal	8	8	C0602 internal	8	7			
	9	75		9	72		9	87			
	10	16		10	15		10	2			
	11	6		11	4		11	1			
	12	2		12	1		12	4			
B3501 external	8	2	B5201 internal	8	54	C0701 internal	8	10			
	9	70		9	44		9	70			
	10	16		10	2		10	7			
	11	9		11	0		11	5			
	12	2		12	0		12	8			

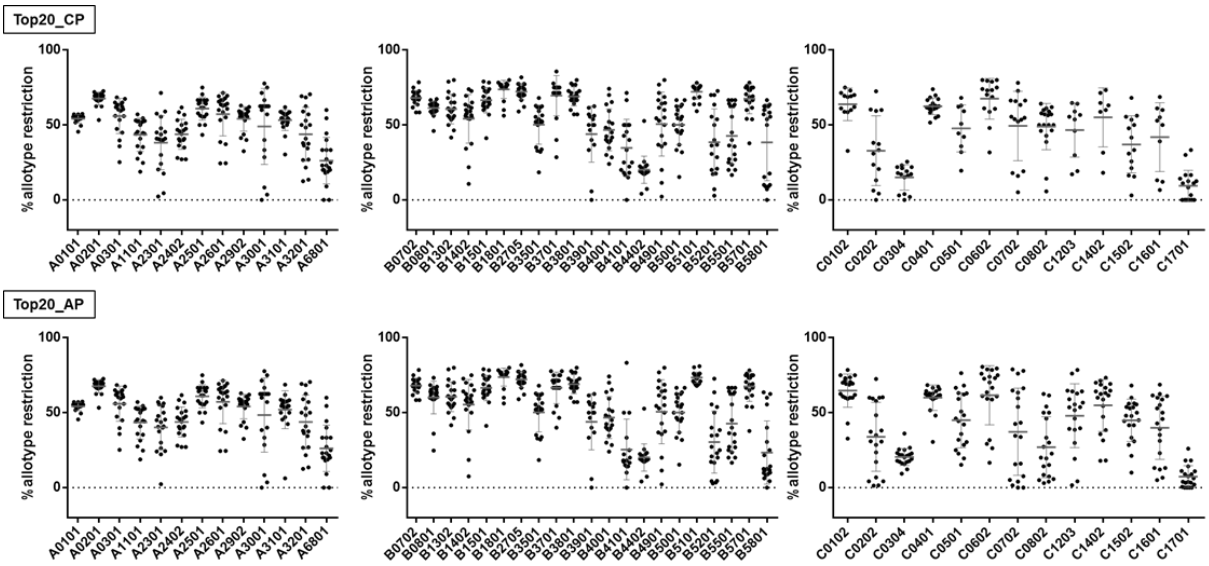
Supplemental Figure S8: Length distribution of mapped ligands from monoallelic cell lines. *Length distribution of peptides from the internal and external monoallelic cell lines, which were isolated, clustered, and mapped to the 4-d-DB.*



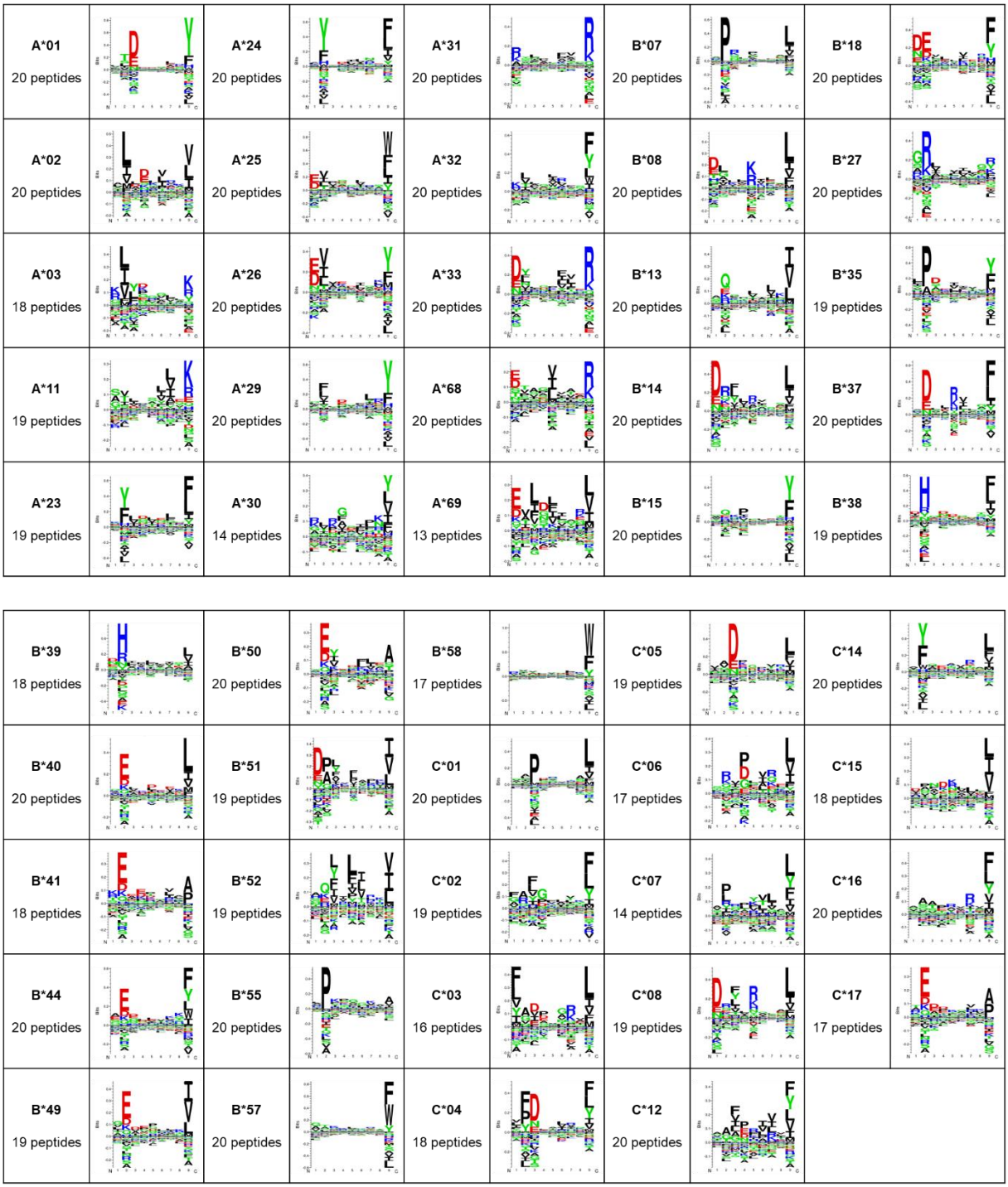
Supplemental Figure S9: Number of peptide positive samples of Top20_CP and AP. *Number of samples (in 4-d-DB) on which each individual peptide from the Top20_CP and AP of each allotype was identified for HLA-A, B and C allotypes.*



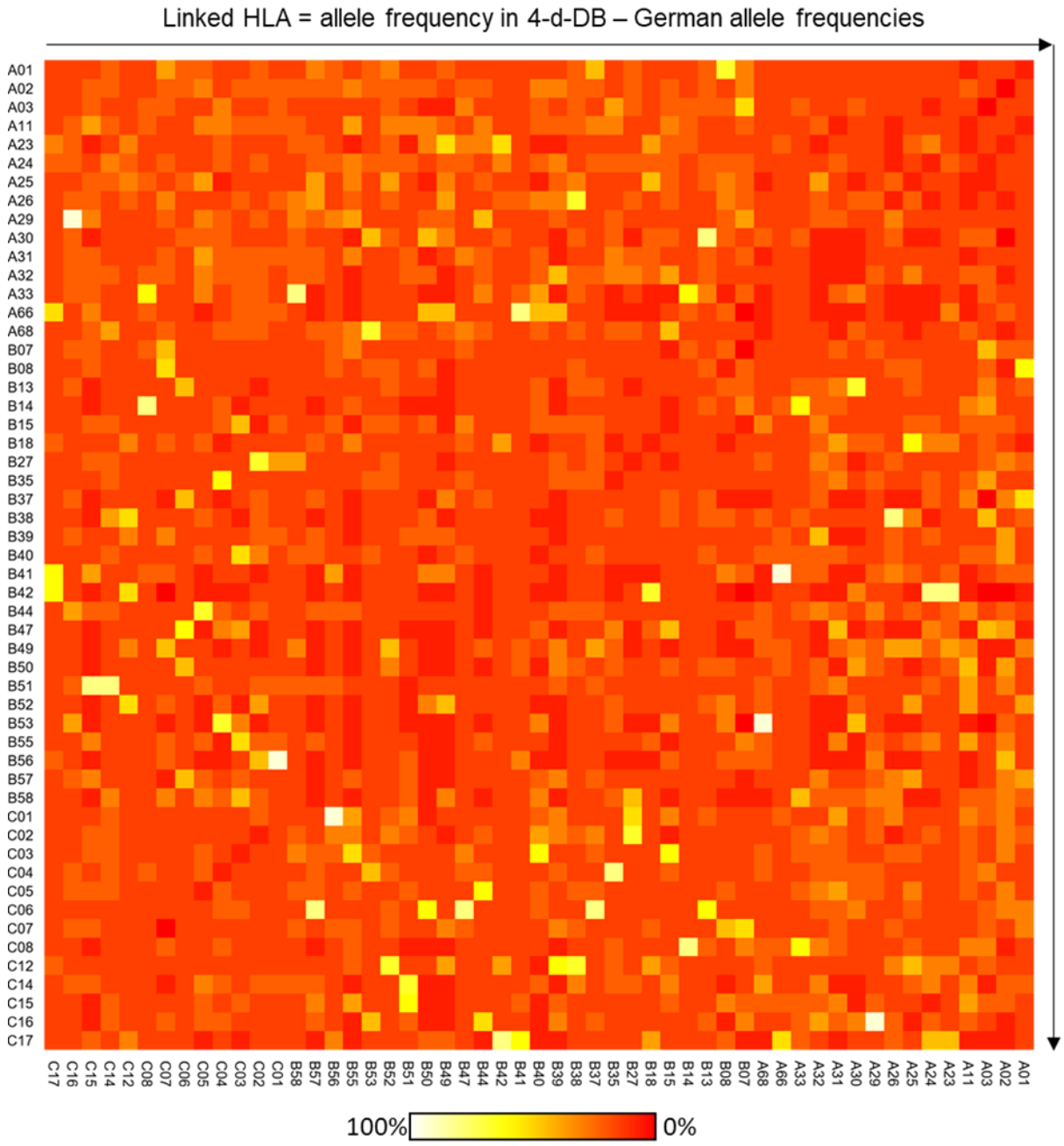
Supplemental Figure S10: Percentage of allotype positive samples of Top20_CP and AP. *Percentage of allotype positive samples (in 4-d-DB) from the samples on which each individual peptide from the Top20_CP and AP of each allotype was identified for HLA-A, B and C allotypes.*



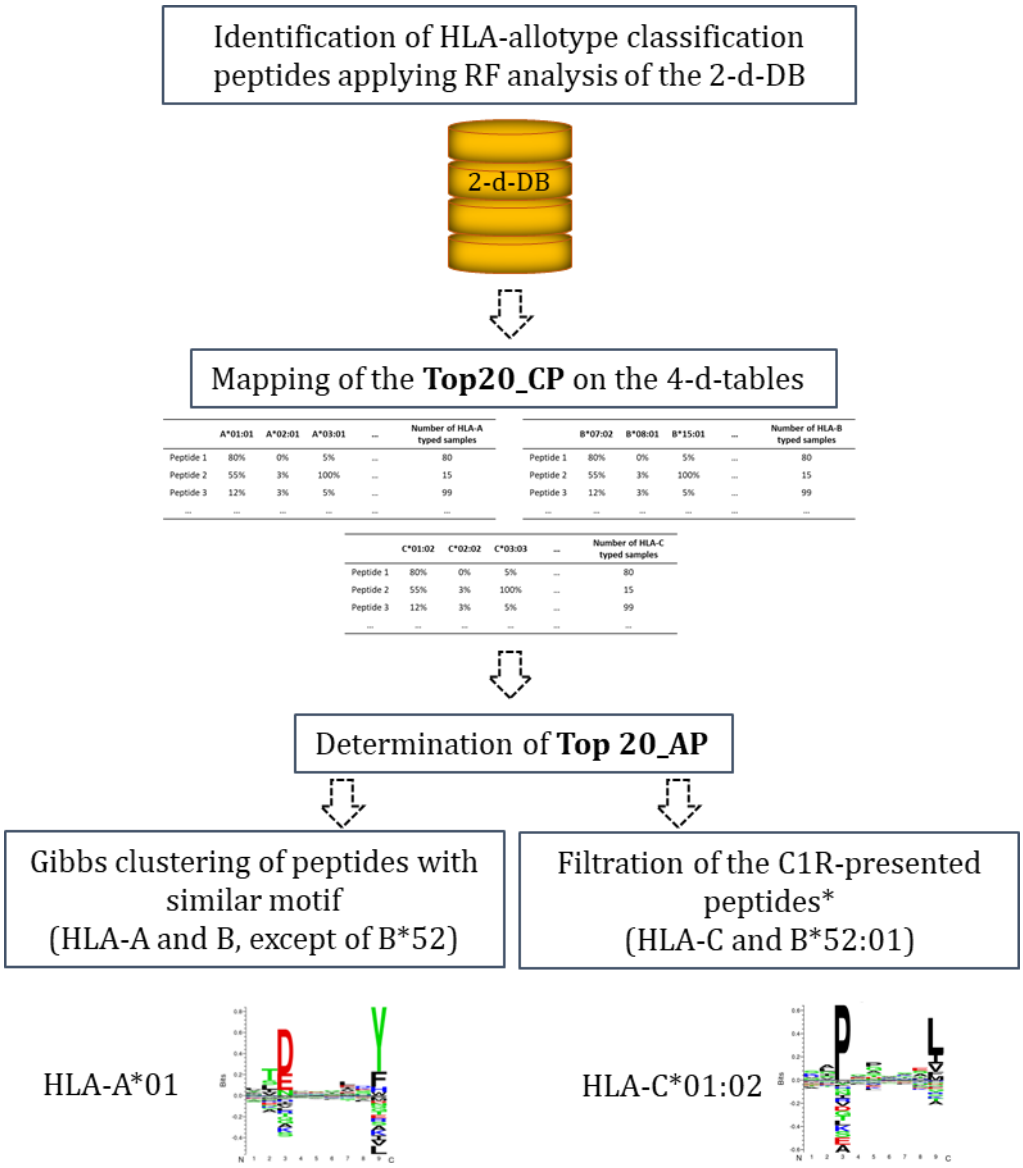
Supplemental Figure S11: Allotype restriction of Top20_CP and AP*. Allotype restriction is indicating the difference of the mean %allotype positive samples of the best fitting and the next highest allotype for each individual peptide of the Top20_CP and AP of each allotype for HLA-A, B and C.*



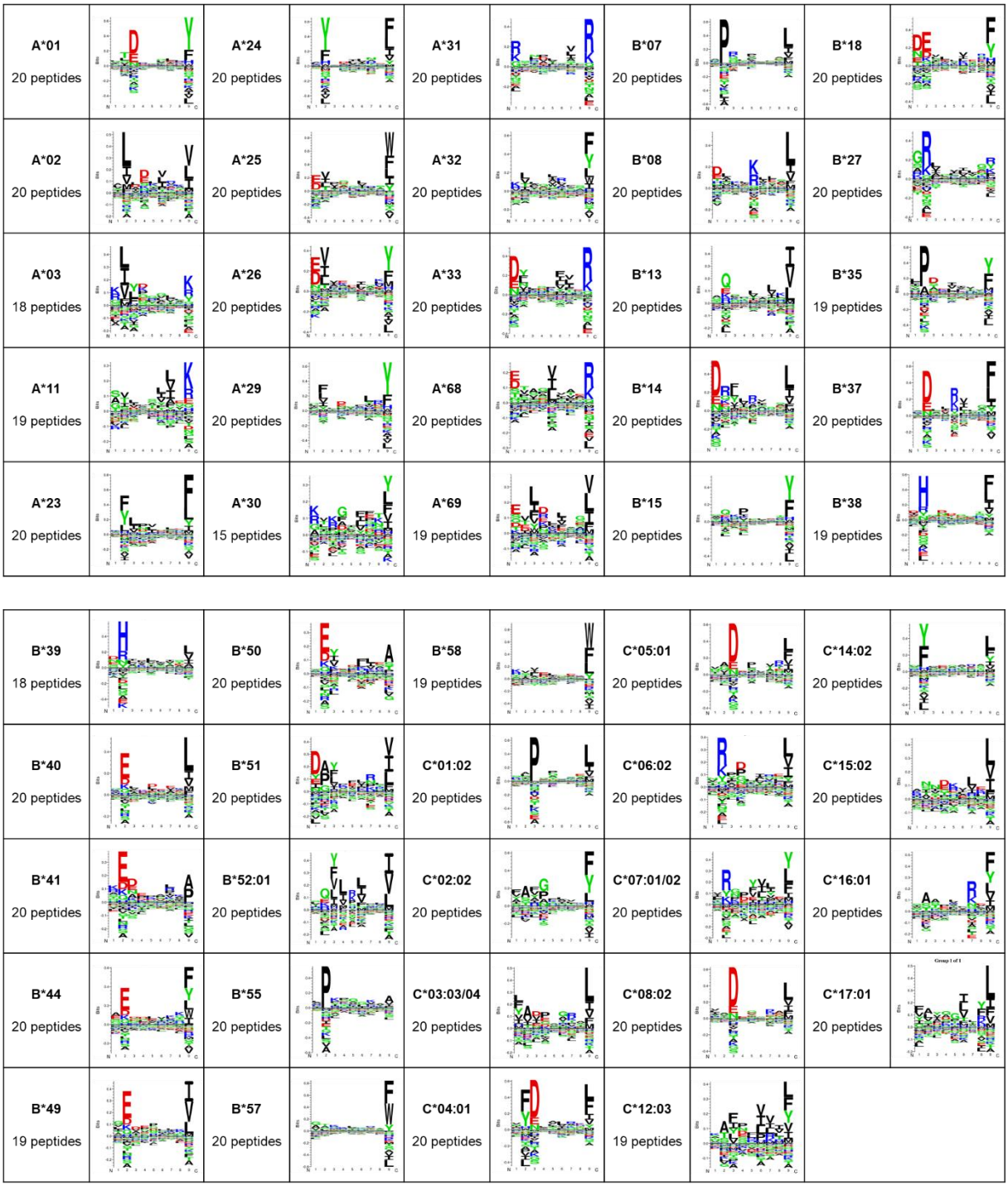
Supplemental Figure S12: Peptide motifs of the Top20_CP for each HLA. *Peptide motifs generated from the Top20_CP for each allotype. Due to the linkage disequilibrium HLA-C peptide motifs contain contaminating peptides with motifs matching other HLA allotypes (except for C*01, *05, *14 and *16).*



Supplemental Figure S13: Linkage disequilibrium of the alleles in the 4-d-DB*. Heat map for HLA with more than five positive samples in the 4-d-DB at two-digit level, indicating how frequently the combination of two alleles occur in a sample of the 4-d DB. The allele frequencies of the German population (allelefrequencies.net: German pop 8, 04.2020) were subtracted from the calculated frequency, resulting in the linked alleles. In the samples of the 4-d-DB, especially HLA-C alleles are frequently present in combination with HLA-B alleles. The heatmap was created by Jonas Scheid.*



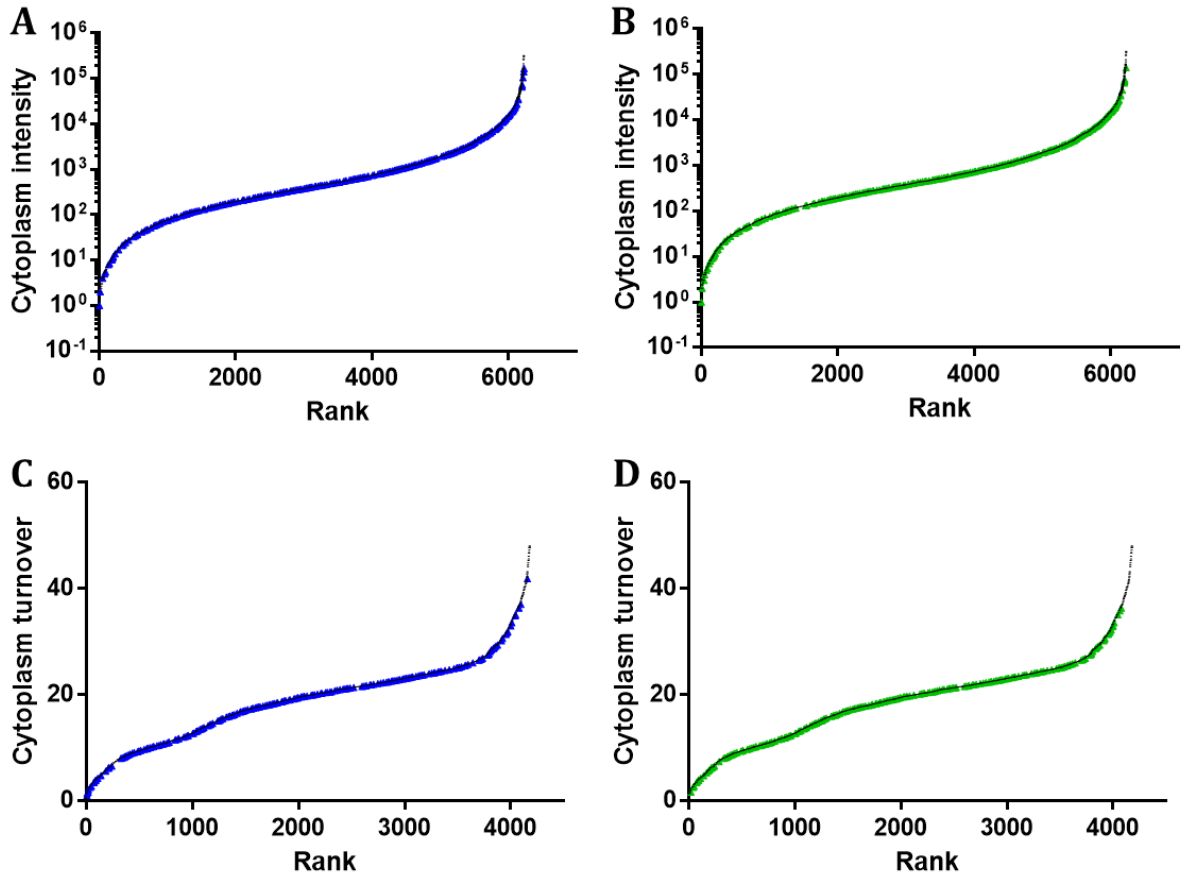
Supplemental Figure S14: Mapping and clustering of Top20_CP to Top20_AP. *The Top20_CP were mapped to the four-digit HLA-A, B and C tables generated from the 4-d-DB. Subsequently, the peptides with the fitting peptide motif were clustered with Gibbs clustering or, in the case of HLA-B*52, and HLA-C, only the peptides overlapping with the peptides of the respective monoallelic cell line, were selected. In case of HLA-C*03 and C*07 an alignment with the eluted peptides of both cell lines combined, C*03:03/04 and C*07:01/02, was performed.*



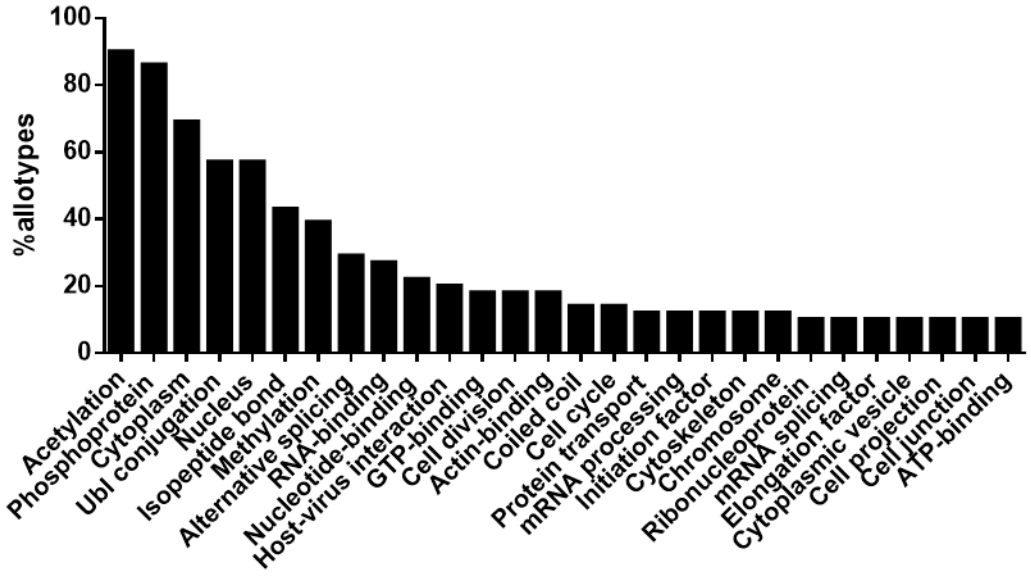
Supplemental Figure S15: Peptide motifs of the Top20_AP for each HLA. *Peptide motifs generated from the Top20_AP for each allotype. The peptide contaminants from the peptide motifs of the Top20_CP could be removed by purification to the Top20_AP.*

Allotype	Length	%	Allotype	Length	%	Allotype	Length	%	Allotype	Length	%	Allotype	Length	%
	8	5		8	5		8	60		8	0		8	25
	9	75		9	60		9	40		9	80		9	70
A01	10	15	A31	10	15	B18	10	0	B50	10	20	C05	10	5
	11	5		11	20		11	0		11	0		11	0
	12	0		12	0		12	0		12	0		12	0
	8	0		8	0		8	0		8	40		8	0
	9	95		9	85		9	90		9	55		9	100
A02	10	0	A32	10	5	B27	10	10	B51	10	5	C06	10	0
	11	0		11	10		11	0		11	0		11	0
	12	5		12	0		12	0		12	0		12	0
	8	0		8	5		8	0		8	75		8	15
	9	65		9	60		9	90		9	25		9	85
A03	10	25	A33	10	15	B35	10	10	B52	10	0	C07	10	0
	11	5		11	20		11	0		11	0		11	0
	12	5		12	0		12	0		12	0		12	0
	8	10		8	0		8	40		8	0		8	5
	9	60		9	75		9	60		9	55		9	90
A11	10	5	A68	10	10	B37	10	0	B55	10	40	C08	10	5
	11	25		11	10		11	0		11	5		11	0
	12	0		12	5		12	0		12	0		12	0
	8	5		8	0		8	0		8	0		8	5
	9	45		9	85		9	70		9	65		9	95
A23	10	35	A69	10	15	B38	10	30	B57	10	15	C12	10	0
	11	15		11	0		11	0		11	20		11	0
	12	0		12	0		12	0		12	0		12	0
	8	0		8	0		8	20		8	5		8	30
	9	90		9	70		9	80		9	75		9	65
A24	10	10	B07	10	15	B39	10	0	B58	10	15	C14	10	5
	11	0		11	15		11	0		11	5		11	0
	12	0		12	0		12	0		12	0		12	0
	8	0		8	25		8	5		8	0		8	0
	9	50		9	75		9	55		9	85		9	100
A25	10	35	B08	10	0	B40	10	30	C01	10	15	C15	10	0
	11	15		11	0		11	10		11	0		11	0
	12	0		12	0		12	0		12	0		12	0
	8	0		8	15		8	0		8	0		8	5
	9	70		9	80		9	80		9	90		9	95
A26	10	15	B13	10	5	B41	10	15	C02	10	10	C16	10	0
	11	15		11	0		11	5		11	0		11	0
	12	0		12	0		12	0		12	0		12	0
	8	0		8	35		8	0		8	0		8	25
	9	90		9	65		9	60		9	85		9	70
A29	10	10	B14	10	0	B44	10	30	C03	10	15	C17	10	0
	11	0		11	0		11	10		11	0		11	5
	12	0		12	0		12	0		12	0		12	0
	8	5		8	15		8	15		8	5			
	9	55		9	75		9	80		9	90			
A30	10	20	B15	10	10	B49	10	5	C04	10	5			
	11	20		11	0		11	0		11	0			
	12	0		12	0		12	0		12	0			

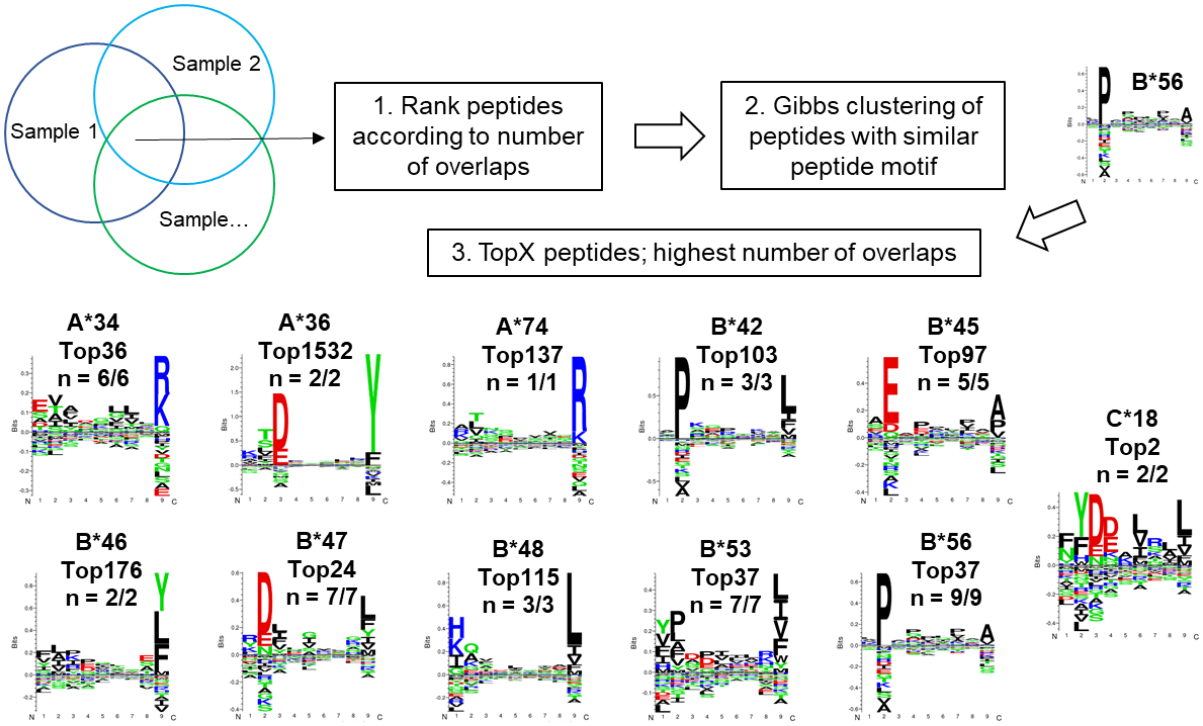
Supplemental Figure S16: Peptide length distribution of the allotypic peptides. *Peptide length distribution of the Top20_AP for each HLA allotype. Most allotypes present a high proportion of nonamers. Some allotypes present a higher proportion of octamers such as HLA-B*52 or deca- and undecamers such as B*35.*



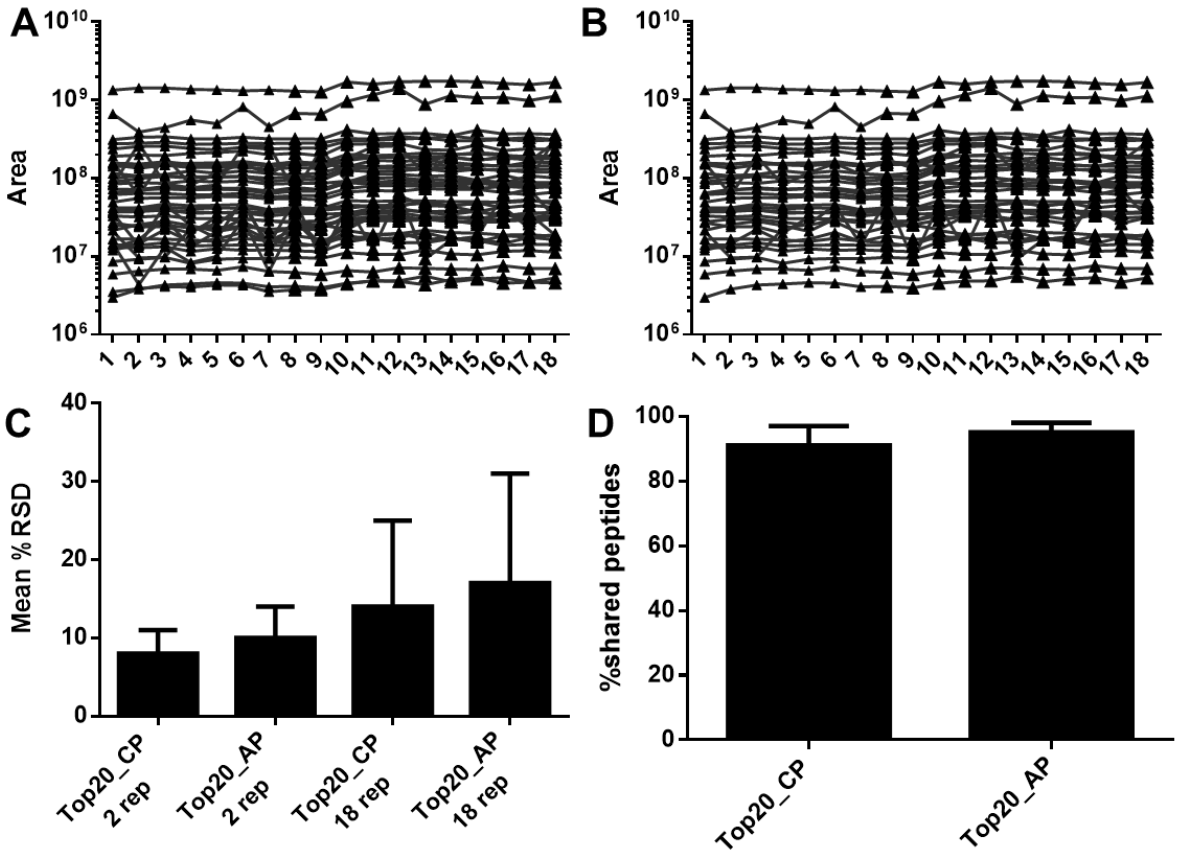
Supplemental Figure S17: Cytoplasmatic protein abundance and turnover rates of Top20_CP and Top20_AP source proteins. *Analysis of protein abundance and protein turnover by overlapping Top20_CP (blue) and AP (green) source proteins with proteome studies of Hela cells (black). For the graphs only proteins with cytoplasmatic abundance and turnover rate above zero were used. Protein abundance: Top20_CP = 495 mapped proteins (A), Top20_AP = 510 mapped proteins (B); protein turnover rates: Top20_CP = 395 mapped proteins (C), Top20_AP = 365 mapped proteins (D).*



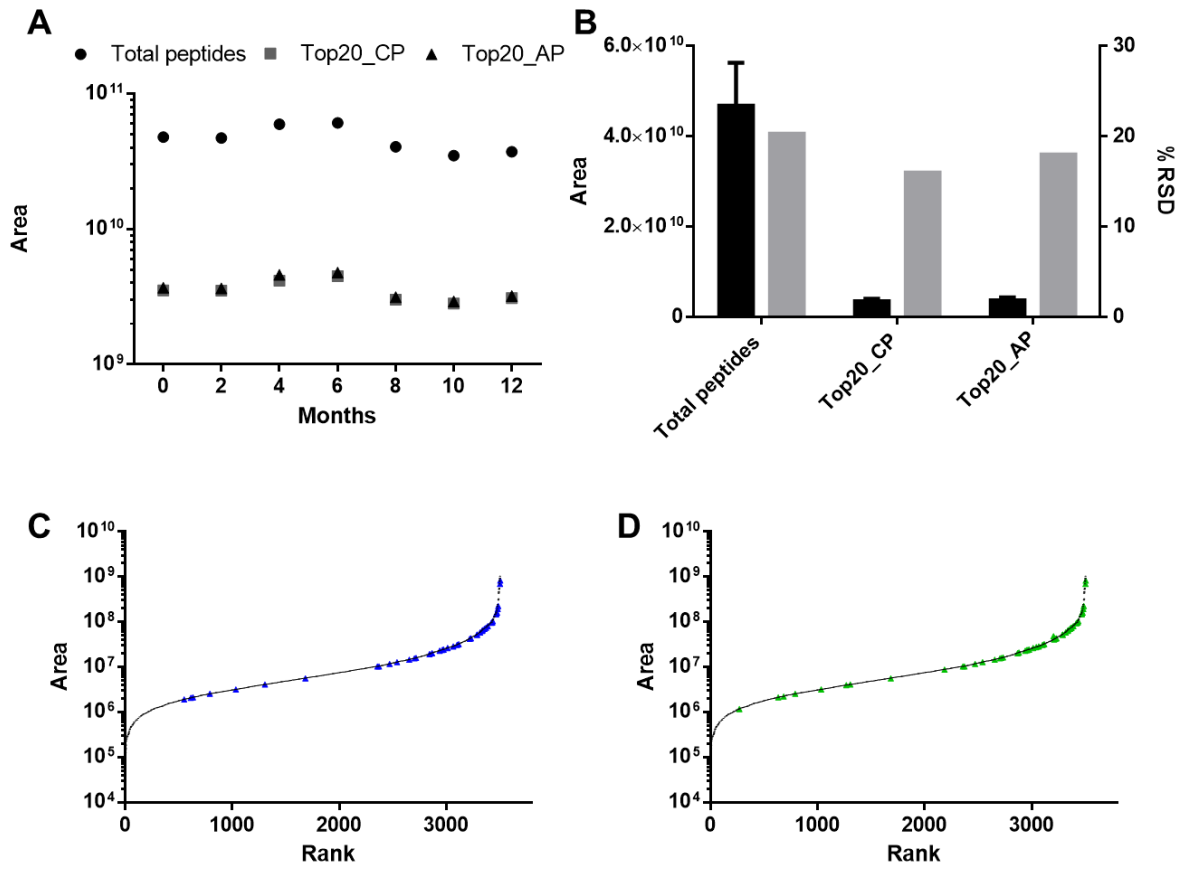
Supplemental Figure S18: UniProt keyword annotations of the Top20_AP source proteins. *Analysis of the UniProt keyword annotations of the Top20_AP source proteins (only annotations with ≥ 5 proteins (10%)).*



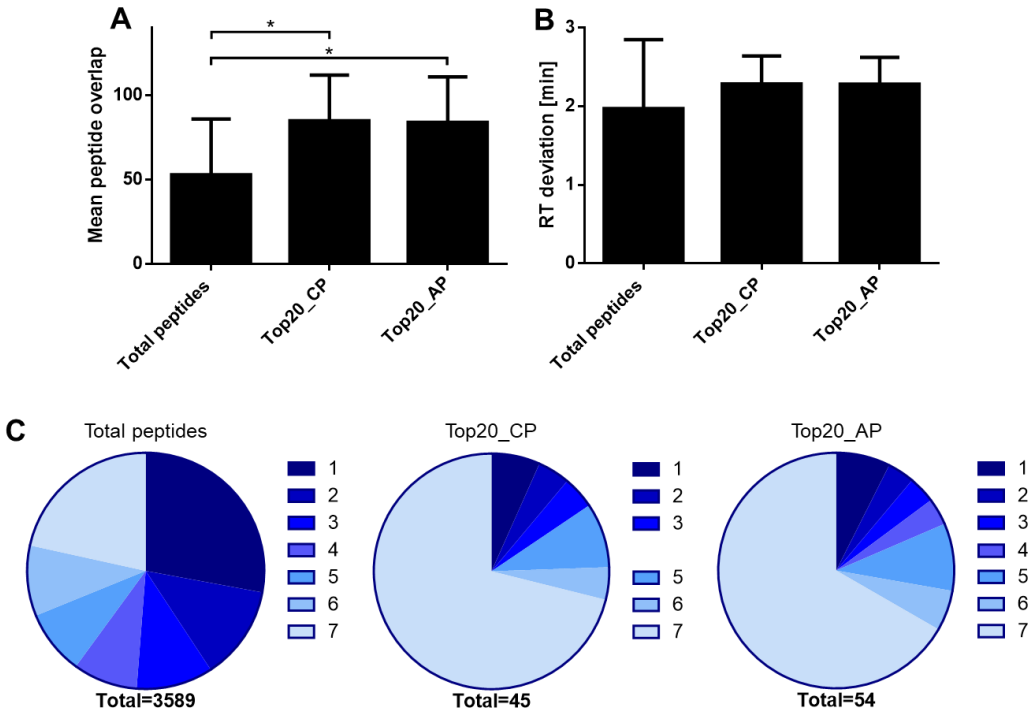
Supplemental Figure S19: TopX peptide lists of Infrequent allotypes. *Generation of HLA specific TopX peptide lists for infrequent allotypes not included in RF training (≤ 10 allotype positive samples in 2-d-DB). A prioritization based on the frequency of the peptides on ≥ 2 HLA positive samples and the peptide motif was performed. For each allotype, the peptide motif, the contained TopX peptides and the number of allotype positive and peptide presenting samples (n) are given.*



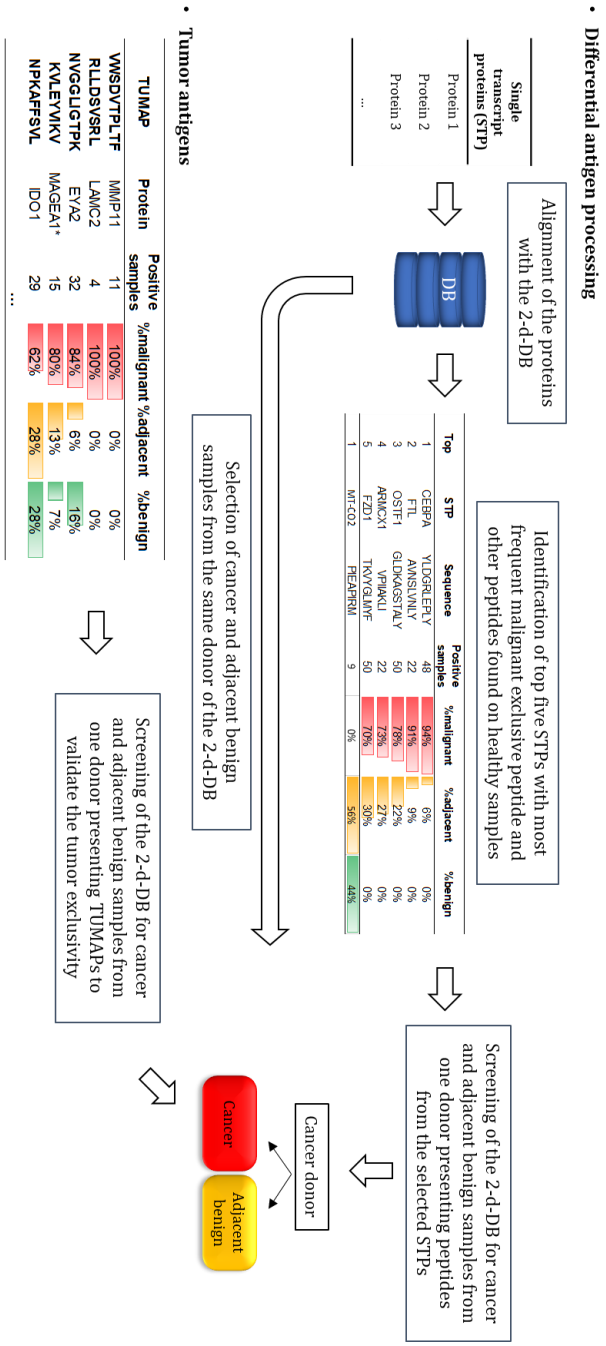
Supplemental Figure S20: Technical influence on the Top20 peptide identification and unlabeled semi-quantification in 18 replicates. *Based on a JY eluted peptide batch measured in 18 replicates over two weeks the technical influence on the Top20_CP and AP peptide identification and unlabeled semi-quantification was investigated. Area of peptides identified in all 18 replicates for Top20_CP (n = 33; A) and Top20_AP (n = 41; B). Mean %RSD of peptide area of two replicates each or all replicates combined for Top20_CP and AP (C). Only peptides identified in all 18 replicates were considered (A-C). The percentage of shared peptides between two replicates each (D).*



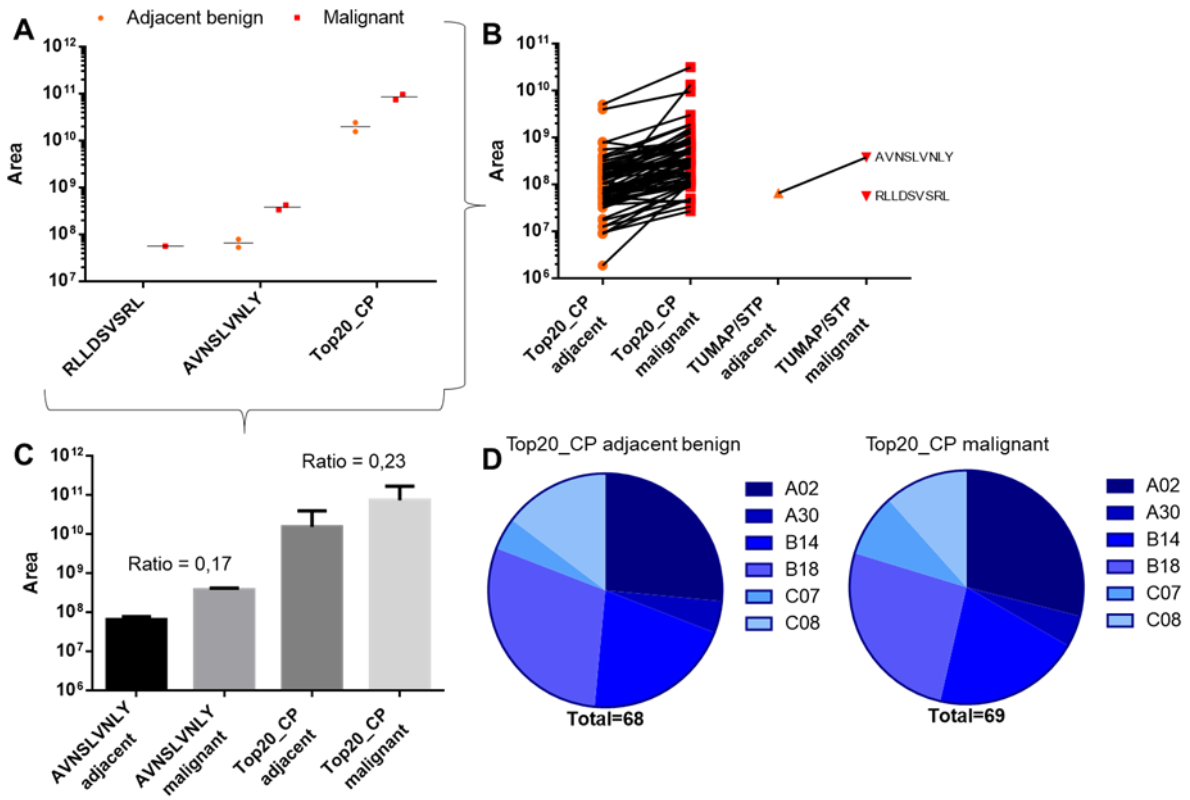
Supplemental Figure S21: Performance of the Top20 peptides over a year. *The Top20_CP and AP were retroperspectively analyzed in a JY eluted peptide batch measured in seven replicates over one year. Sum of the area of all identified peptides, the Top20_CP (n = 44) and AP (n = 53) over the course of a year (A). Average area and %RSD of all peptides and the Top20 in a year (B). Rank and area of the Top20_CP (blue; C) and AP (green; D) highlighted compared to the total peptides (black). Only peptides identified in all seven replicates were considered for Top20_CP and AP (A-D).*



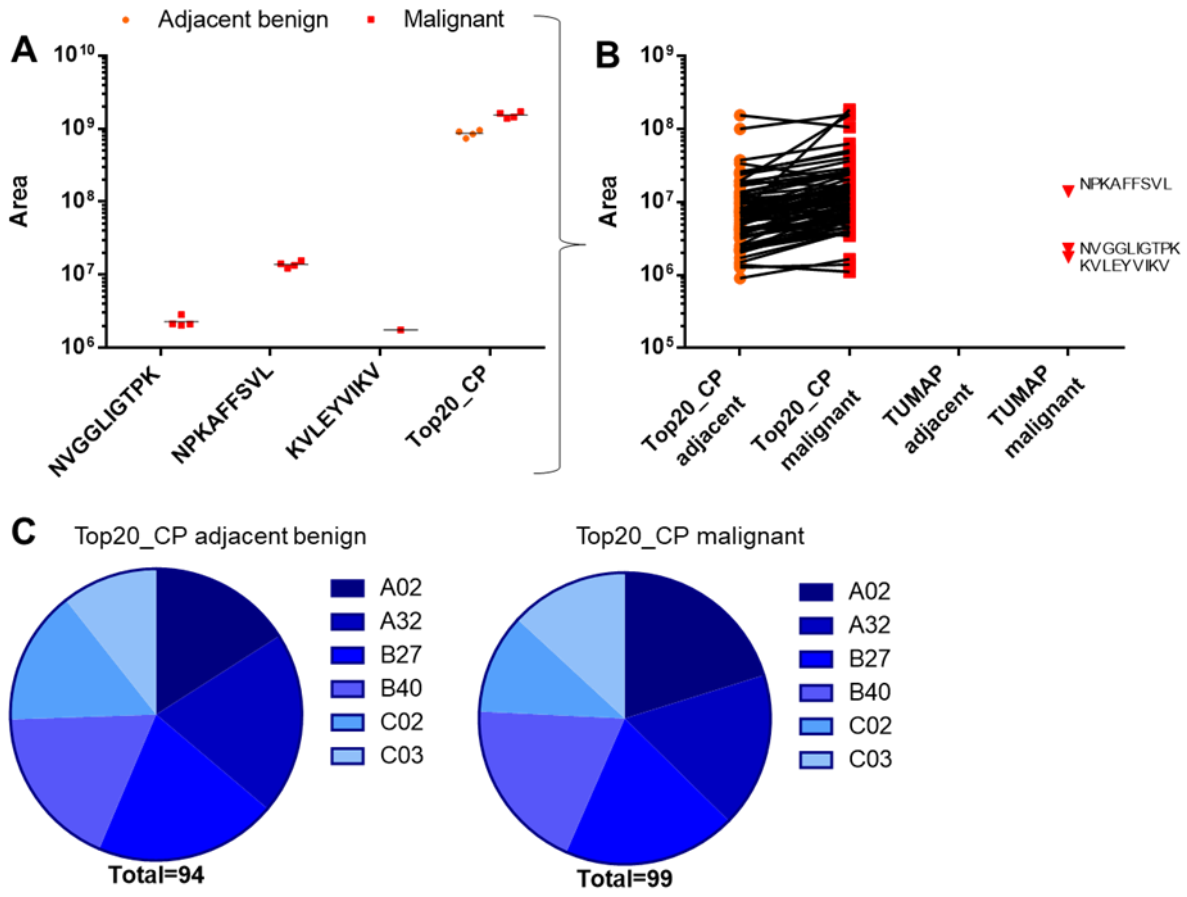
Supplemental Figure S22: Identification of the Top20 peptides over a year. *Peptide overlap of the Top20_CP and AP between seven replicates of a JY eluted peptide batch measured over one year was performed and the mean peptide overlap (A) and the proportion of overlaps depicted (C). Deviation of the liquid chromatography retention time (RT) of the peptides shared in all seven replicates (B).*



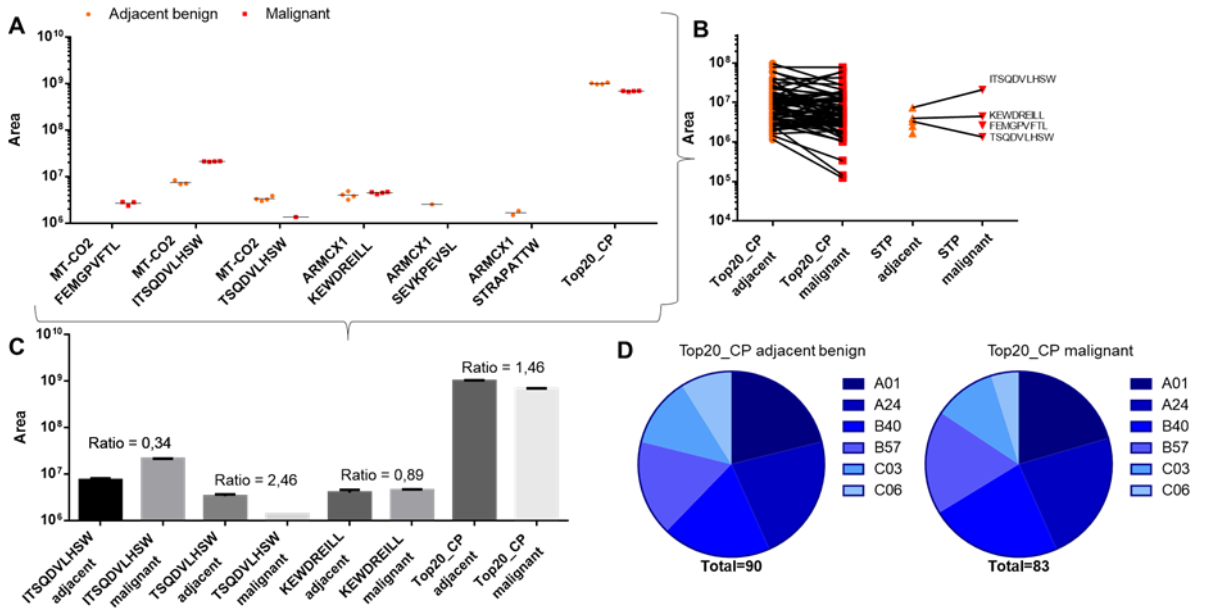
Supplemental Figure S23: Screening for differentially processed antigens and tumor antigens. *To screen for peptides modulated in tumors, which result from DAP, the top five single transcript proteins (STPs retrieved from ensemble.org, 11.2016) were filtered out, which were most frequently identified via a malignant-specific peptide and an additional benign-specific peptide. In addition, the top STP MT-CO2, which was most frequently identified via a benign-specific peptide and other peptides on malignant samples was selected (only the 5 malignant-specific or the benign peptide are depictured). The 2-d-DB was screened for exemplary cancer and adjacent benign samples presenting the desired peptides. Additionally, these samples were also screened for tumor associated peptides from established tumor antigens, to validate their tumor exclusivity. *MAGEA1 is both a tumor antigen and STP.*



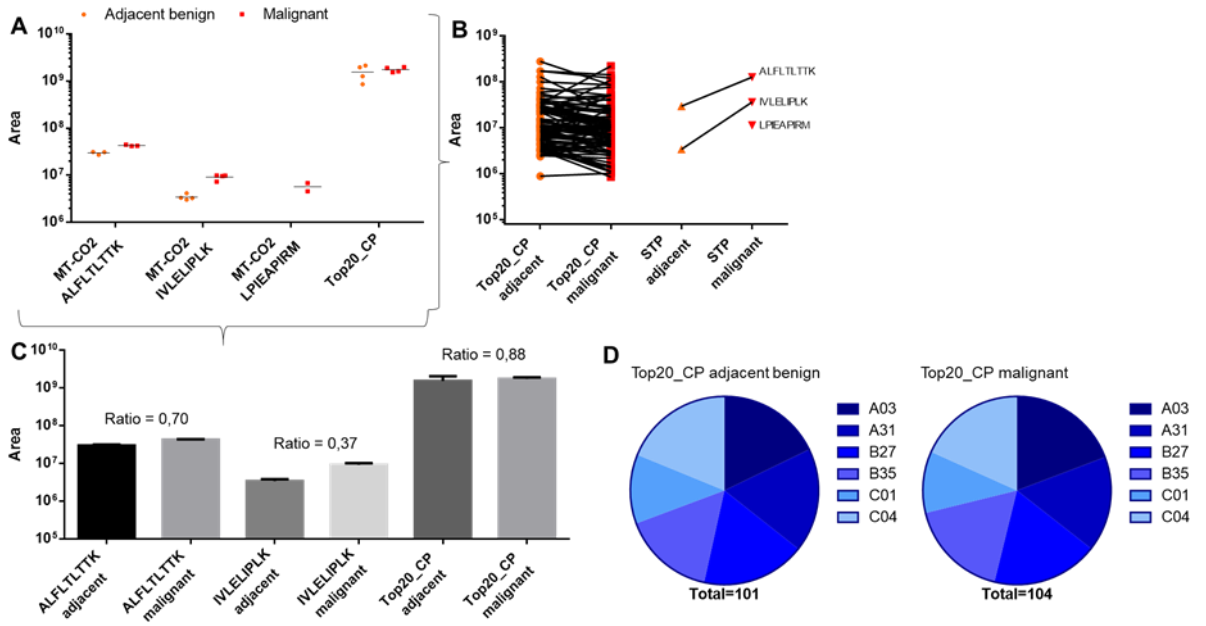
Supplemental Figure S24: TUMAPs and DAP in non-small-cell lung carcinoma and adjacent benign tissue. *Area of the identified TUMAP (RLLDSVSRL), STP peptide (AVNSLVNLY) and the Top20_CP of each replicate (total replicates: n = 2) (A). Individual peptide area of the TUMAP, STP peptide and Top20_CP (B). Ratio of the mean area of the peptides (C). The proportion of the allotype-specific Top20_CP (D).*



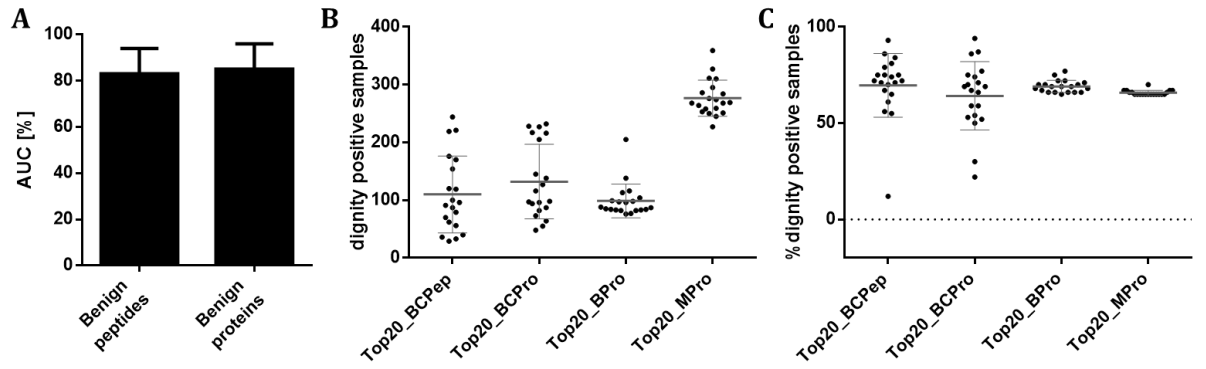
Supplemental Figure S25: TUMAPs and DAP in gastric cancer and adjacent benign tissue. *Area of the identified TUMAPs (NVGGLIGTPK, NPKAFFSVL and KVLEYVIKV) and the Top20_CP of each replicate (total replicates: n = 4) in the adjacent benign and malignant sample (A). Individual peptide area of the TUMAPs and Top20_CP (B). The proportion of the allotype-specific Top20_CP (C).*



Supplemental Figure S26: DAP in renal cell carcinoma I and adjacent benign tissue. *Area of the identified STP (MT-CO2 and ARMCX1) peptides and the Top20_CP of each replicate (total replicates: n = 4) (A). Individual peptide area of the STP peptides and Top20_CP (B). Ratio of the mean area of the peptides (C). The proportion of the allotype-specific Top20_CP (D).*



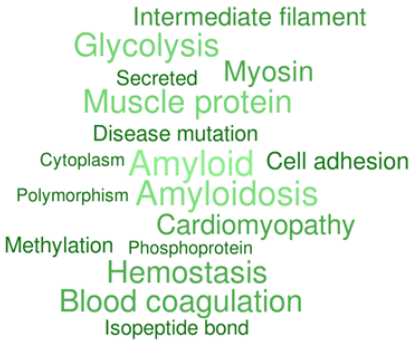
Supplemental Figure S27: DAP in renal cell carcinoma II and adjacent benign tissue. *Area of the identified STP (MT-CO2) peptides and the Top20_CP of each replicate (total replicates: n = 4) (A). Individual peptide area of the STP peptides and Top20_CP in the adjacent benign and malignant sample (B). Ratio of the mean area of the peptides (C). The proportion of the allotype-specific Top20_CP (D).*



Supplemental Figure S28: Dignity classification antigens. *Identification of dignity specific antigens based on benign samples from tumor-free donors in the 2-d-DB. The classification was carried out at peptide (benign peptides) and source protein (benign proteins) level and the AUC's of the classifications in the 2-d-DB were determined (A). Assessment of the number of benign samples presenting the individual peptides (B) and the percentage of benign samples among the peptide presenting samples (C) of the individual Top20 benign classification peptides (Top20_BCPep), as well as the individual Top20 benign classification proteins (Top20_BCPro). Additionally, the manually generated Top20 lists, that are at least 65% positive for benign (Top20_BPro) and malignant (Top20_MPro) samples were considered.*

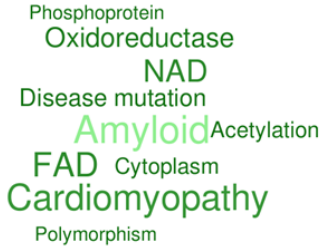
Top20_BCPro

Term	Fold Enrichment
Brain ischemia, stroke, vascular diseases	117
Atherosclerosis, brain ischemia, carotid stenosis, thrombosis	117
Hypertrophic cardiomyopathy	109
Thrombosis	42
Thromboembolism, venous	37
Cardiomyopathy	36



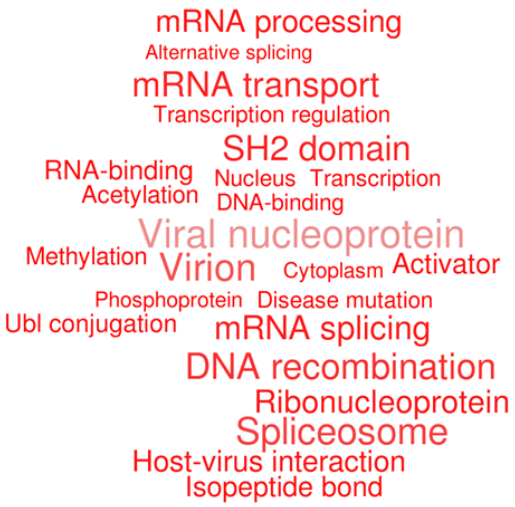
Top20_BPro

Term	Fold Enrichment
Familial amyloid polyneuropathy	763
Alcohol abuse	35
Diabetic nephropathy	33
Hypercholesterolemia	32
Alcohol	22
Null	4
Type 2 diabetes, edema, rosiglitazone	3

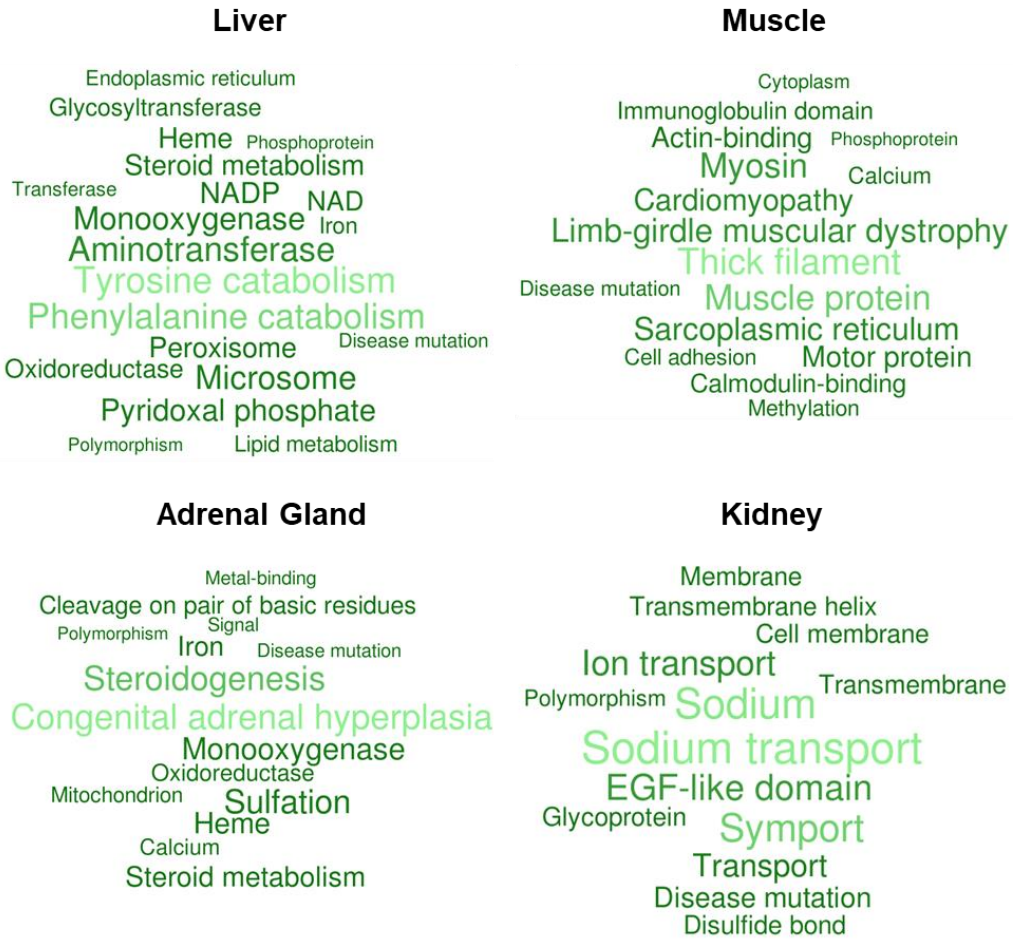


Top20_MPro

Term	Fold Enrichment
Myeloid leukemia	64
Kidney cancer	60
Brain cancer	58
Brain neoplasms, glioma	26
Head and neck cancer	19
Birth weight, leukemia, leukemia, myeloid, Acute, precursor cell lymphoblastic leukemia-lymphoma	13
Meningeal neoplasms, meningioma	12
Lung cancer	8
breast cancer	7
Bladder cancer	7
Chronic obstructive pulmonary disease	6
Multiple sclerosis	6
Null	4
Tobacco use disorder	2



Supplemental Figure S29: Annotations of the dignity classification antigens. *Disease annotations (table) and functional UniProt keyword annotations (word cloud) of the Top20_BCPro, Top20_BPro and Top20_Mpro. Proteins were annotated using DAVID david.ncifcrf.gov and word clouds, word size depending on fold enrichment, were generated using worditout.com.*



Supplemental Figure S30: Annotations of tissue classification antigens*. Functional UniProt keyword annotations (word cloud) of the Top20 tissue proteins. Proteins were annotated using DAVID david.ncifcrf.gov and word clouds, word size depending on fold enrichment, were generated using worditout.com.*

6 Current state of research and outlook

6.1 Current status

This doctoral thesis and the increasing research interest in the field of immunopeptidomics demonstrate that the immunopeptidome is a fascinating "Wikipedia" containing an unimaginable wealth of information which offers us numerous new insights and will not only revolutionize tumor and virus therapy, but also provide countless other possibilities. Impressive examples of the variety and range of possible applications of immunopeptidomic analysis are reflected in the manifold projects during this doctorate, both in research and the GMP quality control for clinical applications, which I was able to work on thanks to the Stevanović working group (WG) in the Department of Immunology, Tübingen and Biochemistry WG in the NMI Natural and Medical Sciences Institute at the University of Tübingen, Reutlingen.

6.1.1 Research

An evolution in immunopeptidomics, that I witnessed in the field, was the leap from thousands of peptide identifications to tens of thousands of peptide identifications per sample, by supplementing our LTQ Orbitrap XL with an Orbitrap Fusion Lumos (Thermo Fisher Scientific, MA, USA). I also replaced the time of flight Q-Tof (Waters, MA, USA) instrument with a modern LTQ Orbitrap XL in the mass spectrometric GMP quality control of the Wirkstoffpeptidlabor and we experienced a significant boost in high mass resolution and a much better distinction between similar compounds.

The scope and possibilities of immunopeptidomic can best be demonstrated through the more than 20 different cooperations I had in the last three years. The publication of most projects is still in the distant future, already published or drafted publications are mentioned in brackets.

I had the great opportunity to experience the process from bench to bedside in two compassionate use vaccinations. In collaboration with the entire Stevanović research group and the Wirkstoffpeptidlabor I was able to predict human papillomavirus (HPV) epitopes, which were synthesized and vaccinated in compassionate use to treat HPV induced skin warts. In cooperation with the University Hospital Tübingen, Tübingen, and the CeGaT GmbH, Tübingen, a vaccination against a submandibular salivary gland tumor was conducted, where tumor specific antigens were identified after immunopeptidome analysis.

In addition to human samples in my doctoral thesis I examined the immunopeptidome of stem cells in murine samples with the Haas working group (WG) (Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM), Heidelberg; manuscript submitted), identified peptide motifs of chicken MHC allotypes and generation of crystal structures with the Kaufmann WG (University of Cambridge, UK) and Härtle WG (Ludwig-Maximilians-University,

Munich) and analyzed the presented peptides in canine samples in cooperation with the Planz WG (University of Tübingen, Tübingen).

I had the opportunity to work with almost ten different types of viruses. Solely in the institute I could cooperate with the Planz WG and discover different influenza epitopes (draft in progress), decode Orf virus (ORFV) epitopes in cooperation with the Amann WG and could show that after vaccination no immune reaction develops against these viruses (Reguzowa et al, publication accepted in Vaccines). Furthermore, I was able to work with HPV in the previously mentioned HPV vaccination and in cooperation with Maren Lübke (Stevanović WG), Jonjic WG (University of Rijeka, Croatia) and Erhard WG (University of Würzburg) I could investigate human cytomegalovirus (HCMV) epitopes. In Tübingen I could also identify measles virus epitopes in a study with the Tabatabai WG (Interdisciplinary Division of Neuro-Oncology University Hospital Tübingen) and could show the advantages of treatment with oncological measles viruses in combination with radiation [63]. For the identification of murine cytomegaloviruses (MCMV) epitopes I had two projects with the WG Jonjic (University of Rijeka) and the WG Schönrich (Charité - Universitätsmedizin Berlin, Berlin). In cooperation with the Solimena WG (Paul Langerhans Institute Dresden of the Helmholtz Center Munich at the University Clinics and Medical Faculty Carl Gustav Carus at TU Dresden, Dresden) I could investigate coxsackieviruses in connection with the pathogenesis of diabetes mellitus type 1. An intensive collaboration with the Kaufmann WG (University of Cambridge) and Härtle WG (Ludwig-Maximilians-University) enabled the identification of the infectious bursal disease virus (IBDV) and Marek's disease virus (MDV) epitopes (publication submitted in PNAS).

Besides the identification of viral epitopes, immunopeptidomics is mainly used for the identification of tumor antigens. In addition to the above mentioned tumor types, I was able to examine glioblastoma and atypical teratoid rhabdoid tumors (ATRT) in cooperation with the Tabatabai WG. In addition, I was able to identify post-translationally modified peptides in ovarian cancer in collaboration with the Santambrogio WG (Albert Einstein College of Medicine, New York) and I could identify tumor antigens in renal cell carcinomas in cooperation with the Sester WG (Saarland University Medical Center and Saarland University Faculty of Medicine). The identified tumor antigens can be administered as peptide vaccination, as in the compassionate use case described above, or as in the latter cooperation for ACT, in which the peptides are used to isolate tumor-specific T cells, expanded *in vitro* and administered back to the patient.

An exciting new field of immunopeptidomics is the search for neoantigens, e.g. peptides derived from in tumor mutated antigens. In cooperation with the Gros WG (Vall d'Hebron Institute of Oncology (VHIO), Barcelona) I was able to identify neoantigens in hematological tumors. In collaboration with the WG Apcher (National Institute of Health and Medical Research, Paris) I was

able to detect a special type of neoantigens, peptides of unspliced antigens, that occur in tumors after splicing inhibition [87].

In addition to the application-related projects, I was also able to employ immunopeptidomics for basic research, for example in collaboration with the Stevanović WG for the deciphering of previously unknown HLA peptide motifs using monoallelic cells and multiallelic tissue [154]; (Chapter 4 and 5) or the influence of individual proteins (T6BP) in the MHC class II presentation pathways with the Arnaud WG (Institut Curie, Paris; draft in progress).

This variety of projects shows how broad the immunopeptidomic applications are currently ranging from basic research to applied research from bench to bedside, although there is still no approved immunopeptidome-derived therapy yet.

6.1.2 Clinical application

Peptide vaccinations have been used in clinical application for a long time and its efficacy has been demonstrated [208,209]. Since 2004 the peptide vaccines must be produced under GMP conditions and a manufacturing license is required. I was allowed to work in the GMP vaccine peptide production facility, the Wirkstoffpeptidlabor. As described in Chapter 2.7., GMP requires the analytical method to be suitable and reliable for your analytical procedures. These requirements have evolved from previous experience with active substances, mainly from analytical procedures in which individual analytes were considered. In the current research environment, however, omic technologies that can be used to examine several analytes at once are playing an increasingly important role, especially for medical diagnosis based on biomarkers [210]. So far there were already proposals for method validation of genomic and transcriptomic technologies, but not for LC-MS/MS based omic technologies [211,212]. In this thesis first guidelines for the method validation of LC-MS/MS based immunopeptidomics were proposed [116], which can also be used as inspiration for the validation of proteomics, lipidomics or metabolomics technologies.

As discussed in Chapter 2.6.1., peptide vaccinations need the support of adjuvants to induce an immune response. In another validation, I was able to develop a protocol for the detection of equal antigen distribution in a vaccine syringe with an antigen-adjuvant emulsion [108] and to develop and validate a LC-MS/MS method to enable the analysis of the new adjuvant XS15, for which the Wirkstoffpeptidlabor recently received manufacturing approval [113]. Besides the LC-MS/MS method development for the adjuvant, I was also able to develop a LC-MS/MS based protocol in cooperation with the Jung WG (University of Tübingen, Tübingen) to successfully detect the proportion of antibody compounds in resulting products of a GMP antibody batch in a protein gel (draft in progress).

Today, there are many planned and ongoing clinical trials using active ingredients, which are based on antigens previously identified through immunopeptidomics, such as peptide vaccines, ACT or bispecific T cell engaging receptors (TCER™). In this doctoral thesis, thanks to the Wirkstoffpeptidlabor I was able to collaborate in the development of peptide vaccines from natural tumor-specific self-peptides (GAPVAC [121], NCT02149225; iVAC-CLL01, NCT02802943 ; PepIVAC-01, IVAC-XS15-CLL01), or neoantigens (NOA-16, NCT02454634; IVAC-ALL-1, NCT03559413, AMPLIFY-NEOVAC, NCT03893903) in clinical studies. Within the framework of ACTolog (NCT02876510), our peptides were also used for the isolation of T cells for ACT and, in addition to the adjuvants used previously, new promising XS15 studies will start with the recently obtained XS15 manufacturing authorization (IVAC-XS15-CLL01).

In the duration of this doctorate only, almost 5000 measurements of test instructions for GMP peptides or adjuvants were performed in our mass spectrometric quality control department, with the majority of analytes originating from immunopeptidome-based projects. This illustrates the current relevance of immunopeptidomics in clinical application.

6.2 Outlook

Significant advances in sequencing and LC-MS/MS technologies, in addition to improved and diverse bioinformatic pipelines, will continue to enhance peptide yields and provide deeper insights into the immunopeptidome [205]. Thus, the repertoire of tumor-associated antigens will be increased and optimized, mutated neoantigens will be detected more effectively and the epitope discovery of more pathogens will be possible. In addition, all areas will be expanded with cryptic peptides from non-encoded open-reading frames in the genome [86]. This will develop immunotherapy by numerous targets.

In addition to therapeutic targets, immunopeptidomics will also be used in other areas to exploit the sensitive biomarker system. The amount of measured immunopeptidomic data will continue to increase. Large databases will allow immunoinformatics to improve the determination of HLA allotypes, tissue type and dignity of samples as described in Chapter 5, and to go deeper and identify peptide patterns caused by altered cellular pathways or cellular stress, which will eventually lead to the identification of biomarkers that could be used for example in T cell response quick test [213]. A cellular test system could be developed to detect and predict cellular influences of biomaterials *in vitro* to minimize animals testing. Which peptide patterns appear in inflammations, which in cellular stress and which are suspected to trigger autoimmune diseases?

Artificial intelligence and more sophisticated algorithms permit to keep track of alterations of multiple peptides at once. Complex cellular changes, such as the gradual alterations in cell aging and the transition to senescence, could be tracked externally through detailed peptide patterns at the immunopeptidome level. A detailed immunopeptidomic investigation of non- and senescent

cells could allow the identification of such biomarkers that could also be used for the therapeutic elimination of senescent cells, which has already led to significant improvements in vitality and lifespan of mice [214].

While writing this dissertation, the SARS-CoV-2 pandemic is keeping the world in suspense. More than hundred companies and institutes worldwide are working on a SARS-CoV-2 vaccine including multiple innovative approaches, such as RNA-based vaccines [215,216], of which there is no approved drug yet. The example of SARS-CoV-2 can demonstrate how a vaccine based on the immunopeptidome may be produced in future using the methods presented in this dissertation. The developed validation of the LC-MS/MS based immunopeptidomic in Chapter 3 allows to reliably identify SARS-CoV-2 epitopes e.g. in virus infected cells, as already shown in the case of measles [63]. Furthermore, the protocol described in Chapter 4 for the identification of MHC peptide motifs and the generation of MHC prediction matrices can theoretically predict SARS-CoV-2 epitopes for the MHC of all vertebrates. The discovered epitopes could be used for therapies against SARS-CoV-2 [217]. Using the method described in Chapter 5, a database of immunopeptidomes from SARS-CoV-2 infected and non-infected samples might be used to develop a classification that enables the detection of infected tissue. Furthermore, the antigens can be determined, which might be used as biomarkers for the identification of SARS-CoV-2 infections.

An advantage of possible immunopeptidome based SARS-CoV-2 therapies is the usage of specific T cell epitopes, which minimizes the potential risk of side effects such as autoimmune diseases [218]. Even the most innovative current approaches from vaccine manufacturers against SARS-CoV-2 are based on whole proteins instead of the individual peptide epitopes [215,216]. Thus, the next step after the development of protein-based vaccines might be the development towards epitope-reduced vaccines [219].

7 Abbreviations

aa	amino acid
ACT	adoptive T cell transfer
APC	antigen presenting cell
AUC	area under the curve
β2m	β2-microglobulin
CD	cluster of differentiation
CTL	cytotoxic T lymphocyte
DAMPs	Damage-Associated Molecular Patterns
DC	dendritic cell
DRiP	defective ribosomal product
EMA	European Medicines Agency
ER	endoplasmic reticulum
FDA	Food and Drug Administration
FDR	false discovery rate
GMP	Good Manufacturing Practice
HCMV	human cytomegalovirus
HLA	human leukocyte antigen
HPV	human papillomavirus
IBDV	infectious bursal disease virus
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IFN	interferon
IL	interleukin
LC	liquid chromatography
LC-MS/MS	liquid chromatographic tandem mass spectrometry

LOD	limit of detection
LOQ	limit of quantification
MCMV	murine cytomegaloviruses
MDV	Marek's disease virus
MHC	major histocompatibility complex
MIIC	MHC class II compartment
MPLA	monophosphoryl lipid A
mRNA	messenger ribonucleic acid
NK cell	natural killer cell
ORFV	Orf virus
PAMPs	pathogen-associated patterns
PBMC	peripheral blood mononuclear cell
pHLA	peptide-HLA complex
PLC	peptide loading complex
PRRs	pattern recognition receptors
RCC	Renal cell carcinoma
ROC	receiver operating characteristic
RT	room temperature
TAP	transporter associated with antigen processing
TCR	T cell receptor
T_H cell	T helper cell
TLR	Toll-like receptor
Treg cell	regulatory T cell
TUMAP	tumor-associated peptide
WG	working group

8 References

1. Parkin, J. & Cohen, B. An overview of the immune system. *Lancet* **357,** 1777–1789 (2001).
2. Dranoff, G. Cytokines in cancer pathogenesis and cancer therapy. *Nat. Rev. Cancer* **4,** 11–22 (2004).
3. Rasmussen, S. B., Reinert, L. S. & Paludan, S. R. Innate recognition of intracellular pathogens: detection and activation of the first line of defense. *APMIS* **117,** 323–337 (2009).
4. Medzhitov, R. & Janeway, C. Innate immune recognition: Mechanisms and pathways. *Immunol. Rev.* **173,** 89–97 (2000).
5. Tang, D., Kang, R., Coyne, C. B., Zeh, H. J. & Lotze, M. T. PAMPs and DAMPs: Signal 0s that spur autophagy and immunity. *Immunol. Rev.* **249,** 158–175 (2012).
6. Rossi, M. & Young, J. W. Human Dendritic Cells: Potent Antigen-Presenting Cells at the Crossroads of Innate and Adaptive Immunity. *J. Immunol.* **175,** 1373–1381 (2005).
7. Matzinger, P. Tolerance, Danger, and the Extended Family. *Annu. Rev. Immunol.* **12,** 991–1045 (1994).
8. Borghans JA, Noest AJ, D. B. R. How Specific Should Immunological Memory Be? *J Immunol.* **163,** 569–575 (1999).
9. Medzhitov, R. & Janeway, C. A. Innate immunity: The virtues of a nonclonal system of recognition. *Cell* **91,** 295–298 (1997).
10. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* **14,** 377–91 (2014).
11. den Haan, J. M. M., Arens, R. & van Zelm, M. C. The activation of the adaptive immune system: cross-talk between antigen-presenting cells, T cells and B cells. *Immunol. Lett.* **162,** 103–12 (2014).
12. Grakoui, A., Bromley, S. K., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M. & Dustin, M. L. The immunological synapse: A molecular machine controlling T cell activation. *Science* **285,** 221–227 (1999).
13. Janeway, C. A. The T Cell Receptor as a Multicomponent Signalling Machine: CD4/CD8 Coreceptors and CD45 in T Cell Activation. *Annu. Rev. Immunol.* **10,** 645–674 (1992).
14. Schwartz, R. H. T cell clonal anergy. *Curr. Opin. Immunol.* **9,** 351–357 (1997).
15. Malissen, B. & Bongrand, P. Early T cell activation: integrating biochemical, structural, and biophysical cues. *Annu. Rev. Immunol.* **33,** 539–61 (2015).
16. Pennock, N. D., White, J. T., Cross, E. W., Cheney, E. E., Tamburini, B. A. & Kedl, R. M. T cell responses: naive to memory and everything in between. *Adv. Physiol. Educ.* **37,** 273–83 (2013).
17. Curtsinger, J. M., Schmidt, C. S., Mondino, A., Lins, D. C., Kedl, R. M., Jenkins, M. K. & Mescher, M. F. Inflammatory cytokines provide a third signal for activation of naive CD4+ and CD8+ T cells. *J. Immunol.* **162,** 3256–62 (1999).
18. Turner, S. J., Doherty, P. C., McCluskey, J. & Rossjohn, J. Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* **6,** 883–894 (2006).
19. Kaech, S. M., Wherry, E. J. & Ahmed, R. Effector and memory T-cell differentiation: Implications for vaccine development. *Nat. Rev. Immunol.* **2,** 251–262 (2002).
20. Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nat. Rev. Immunol.* **15,** 149–159 (2015).

21. Murphy, K. M., Weaver, C., Janeway, C. & Seidler, L. *Janeway Immunology*. (2018).

22. Klein, J. & Sato, A. The HLA System. *N. Engl. J. Med.* **343,** 702–709 (2000).

23. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11,** 823–36 (2011).

24. Vizcaíno, J. A., Kubiniok, P., Kovalchik, K. A., Ma, Q., Duquette, J. D., Mongrain, I., Deutsch, E. W., Peters, B., Sette, A., Sirois, I. & Caron, E. The human immunopeptidome project: A roadmap to predict and treat immune diseases. *Mol. Cell. Proteomics* **19,** 31–49 (2020).

25. Caron E, Charbonneau R, Huppé G, Brochu S, P. C. The Structure and Location of SIMP/STT3B Account for Its Prominent Imprint on the MHC I Immunopeptidome - PubMed. *Int Immunol.* **17,** 1583–1596 (2005).

26. Bjorkman, P. J. & Parham, P. Structure, function, and diversity of class I major histocompatibility complex molecules. *Annu. Rev. Biochem.* **59,** 253–88 (1990).

27. Yaneva, R., Schneeweiss, C., Zacharias, M. & Springer, S. Peptide binding to MHC class I and II proteins: new avenues from new methods. *Mol. Immunol.* **47,** 649–57 (2010).

28. Bouvier, M. & Wiley, D. C. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science* **265,** 398–402 (1994).

29. Matsumura, M., Fremont, D. H., Peterson, P. A. & Wilson, I. A. Emerging Principles for the Recognition of Peptide Antigens by MHC Class I Molecules. *Science* **257,** 927–934 (1992).

30. Rötzschke, O., Falk, K., Stevanović, S., Jung, G. & Rammensee, H.-G. Peptide motifs of closely related HLA class I molecules encompass substantial differences. *Eur. J. Immunol.* **22,** 2453–2456 (1992).

31. Matsui, M., Moots, R. J., McMichael, A. J. & Frelinger, J. A. Significance of the six peptide-binding pockets of HLA-A2.1 in influenza a matrix peptide-specific cytotoxic T-lymphocyte reactivity. *Hum. Immunol.* **41,** 160–166 (1994).

32. Rammensee, H.-G., Bachmann, J. & Stevanović, S. *MHC Ligands and Peptide Motifs*. (Springer, Berlin, Heidelberg, 1997).

33. Chelvanayagam, G. A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics* **45,** 15–26 (1996).

34. Huyton, T., Ladas, N., Schumacher, H., Blasczyk, R. & Bade-Doeding, C. Pocketcheck: Updating the HLA class I peptide specificity roadmap. *Tissue Antigens* **80,** 239–248 (2012).

35. Rammensee, H. G., Friede, T. & Stevanovic, S. MHC ligands and peptide motifs: first listing. *Immunogenetics* **41,** 178–228 (1995).

36. Wu, J. Q., Habegger, L., Noisa, P., Szekely, A., Qiu, C., Hutchison, S., Raha, D., Egholm, M., Lin, H., Weissman, S., Cui, W., Gerstein, M. & Snyder, M. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 5254–5259 (2010).

37. Van Der Velden, A. W. & Thomas, A. A. M. The role of the 5' untranslated region of an mRNA in translation regulation during development. *Int. J. Biochem. Cell Biol.* **31,** 87–106 (1999).

38. Treutlein, B., Gokce, O., Quake, S. R. & Südhof, T. C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E1291–E1299 (2014).

39. Breitbart, R. E., Andreadis, A. & Nadal-Ginard, B. Alternative Splicing: A Ubiquitous Mechanism for the Generation of Multiple Protein Isoforms from Single Genes. *Annu. Rev.*

Biochem. **56,** 467–495 (1987).

40. Brett, D., Pospisil, H., Valcárcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nat. Genet.* **30,** 29–30 (2002).

41. Schubert, U., Antón, L. C., Gibbs, J., Norbury, C. C., Yewdell, J. W. & Bennink, J. R. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **404,** 770–774 (2000).

42. Yewdell, J. W., Reits, E. & Neefjes, J. Making sense of mass destruction: Quantitating MHC class I antigen presentation. *Nat. Rev. Immunol.* **3,** 952–961 (2003).

43. Bulik, S., Peters, B. & Holzhütter, H.-G. Quantifying the Contribution of Defective Ribosomal Products to Antigen Production: A Model-Based Computational Analysis. *J. Immunol.* **175,** 7957–7964 (2005).

44. Hershko, A. & Ciechanover, A. THE UBIQUITIN SYSTEM. *Annu. Rev. Biochem.* **67,** 425–479 (1998).

45. Voges, D., Zwickl, P. & Baumeister, W. The 26S Proteasome: A Molecular Machine Designed for Controlled Proteolysis. *Annu. Rev. Biochem.* **68,** 1015–1068 (1999).

46. Tanaka, K. The proteasome: From basic mechanisms to emerging roles. *Keio J. Med.* **62,** 1–12 (2013).

47. Kobayashi, K. S. & Van Den Elsen, P. J. NLRC5: A key regulator of MHC class I-dependent immune responses. *Nat. Rev. Immunol.* **12,** 813–820 (2012).

48. Schoenborn, J. R. & Wilson, C. B. Regulation of Interferon-γ During Innate and Adaptive Immune Responses. *Adv. Immunol.* **96,** 41–101 (2007).

49. Vigneron, N. & Van den Eynde, B. J. Proteasome subtypes and the processing of tumor antigens: Increasing antigenic diversity. *Curr. Opin. Immunol.* **24,** 84–91 (2012).

50. Gaczynska, M., Rock, K. L. & Goldberg, A. L. Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes. *Nature* **365,** 264–267 (1993).

51. Gray, C. W., Slaughter, C. A. & DeMartino, G. N. PA28 activator protein forms regulatory caps on proteasome stacked rings. *J. Mol. Biol.* **236,** 7–15 (1994).

52. Groettrup, M., Soza, A., Eggers, M., Kuehn, L., Dick, T. P., Schild, H., Rammensee, H. G., Koszinowski, U. H. & Kloetzel, P. M. A role for the proteasome regulator PA28α in antigen presentation. *Nature* **381,** 166–168 (1996).

53. Newey, A., Griffiths, B., Michaux, J., Pak, H. S., Stevenson, B. J., Woolston, A., Semiannikova, M., Spain, G., Barber, L. J., Matthews, N., Rao, S., Watkins, D., Chau, I., Coukos, G., Racle, J., Gfeller, D., Starling, N., Cunningham, D., Bassani-Sternberg, M. & Gerlinger, M. Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class i neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. *J. Immunother. Cancer* **7,** 1–15 (2019).

54. Javitt, A., Barnea, E., Kramer, M. P., Wolf-Levy, H., Levin, Y., Admon, A. & Merbl, Y. Pro-inflammatory cytokines alter the immunopeptidome landscape by modulation of HLA-B expression. *Front. Immunol.* **10,** (2019).

55. Cresswell, P., Androlewicz, M. J. & Ortmann, B. Assembly and transport of class I MHC-peptide complexes. *Ciba Found. Symp.* **187,** 150–62; discussion 162-9 (1994).

56. Stevanovic, S. & Schild, H. Quantitative aspects of T cell activation - Peptide generation and editing by MHC class I molecule. *Semin. Immunol.* **11,** 375–384 (1999).

57. Yewdell, J. W. Not such a dismal science: the economics of protein synthesis, folding, degradation and antigen processing. *Trends Cell Biol.* **11,** 294–7 (2001).

58. Reits, E., Griekspoor, A., Neijssen, J., Groothuis, T., Jalink, K., Van Veelen, P., Janssen, H., Calafat, J., Drijfhout, J. W. & Neefjes, J. Peptide Diffusion, Protection, and Degradation in Nuclear and Cytoplasmic Compartments before Antigen Presentation by MHC Class I. *Immunity* **18,** 97–108 (2003).

59. Guermonprez, P. & Amigorena, S. Pathways for antigen cross presentation. *Springer Semin. Immunopathol.* **26,** 257–271 (2005).

60. Ackerman, A. L. & Cresswell, P. Cellular mechanisms governing cross-presentation of exogenous antigens. *Nat. Immunol.* **5,** 678–684 (2004).

61. Dengjel, J., Schoor, O., Fischer, R., Reich, M., Kraus, M., Müller, M., Kreymborg, K., Altenberend, F., Brandenburg, J., Kalbacher, H., Brock, R., Driessen, C., Rammensee, H. G. & Stevanovic, S. Autophagy promotes MHC class II presentation of peptides from intracellular source proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7922–7927 (2005).

62. Hillen, N. & Stevanovic, S. Contribution of mass spectrometry-based proteomics to immunology. *Expert Rev. Proteomics* **3,** 653–64 (2006).

63. Rajaraman, S., Canjuga, D., Ghosh, M., Codrea, M. C., Sieger, R., Wedekink, F., Tatagiba, M., Koch, M., Lauer, U. M., Nahnsen, S., Rammensee, H.-G., Mühlebach, M. D., Stevanovic, S. & Tabatabai, G. Measles Virus-Based Treatments Trigger a Pro-inflammatory Cascade and a Distinctive Immunopeptidome in Glioblastoma. *Mol. Ther. - Oncolytics* **12,** 147–161 (2019).

64. Freudenmann, L. K., Marcu, A. & Stevanović, S. Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology* **154,** 331–345 (2018).

65. Ghosh, M., Di Marco, M. & Stevanović, S. Identification of MHC Ligands and Establishing MHC Class I Peptide Motifs. *Methods Mol. Biol.* **1988,** 137–147 (2019).

66. Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A. & Stevanović, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50,** 213–219 (1999).

67. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A. & Peters, B. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47,** D339–D343 (2019).

68. Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S. & Nielsen, M. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61,** 1–13 (2009).

69. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B. & Nielsen, M. NetMHC pan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol.* (2017).

70. Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B. & Nielsen, M. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154,** 394–406 (2018).

71. Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R. T., Racle, J., Coukos, G. & Bassani-Sternberg, M. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. *J. Immunol.* **201,** 3705–3716 (2018).

72. Racle, J., Michaux, J., Rockinger, G. A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., Bassani-Sternberg, M. & Gfeller, D. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.*

37, 1283–1286 (2019).

73. Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M. & Nielsen, M. NNAlign-MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved t-cell epitope predictions. *Mol. Cell. Proteomics* **18,** 2459–2477 (2019).

74. Burnet, M. Cancer-A Biological Approach I. The Processes Of Control. *Br. Med. J.* **1,** 779–786 (1957).

75. Dunn, G. P., Old, L. J. & Schreiber, R. D. The immunobiology of cancer immunosurveillance and immunoediting. *Immunity* **21,** 137–148 (2004).

76. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. *Science* **331,** 1565–1570 (2011).

77. Zitvogel, L., Tesniere, A. & Kroemer, G. Cancer despite immunosurveillance: Immunoselection and immunosubversion. *Nat. Rev. Immunol.* **6,** 715–727 (2006).

78. Hahn, W. C. & Weinberg, R. A. Rules for Making Human Tumor Cells. *N. Engl. J. Med.* **347,** 1593–1603 (2002).

79. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).

80. Valastyan, S. & Weinberg, R. A. Tumor metastasis: Molecular insights and evolving paradigms. *Cell* **147,** 275–292 (2011).

81. Alizadeh, A. M., Shiri, S. & Farsinejad, S. Metastasis review: from bench to bedside. *Tumor Biol.* **35,** 8483–8523 (2014).

82. Cobbold, M., De La Peña, H., Norris, A., Polefrone, J. M., Qian, J., English, A. M., Cummings, K. L., Penny, S., Turner, J. E., Cottine, J., Abelin, J. G., Malaker, S. A., Zarling, A. L., Huang, H. W., Goodyear, O., Freeman, S. D., Shabanowitz, J., Pratt, G., Craddock, C., Williams, M. E., Hunt, D. F. & Engelhard, V. H. MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia. *Sci. Transl. Med.* **5,** 203ra125 (2013).

83. Marino, F., Bern, M., Mommen, G. P. M., Leney, A. C., Van Gaans-Van Den Brink, J. A. M., Bonvin, A. M. J. J., Becker, C., Van Els, C. A. C. M. & Heck, A. J. R. Extended O-GlcNAc on HLA Class-I-Bound Peptides. *J. Am. Chem. Soc.* **137,** 10922–10925 (2015).

84. Brentville, V. A., Metheringham, R. L., Gunn, B., Symonds, P., Daniels, I., Gijon, M., Cook, K., Xue, W. & Durrant, L. G. Citrullinated vimentin presented on MHC-II in tumor cells is a target for CD4+ T-Cell-mediated antitumor immunity. *Cancer Res.* **76,** 548–560 (2016).

85. Rammensee, H.-G. & Singh-Jasuja, H. HLA ligandome tumor antigen discovery for personalized vaccine approach. *Expert Rev. Vaccines* **12,** 1211–7 (2013).

86. Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., Stevanovic, S., Zimmer, R. & Dölken, L. Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* **15,** 363–366 (2018).

87. Pierson, A., Darrigrand, R., Rouillon, M., Boulpicante, M., Renko, Z. D., Garcia, C., Ghosh, M., Laiguillon, M.-C., Lobry, C., Alami, M. & Apcher, S. Splicing inhibition enhances the antitumor immune response through increased tumor antigen presentation and altered MHC-I immunopeptidome. *bioRxiv* (2019). doi:10.1101/512681

88. Menez-Jamet, J., Gallou, C., Rougeot, A. & Kosmatopoulos, K. Optimized tumor cryptic peptides: The basis for universal neoantigen-like tumor vaccines. *Ann. Transl. Med.* **4,** (2016).

89. Chen, D. S. & Mellman, I. Oncology meets immunology: The cancer-immunity cycle. *Immunity* **39,** 1–10 (2013).

90. Beatty, G. L. & Gladney, W. L. Immune escape mechanisms as a guide for cancer immunotherapy. *Clin. Cancer Res.* **21,** 687–692 (2015).

91. Joyce, J. A. & Fearon, D. T. T cell exclusion, immune privilege, and the tumor microenvironment. *Science* **348,** 74–80 (2015).

92. Khong, H. T. & Restifo, N. P. Natural selection of tumor variants in the generation of 'tumor escape' phenotypes. *Nat. Immunol.* **3,** 999–1005 (2002).

93. Vesely, M. D., Kershaw, M. H., Schreiber, R. D. & Smyth, M. J. Natural Innate and Adaptive Immunity to Cancer. *Annu. Rev. Immunol.* **29,** 235–271 (2011).

94. Coley, W. B. The treatment of malignant tumors by repeated inoculations of erysipelas: With a report of ten original cases. *Clin. Orthop. Relat. Res.* 3–11 (1991).

95. Wiemann, B. & Starnes, C. O. Coley's toxins, tumor necrosis factor and cancer research: A historical perspective. *Pharmacol. Ther.* **64,** 529–564 (1994).

96. Zhang, H. & Chen, J. Current status and future directions of cancer immunotherapy. *J. Cancer* **9,** 1773–1781 (2018).

97. Schirrmacher, V. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (Review). *Int. J. Oncol.* **54,** 407–419 (2019).

98. Kennedy, L. B. & Salama, A. K. S. A review of cancer immunotherapy toxicity. *CA. Cancer J. Clin.* **70,** 86–104 (2020).

99. Kroschinsky, F., Stölzel, F., von Bonin, S., Beutel, G., Kochanek, M., Kiehl, M. & Schellongowski, P. New drugs, new toxicities: Severe side effects of modern targeted and immunotherapy of cancer and their management. *Crit. Care* **21,** 1–11 (2017).

100. Ott, P. A., Dotti, G., Yee, C. & Goff, S. L. An Update on Adoptive T-Cell Therapy and Neoantigen Vaccines. *Am. Soc. Clin. Oncol. Educ. book.* **39,** e70–e78 (2019).

101. Ophir, E., Bobisse, S., Coukos, G., Harari, A. & Kandalaft, L. E. Personalized approaches to active immunotherapy in cancer. *Biochim. Biophys. Acta - Rev. Cancer* **1865,** 72–82 (2016).

102. Rausch, S., Gouttefangeas, C., Hennenlotter, J., Laske, K., Walter, K., Feyerabend, S., Chandran, P. A., Kruck, S., Singh-Jasuja, H., Frick, A., Kröger, N., Stevanović, S., Stenzl, A., Rammensee, H.-G. & Bedke, J. Results of a Phase 1/2 Study in Metastatic Renal Cell Carcinoma Patients Treated with a Patient-specific Adjuvant Multi-peptide Vaccine after Resection of Metastases. *Eur. Urol. Focus* (2017).

103. Mendrzyk, R., Ulges, A., Demberg, T., Stephens, G., Reinhardt, C., Walter, S. & Maurer, D. Abstract A015: Cellular immunomonitoring for personalized adoptive cellular therapy trial ACTolog® (IMA101-101). in *Cancer Immunol. Res.* **7,** A015–A015 (2019).

104. Gouttefangeas, C. & Rammensee, H.-G. Personalized cancer vaccines: adjuvants are important, too. *Cancer Immunol. Immunother.* **67,** 1911–1918 (2018).

105. Melief, C. J. M. & Van Der Burg, S. H. Immunotherapy of established (pre)malignant disease by synthetic long peptide vaccines. *Nat. Rev. Cancer* **8,** 351–360 (2008).

106. Marco, M. Di, Peper, J. K. & Rammensee, H. G. Identification of immunogenic epitopes by MS/MS. *Cancer J. (United States)* **23,** 102–107 (2017).

107. Toes, R. E., Blom, R. J., Offringa, R., Kast, W. M. & Melief, C. J. Enhanced tumor outgrowth

after peptide vaccination. Functional deletion of tumor-specific CTL induced by peptide vaccination can lead to the inability to reject tumors. *J. Immunol.* **156,** 3911–8 (1996).

108. Ghosh, M., Gauger, M., Denk, M., Rammensee, H.-G. & Stevanović, S. Antigens in water-in-oil emulsion: a simple antigen extraction method for analysis and proof of equal antigen distribution in vaccination syringe after mixture. *bioRxiv* (2020). doi:10.1101/2020.01.22.916189

109. Singh, M. & O'Hagan, D. Advances in vaccine adjuvants. *Nat. Biotechnol.* **17,** 1075–1081 (1999).

110. Kenter, G. G., Welters, M. J. P., Valentijn, A. R. P. M., Lowik, M. J. G., Berends-van Der Meer, D. M. A., Vloon, A. P. G., Essahsah, F., Fathers, L. M., Offringa, R., Drijfhout, J. W., Wafelman, A. R., Oostendorp, J., Fleuren, G. J., Van Der Burg, S. H. & Melief, C. J. M. Vaccination against HPV-16 oncoproteins for vulvar intraepithelial neoplasia. *N. Engl. J. Med.* **361,** 1838–1847 (2009).

111. Aucouturier, J., Dupuis, L., Deville, S., Ascarateil, S. & Ganne, V. Montanide ISA 720 and 51: A new generation of water in oil emulsions as adjuvants for human vaccines. *Expert Rev. Vaccines* **1,** 111–118 (2002).

112. Black, M., Trent, A., Tirrell, M. & Olive, C. Advances in the design and delivery of peptide subunit vaccines with a focus on Toll-like receptor agonists. *Expert Rev. Vaccines* **9,** 157–173 (2010).

113. Rammensee, H. G., Wiesmüller, K. H., Chandran, P. A., Zelba, H., Rusch, E., Gouttefangeas, C., Kowalewski, D. J., Di Marco, M., Haen, S. P., Walz, J. S., Gloria, Y. C., Bödder, J., Schertel, J. M., Tunger, A., Müller, L., Kießler, M., Wehner, R., Schmitz, M., Jakobi, M., Schneiderhan-Marra, N., Klein, R., Laske, K., Artzner, K., Backert, L., Schuster, H., Schwenck, J., Weber, A. N. R., Pichler, B. J., Kneilling, M., La Fougère, C., Forchhammer, S., Metzler, G., Bauer, J., Weide, B., Schippert, W., Stevanović, S. & Löffler, M. W. A new synthetic toll-like receptor 1/2 ligand is an efficient adjuvant for peptide vaccination in a human volunteer. *J. Immunother. Cancer* **7,** 307 (2019).

114. Didierlaurent, A. M., Laupèze, B., Di Pasquale, A., Hergli, N., Collignon, C. & Garçon, N. Adjuvant system AS01: helping to overcome the challenges of modern vaccines. *Expert Rev. Vaccines* **16,** 55–63 (2017).

115. ICH. Validation of Analytical Procedures: Text and Methodology Q2(R1), International Conference on Harmonisation (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use. (2005). at <http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf>

116. Ghosh, M., Gauger, M., Marcu, A., Nelde, A., Denk, M., Schuster, H., Rammensee, H. G. & Stevanovic, S. Guidance document: validation of a high-performance liquid chromatography-tandem mass spectrometry immunopeptidomics assay for the identification of HLA class I ligands suitable for pharmaceutical therapies. *Mol. Cell. Proteomics* **19,** 432–443 (2020).

117. Fortier, M.-H., Caron, E., Hardy, M.-P., Voisin, G., Lemieux, S., Perreault, C. & Thibault, P. The MHC class I peptide repertoire is molded by the transcriptome. *J. Exp. Med.* **205,** 595–610 (2008).

118. Mommen, G. P. M., Frese, C. K., Meiring, H. D., van Gaans-van den Brink, J., de Jong, A. P. J. M., van Els, C. A. C. M. & Heck, A. J. R. Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci. U. S. A.* **111,** 4507–12 (2014).

119. Caron, E., Vincent, K., Fortier, M.-H., Laverdure, J.-P., Bramoulle, A., Hardy, M.-P., Voisin, G., Roux, P. P., Lemieux, S., Thibault, P. & Perreault, C. The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol. Syst. Biol.* **7,** 533–533 (2014).

120. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14,** 658–73 (2015).

121. Hilf, N., Kuttruff-Coqui, S., Frenzel, K., Bukur, V., Stevanović, S., Gouttefangeas, C., Platten, M., Tabatabai, G., Dutoit, V., van der Burg, S. H., thor Straten, P., Martínez-Ricarte, F., Ponsati, B., Okada, H., Lassen, U., Admon, A., Ottensmeier, C. H., Ulges, A., Kreiter, S., von Deimling, A., Skardelly, M., Migliorini, D., Kroep, J. R., Idorn, M., Rodon, J., Piró, J., Poulsen, H. S., Shraibman, B., McCann, K., Mendrzyk, R., Löwer, M., Stieglbauer, M., Britten, C. M., Capper, D., Welters, M. J. P., Sahuquillo, J., Kiesel, K., Derhovanessian, E., Rusch, E., Bunse, L., Song, C., Heesch, S., Wagner, C., Kemmer-Brück, A., Ludwig, J., Castle, J. C., Schoor, O., Tadmor, A. D., Green, E., Fritsche, J., Meyer, M., Pawlowski, N., Dorner, S., Hoffgaard, F., Rössler, B., Maurer, D., Weinschenk, T., Reinhardt, C., Huber, C., Rammensee, H.-G., Singh-Jasuja, H., Sahin, U., Dietrich, P.-Y. & Wick, W. Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature* **565,** 240–245 (2019).

122. Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E. & Engelhard, V. H. HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science* **255,** 1264–1266 (1992).

123. Lemmel, C., Weik, S., Eberle, U., Dengjel, J., Kratt, T., Becker, H.-D., Rammensee, H.-G. & Stevanović, S. Differential quantitative analysis of MHC ligands by mass spectrometry using stable isotope labeling. *Nat. Biotechnol.* **22,** 450–454 (2004).

124. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., H. Busch, D., Peschel, C., Rad, R., Cox, J., Mann, M. & Krackhardt, A. M. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7,** 13404 (2016).

125. Falk, K., Rötzschke, O., Stevanović, S., Jung, G. & Rammensee, H. G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **351,** 290–6 (1991).

126. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A. & Wu, C. J. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46,** 315–326 (2017).

127. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanović, S., Rammensee, H.-G. & Kohlbacher, O. MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. *J. Proteome Res.* **18,** 3876–3884 (2019).

128. Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., Ojo, N. C., Caldwell, K., Abhyankar, J., Boucher, T., Hart, M. G., Makarov, V., De Montpreville, V. T., Mercier, O., Chan, T. A., Scagliotti, G., Bironzo, P., Novello, S., Karachaliou, N., Rosell, R., Anderson, I., Gabrail, N., Hrom, J., Limvarapuss, C., Choquette, K., Spira, A., Rousseau, R., Voong, C., Rizvi, N. A., Fadel, E., Frattini, M., Jooss, K., Skoberne, M., Francis, J. & Yelensky, R. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* **37,** 55–63 (2019).

129. Brown, S. D. & Holt, R. A. Neoantigen characteristics in the context of the complete

predicted MHC class I self-immunopeptidome. *Oncoimmunology* **8,** 1556080 (2019).

130. Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., Modrusan, Z., Mellman, I., Lill, J. R. & Delamarre, L. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515,** 572–576 (2014).

131. Backert, L. & Kohlbacher, O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med.* **7,** 119 (2015).

132. Keskin, D. B., Anandappa, A. J., Sun, J., Tirosh, I., Mathewson, N. D., Li, S., Oliveira, G., Giobbie-Hurder, A., Felt, K., Gjini, E., Shukla, S. A., Hu, Z., Li, L., Le, P. M., Allesøe, R. L., Richman, A. R., Kowalczyk, M. S., Abdelrahman, S., Geduldig, J. E., Charbonneau, S., Pelton, K., Iorgulescu, J. B., Elagina, L., Zhang, W., Olive, O., McCluskey, C., Olsen, L. R., Stevens, J., Lane, W. J., Salazar, A. M., Daley, H., Wen, P. Y., Chiocca, E. A., Harden, M., Lennon, N. J., Gabriel, S., Getz, G., Lander, E. S., Regev, A., Ritz, J., Neuberg, D., Rodig, S. J., Ligon, K. L., Suvà, M. L., Wucherpfennig, K. W., Hacohen, N., Fritsch, E. F., Livak, K. J., Ott, P. A., Wu, C. J. & Reardon, D. A. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565,** 234–239 (2019).

133. Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., Javeri, I., Nellaiappan, K., Salazar, A. M., Daley, H., Seaman, M., Buchbinder, E. I., Yoon, C. H., Harden, M., Lennon, N., Gabriel, S., Rodig, S. J., Barouch, D. H., Aster, J. C., Getz, G., Wucherpfennig, K., Neuberg, D., Ritz, J., Lander, E. S., Fritsch, E. F., Hacohen, N. & Wu, C. J. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547,** 217–221 (2017).

134. Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., Bukur, V., Tadmor, A. D., Luxemburger, U., Schrörs, B., Omokoko, T., Vormehr, M., Albrecht, C., Paruzynski, A., Kuhn, A. N., Buck, J., Heesch, S., Schreeb, K. H., Müller, F., Ortseifer, I., Vogler, I., Godehardt, E., Attig, S., Rae, R., Breitkreuz, A., Tolliver, C., Suchan, M., Martic, G., Hohberger, A., Sorn, P., Diekmann, J., Ciesla, J., Waksmann, O., Brück, A.-K., Witt, M., Zillgen, M., Rothermel, A., Kasemann, B., Langer, D., Bolte, S., Diken, M., Kreiter, S., Nemecek, R., Gebhardt, C., Grabbe, S., Höller, C., Utikal, J., Huber, C., Loquai, C. & Türeci, Ö. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547,** 222–226 (2017).

135. Walter, S., Weinschenk, T., Stenzl, A., Zdrojowy, R., Pluzanska, A., Szczylik, C., Staehler, M., Brugger, W., Dietrich, P.-Y., Mendrzyk, R., Hilf, N., Schoor, O., Fritsche, J., Mahr, A., Maurer, D., Vass, V., Trautwein, C., Lewandrowski, P., Flohr, C., Pohla, H., Stanczak, J. J., Bronte, V., Mandruzzato, S., Biedermann, T., Pawelec, G., Derhovanessian, E., Yamagishi, H., Miki, T., Hongo, F., Takaha, N., Hirakawa, K., Tanaka, H., Stevanovic, S., Frisch, J., Mayer-Mokler, A., Kirner, A., Rammensee, H.-G., Reinhardt, C. & Singh-Jasuja, H. Multipeptide immune response to cancer vaccine IMA901 after single-dose cyclophosphamide associates with longer patient survival. *Nat. Med.* **18,** 1254–61 (2012).

136. Kowalewski, D. J. & Stevanović, S. Biochemical large-scale identification of MHC class I ligands. *Methods Mol. Biol.* **960,** 145–57 (2013).

137. Nelde, A., Kowalewski, D. J. & Stevanović, S. Purification and Identification of Naturally Presented MHC Class I and II Ligands. *Methods Mol. Biol.* **1988,** 123–136 (2019).

138. Lathrop, J. T., Jeffery, D. A., Shea, Y. R., Scholl, P. F. & Chan, M. M. US Food and Drug Administration Perspectives on Clinical Mass Spectrometry. *Clin. Chem.* **62,** 41–47 (2016).

139. Lynch, K. L. CLSI C62-A: A New Standard for Clinical Mass Spectrometry. *Clin. Chem.* **62,** 24–9 (2016).

140. U.S. Food and Drug Administration Department of Health and Human Services. Guidance for Industry: Bioanalytical Method Validation. (2018). at <https://www.fda.gov/downloads/Drugs/.../Guidances/ucm070107.pdf>

141. European Medicines Agency Committee For Medicinal Products For Human Use. Guideline on Bioanalytical Method Validation. (2018). at <http://www.ema.europa.eu/docs/en%0AGB/document%0Alibrary/Scientific%0Aguideline/%0A2011/08/WC500109686.pdf>

142. Vogeser, M. & Seger, C. Quality management in clinical application of mass spectrometry measurement systems. *Clin. Biochem.* **49,** 947–954 (2016).

143. Schuster, H., Peper, J. K., Bösmüller, H.-C., Röhle, K., Backert, L., Bilich, T., Ney, B., Löffler, M. W., Kowalewski, D. J., Trautwein, N., Rabsteyn, A., Engler, T., Braun, S., Haen, S. P., Walz, J. S., Schmid-Horch, B., Brucker, S. Y., Wallwiener, D., Kohlbacher, O., Fend, F., Rammensee, H.-G., Stevanović, S., Staebler, A. & Wagner, P. The immunopeptidomic landscape of ovarian carcinomas. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E9942–E9951 (2017).

144. Nelde, A., Kowalewski, D. J., Backert, L., Schuster, H., Werner, J.-O., Klein, R., Kohlbacher, O., Kanz, L., Salih, H. R., Rammensee, H.-G., Stevanović, S. & Walz, J. S. HLA ligandome analysis of primary chronic lymphocytic leukemia (CLL) cells under lenalidomide treatment confirms the suitability of lenalidomide for combination with T-cell-based immunotherapy. *Oncoimmunology* **7,** e1316438 (2018).

145. Barnstable, C., Bodmer, W. F., Brown, G., Galfre, G., Milstein, C., Williams, A. F. & Ziegler, A. Production of monoclonal antibodies to group A erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. *Cell* **14,** 9–20 (1978).

146. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–89 (1994).

147. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4,** 923–925 (2007).

148. Organisation for economic co-operation and development. *OECD Principles of Good Laboratory Practice*. (1998). at <http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/mc/chem(98)17&doclanguage=en>

149. Panchaud, A., Scherl, A., Shaffer, S. A., von Haller, P. D., Kulasekara, H. D., Miller, S. I. & Goodlett, D. R. Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean. *Anal. Chem.* **81,** 6481–6488 (2009).

150. Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **9,** 2252–61 (2010).

151. Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G., Toth, M. J. & MacCoss, M. J. Deconvolution of Mixture Spectra from Ion-Trap Data-Independent-Acquisition Tandem Mass Spectrometry. *Anal. Chem.* **82,** 833–841 (2010).

152. Carvalho, P. C., Han, X., Xu, T., Cociorva, D., Carvalho, M. da G., Barbosa, V. C. & Yates, J. R. XDIA: improving on the label-free data-independent analysis. *Bioinformatics* **26,** 847–848 (2010).

153. Panchaud, A., Jung, S., Shaffer, S. A., Aitchison, J. D. & Goodlett, D. R. Faster, Quantitative, and Accurate Precursor Acquisition Independent From Ion Count. *Anal. Chem.* **83,** 2250–2257 (2011).

154. Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H.-G. & Stevanović, S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. *J. Immunol.* **199,** 2639–2651 (2017).

155. Walz, S., Stickel, J. S., Kowalewski, D. J., Schuster, H., Weisel, K., Backert, L., Kahn, S., Nelde, A., Stroh, T., Handel, M., Kohlbacher, O., Kanz, L., Salih, H. R., Rammensee, H.-G. & Stevanović, S. The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for T-cell-based immunotherapy. *Blood* **126,** 1203–13 (2015).

156. Kowalewski, D. J., Schuster, H., Backert, L., Berlin, C., Kahn, S., Kanz, L., Salih, H. R., Rammensee, H.-G., Stevanovic, S. & Stickel, J. S. HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). *Proc. Natl. Acad. Sci. U. S. A.* **112,** E166-75 (2015).

157. Neidert, M. C., Kowalewski, D. J., Silginer, M., Kapolou, K., Backert, L., Freudenmann, L. K., Peper, J. K., Marcu, A., Wang, S. S.-Y., Walz, J. S., Wolpert, F., Rammensee, H.-G., Henschler, R., Lamszus, K., Westphal, M., Roth, P., Regli, L., Stevanović, S., Weller, M. & Eisele, G. The natural HLA ligandome of glioblastoma stem-like cells: antigen discovery for T cell-based immunotherapy. *Acta Neuropathol.* **135,** 923–938 (2018).

158. Berlin, C., Kowalewski, D. J., Schuster, H., Mirza, N., Walz, S., Handel, M., Schmid-Horch, B., Salih, H. R., Kanz, L., Rammensee, H.-G., Stevanović, S. & Stickel, J. S. Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. *Leukemia* **29,** 647–659 (2015).

159. Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., Martinez-Bartolomé, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R. & Hermjakob, H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226 (2014).

160. Rammensee, H.-G., Weinschenk, T., Gouttefangeas, C. & Stevanović, S. Towards patient-specific tumor antigen selection for vaccination. *Immunol. Rev.* **188,** 164–76 (2002).

161. Singh-Jasuja, H., Emmerich, N. P. N. & Rammensee, H. G. The Tübingen approach: Identification, selection, and validation of tumor-associated HLA peptides for cancer therapy. *Cancer Immunol. Immunother.* **53,** 187–195 (2004).

162. Germain, R. N. & Margulies, D. H. The Biochemistry and Cell Biology of Antigen Processing and Presentation. *Annu. Rev. Immunol.* **11,** 403–450 (1993).

163. Germain, R. N. The Biochemistry and Cell Biology of Antigen Presentation by MHC Class I and Class II Molecules: Implications for Development of Combination Vaccines. *Ann. N. Y. Acad. Sci.* **754,** 114–125 (1995).

164. Serwold, T., Gonzalez, F., Kim, J., Jacob, R. & Shastri, N. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature* **419,** 480–3 (2002).

165. Rammensee, H. G., Falk, K. & Rötzschke, O. Peptides Naturally Presented by MHC Class I Molecules. *Annu. Rev. Immunol.* **11,** 213–244 (1993).

166. Bouvier, M. & Wiley, D. C. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science* **265,** 398–402 (1994).

167. Andreatta, M., Alvarez, B. & Nielsen, M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45,** W458–W463 (2017).

168. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo

counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40,** W281-7 (2012).

169. Dönnes, P. & Kohlbacher, O. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res.* **34,** W194-7 (2006).

170. Soam, S. S., Bhasker, B. & Mishra, B. N. Improved prediction of MHC class i binders/non-binders peptides through artificial neural network using variable learning rate: SARS Corona virus, a case study. in *Adv. Exp. Med. Biol.* **696,** 223–229 (2011).

171. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G. & Gfeller, D. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Comput. Biol.* **13,** e1005725 (2017).

172. Trautwein, N. & Stevanović, S. in *Methods Mol. Biol.* **960,** 159–168 (2013).

173. Nelde, A., Kowalewski, D. J. & Stevanović, S. Purification and identification of naturally presented MHC class I and II ligands. *Methods Mol. Biol.* **1988,** 123–136 (2019).

174. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **627,** 71–81 (2008).

175. Schubert, B., Walzer, M., Brachvogel, H.-P., Szolek, A., Mohr, C. & Kohlbacher, O. FRED 2: an immunoinformatics framework for Python. *Bioinformatics* **32,** 2044–2046 (2016).

176. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation. *Mol. Cell. Proteomics* **14,** 658–673 (2015).

177. de Bakker, P. I. W., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., Ke, X., Monsuur, A. J., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E. C., Gao, X., Galver, L., Hart, J., Hafler, D. A., Pericak-Vance, M., Todd, J. A., Daly, M. J., Trowsdale, J., Wijmenga, C., Vyse, T. J., Beck, S., Murray, S. S., Carrington, M., Gregory, S., Deloukas, P. & Rioux, J. D. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38,** 1166–1172 (2006).

178. Svejgaard, A. & Ryder, L. P. HLA and disease associations: Detecting the strongest association. *Tissue Antigens* **43,** 18–27 (1994).

179. Tiwari, J. L. & Terasaki, P. I. *HLA and Disease Associations*. (Springer New York, 1985).

180. Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A. & Peters, B. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43,** D405–D412 (2015).

181. Weiskopf, D., Yauch, L. E., Angelo, M. A., John, D. V., Greenbaum, J. A., Sidney, J., Kolla, R. V., De Silva, A. D., de Silva, A. M., Grey, H., Peters, B., Shresta, S. & Sette, A. Insights into HLA-Restricted T Cell Responses in a Novel Mouse Model of Dengue Virus Infection Point toward New Implications for Vaccine Design. *J. Immunol.* **187,** 4268–4279 (2011).

182. Freudenmann, L. K., Marcu, A. & Stevanović, S. Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology* **154,** 331–345 (2018).

183. Marcu, A., Bichmann, L., Kuchenbecker, L., Backert, L., Kowalewski, D. J., Freudenmann, L. K., Löffler, M. W., Lübke, M., Walz, J. S., Velz, J., Moch, H., Regli, L., Silginer, M., Weller, M., Schlosser, A., Kohlbacher, O., Stevanović, S., Rammensee, H.-G. & Neidert, M. C. The HLA Ligand Atlas. A resource of natural HLA ligands presented on benign tissues. *bioRxiv* (2019). doi:10.1101/778944

184. Nelde, A., Kowalewski, D. J., Backert, L., Schuster, H., Werner, J.-O., Klein, R., Kohlbacher, O.,

Kanz, L., Salih, H. R., Rammensee, H.-G., Stevanović, S. & Walz, J. S. HLA ligandome analysis of primary chronic lymphocytic leukemia (CLL) cells under lenalidomide treatment confirms the suitability of lenalidomide for combination with T-cell-based immunotherapy. *Oncoimmunology* **7,** e1316438 (2018).

185. Kowalewski, D. J., Schemionek, M., Kanz, L., Salih, H. R., Brümmendorf, T. H., Vucinic, V., Niederwieser, D., Rammensee, H.-G., Stevanovic, S. & Stickel, J. S. Mapping the HLA Ligandome Landscape of Chronic Myeloid Leukemia (CML)—Towards Peptide Based Immunotherapy. *Blood* **124,** (2014).

186. Reustle, A., Di Marco, M., Meyerhoff, C., Nelde, A., Walz, J. S., Winter, S., Kandabarau, S., Büttner, F., Haag, M., Backert, L., Kowalewski, D. J., Rausch, S., Hennenlotter, J., Stühler, V., Scharpf, M., Fend, F., Stenzl, A., Rammensee, H.-G., Bedke, J., Stevanović, S., Schwab, M. & Schaeffeler, E. Integrative -omics and HLA-ligandomics analysis to identify novel drug targets for ccRCC immunotherapy. *Genome Med.* **12,** (2020).

187. Löffler, M. W., Kowalewski, D. J., Backert, L., Bernhardt, J., Adam, P., Schuster, H., Dengler, F., Backes, D., Kopp, H. G., Beckert, S., Wagner, S., Königsrainer, I., Kohlbacher, O., Kanz, L., Königsrainer, A., Rammensee, H. G., Stevanovic, S. & Haen, S. P. Mapping the HLA ligandome of colorectal cancer reveals an imprint of malignant cell transformation. *Cancer Res.* **78,** 4627–4641 (2018).

188. Gubin, M. M., Zhang, X., Schuster, H., Caron, E., Ward, J. P., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C. D., Krebber, W.-J., Mulder, G. E., Toebes, M., Vesely, M. D., Lam, S. S. K., Korman, A. J., Allison, J. P., Freeman, G. J., Sharpe, A. H., Pearce, E. L., Schumacher, T. N., Aebersold, R., Rammensee, H.-G., Melief, C. J. M., Mardis, E. R., Gillanders, W. E., Artyomov, M. N. & Schreiber, R. D. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515,** 577–81 (2014).

189. Matsushita, H., Vesely, M. D., Koboldt, D. C., Rickert, C. G., Uppaluri, R., Magrini, V. J., Arthur, C. D., White, J. M., Chen, Y.-S., Shea, L. K., Hundal, J., Wendl, M. C., Demeter, R., Wylie, T., Allison, J. P., Smyth, M. J., Old, L. J., Mardis, E. R. & Schreiber, R. D. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482,** 400–404 (2012).

190. Robbins, P. F., Lu, Y.-C., El-Gamil, M., Li, Y. F., Gross, C., Gartner, J., Lin, J. C., Teer, J. K., Cliften, P., Tycksen, E., Samuels, Y. & Rosenberg, S. A. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19,** 747–752 (2013).

191. Gough, S. C. L. & Simmonds, M. J. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr. Genomics* **8,** 453–65 (2007).

192. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G. & Gfeller, D. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Comput. Biol.* **13,** e1005725 (2017).

193. Del Río, S., López, V., Benítez, J. M. & Herrera, F. On the use of MapReduce for imbalanced big data using Random Forest. *Inf. Sci. (Ny).* **285,** 112–137 (2014).

194. Faviel F Gonzalez-Galarza, Antony McCabe, Eduardo J Melo dos Santos, James Jones, Louise Takeshita, Nestor D Ortega-Rivera, Glenda M Del Cid-Pavon, Kerry Ramsbottom, Gurpreet Ghattaoraya, Ana Alfirevic, Derek Middleton, A. R. J. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48,** D783-788 (2020).

195. Boisvert, F. M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., Barton, G. & Lamond, A. I. A quantitative spatial proteomics analysis of proteome turnover in human

cells. *Mol. Cell. Proteomics* **11,** (2012).

196. Chong, C., Marino, F., Pak, H., Racle, J., Daniel, R. T., Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. High-throughput and sensitive immunopeptidomics platform reveals profound interferon γ-mediated remodeling of the human leukocyte antigen (HLA) ligandome. *Mol. Cell. Proteomics* **17,** 533–548 (2018).

197. Mishto, M. & Liepe, J. Post-Translational Peptide Splicing and T Cell Responses. *Trends Immunol.* **38,** 904–915 (2017).

198. Ghosh, M., Gauger, M., Marcu, A., Nelde, A., Denk, M., Schuster, H., Rammensee, H. & Stevanovic, S. Validation of a high-performance liquid chromatography-tandem mass spectrometry immunopeptidomics assay for the identification of HLA class I ligands suitable for pharmaceutical therapies. *bioRxiv* (2019). doi:10.1101/821249

199. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The Genetic Association Database. *Nat. Genet.* **36,** 431–432 (2004).

200. Gfeller, D. & Bassani-Sternberg, M. Predicting antigen presentation-What could we learn from a million peptides? *Front. Immunol.* **9,** (2018).

201. Del Guercio, M. F., Sidney, J., Hermanson, G., Perez, C., Grey, H. M., Kubo, R. T. & Sette, A. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J. Immunol.* **154,** 685–93 (1995).

202. Sidney, J., del Guercio, M. F., Southwood, S., Engelhard, V. H., Appella, E., Rammensee, H. G., Falk, K., Rötzschke, O., Takiguchi, M. & Kubo, R. T. Several HLA alleles share overlapping peptide specificities. *J. Immunol.* **154,** 247–59 (1995).

203. Stopfer, L. E., Mesfin, J. M., Joughin, B. A., Lauffenburger, D. A. & White, F. Multiplexed relative and absolute quantitative immunopeptidomics reveals MHC I repertoire alterations induced by CDK4/6 inhibition. *bioRxiv* (2020). doi:10.1101/2020.03.03.968750

204. Bilich, T., Nelde, A., Bichmann, L., Roerden, M., Salih, H. R., Kowalewski, D. J., Schuster, H., Tsou, C. C., Marcu, A., Neidert, M. C., Lübke, M., Rieth, J., Schemionek, M., Brümmendorf, T. H., Vucinic, V., Niederwieser, D., Bauer, J., Märklin, M., Peper, J. K., Klein, R., Kohlbacher, O., Kanz, L., Rammensee, H. G., Stevanović, S. & Walz, J. S. The HLA ligandome landscape of chronic myeloid leukemia delineates novel T-cell epitopes for immunotherapy. *Blood* **133,** 550–565 (2019).

205. Zhang, X., Qi, Y., Zhang, Q. & Liu, W. Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomed. Pharmacother.* **120,** 109542 (2019).

206. Hickman, H. D. & Yewdell, J. W. Mining the plasma immunopeptidome for cancer peptides as biomarkers and beyond. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 18747–18748 (2010).

207. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How Machine Learning Will Transform Biomedicine. *Cell* **181,** 92–101 (2020).

208. Rosenberg, S. A., Yang, J. C., Schwartzentruber, D. J., Hwu, P., Marincola, F. M., Topalian, S. L., Restifo, N. P., Dudley, M. E., Schwarz, S. L., Spiess, P. J., Wunderlich, J. R., Parkhurst, M. R., Kawakami, Y., Seipp, C. A., Einhorn, J. H. & White, D. E. Immunologic and therapeutic evaluation of a synthetic peptide vaccine for the treatment of patients with metastatic melanoma. *Nat. Med.* **4,** 321–327 (1998).

209. Cormier, J. N., Salgaller, M. L., Prevette, T., Barracchini, K. C., Rivoltini, L., Restifo, N. P., Rosenberg, S. A. & Marincola, F. M. Enhancement of cellular immunity in melanoma patients immunized with a peptide from MART-1/Melan A. *Cancer J. Sci. Am.* **3,** 37–44 (1997).

210. Matthews, H., Hanison, J. & Nirmalan, N. 'Omics'-informed drug and biomarker discovery: Opportunities, challenges and future perspectives. *Proteomes* **4,** (2016).

211. Xu, J., Thakkar, S., Gong, B. & Tong, W. The FDA's Experience with Emerging Genomics Technologies—Past, Present, and Future. *AAPS J.* **18,** 814–818 (2016).

212. Jo, H. Y., Han, H. W., Jung, I., Ju, J. H., Park, S. J., Moon, S., Geum, D., Kim, H., Park, H. J., Kim, S., Stacey, G. N., Koo, S. K., Park, M. H. & Kim, J. H. Development of genetic quality tests for good manufacturing practice-compliant induced pluripotent stem cells and their derivatives. *Sci. Rep.* **10,** 1–15 (2020).

213. Dimitrov, S., Gouttefangeas, C., Besedovsky, L., Jensen, A. T. R., Chandran, P. A., Rusch, E., Businger, R., Schindler, M., Lange, T., Born, J. & Rammensee, H. G. Activated integrins identify functional antigen-specific CD8+ T cells within minutes after antigen stimulation. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E5536–E5545 (2018).

214. Kirkland, J. L., Tchkonia, T., Zhu, Y., Niedernhofer, L. J. & Robbins, P. D. The Clinical Potential of Senolytic Drugs. *J. Am. Geriatr. Soc.* **65,** 2297–2301 (2017).

215. Vaccines against Coronavirus – the current status of research development | vfa. at <https://www.vfa.de/de/englische-inhalte/vaccines-to-protect-against-covid-19>

216. COVID-19 vaccine - Wikipedia. at <https://en.wikipedia.org/wiki/COVID-19_vaccine>

217. Designing a therapeutic SARS-CoV-2 T-cell-inducing vaccine for high-risk patient groups. (2020). doi:10.21203/RS.3.RS-27316/V1

218. Rajčáni, J. & Szathmary, S. Peptide Vaccines: New Trends for Avoiding the Autoimmune Response. *Open Infect. Dis. J.* **10,** 47–62 (2018).

219. Oscherwitz, J. The promise and challenge of epitope-focused vaccines. *Hum. Vaccines Immunother.* **12,** 2113–2116 (2016).

220. Darwin, C. Charles Darwin's Zoology Notes and Specimen Lists from H. M. S. Beagle. *2000*

9 Publications

Accepted publications

Reguzova, A., **Ghosh, M.**, Müller, M., Rziha, H.-J. & Amann, R. (2020) Orf Virus-based Vaccine Vector D1701-V Induces Strong CD8+ T Cell Response Against the Transgene But Not Against ORFV-derived Epitopes. Vaccines 8, E295 (2020).

Ghosh, M., Gauger, M., Marcu, A., Nelde, A., Denk, M., Schuster, H., Rammensee, H. G. & Stevanovic, S. Guidance document: validation of a high-performance liquid chromatography-tandem mass spectrometry immunopeptidomics assay for the identification of HLA class I ligands suitable for pharmaceutical therapies. Mol. Cell. Proteomics 19, 432–443 (2020).

Bichmann, L., Nelde, A., **Ghosh, M.**, Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanović, S., Rammensee, H.-G. & Kohlbacher, O. MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. J. Proteome Res. 18, 3876–3884 (2019).

Ghosh, M., Di Marco, M. & Stevanović, S. Identification of MHC Ligands and Establishing MHC Class I Peptide Motifs. Methods Mol. Biol. 1988, 137–147 (2019).

Rajaraman, S., Canjuga, D., **Ghosh, M.**, Codrea, M. C., Sieger, R., Wedekink, F., Tatagiba, M., Koch, M., Lauer, U. M., Nahnsen, S., Rammensee, H.-G., Mühlebach, M. D., Stevanovic, S. & Tabatabai, G. Measles Virus-Based Treatments Trigger a Pro-inflammatory Cascade and a Distinctive Immunopeptidome in Glioblastoma. Mol. Ther. - Oncolytics 12, 147–161 (2019).

Di Marco, M., Schuster, H., Backert, L., **Ghosh, M.**, Rammensee, H.-G. & Stevanović, S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. J. Immunol. 199, 2639–2651 (2017).

"Without speculation there is no [...] observation." [220] - Darwin, Charles Robert -