Applied immunoinformatics: HLA peptidome analysis for cancer immunotherapy

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. Linus Backert

aus Lauterbach (Hessen)

Tübingen
2018

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

Tag der mündlichen Qualifikation:	14.11.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Stefan Stevanović

Abstract

Despite different therapeutic approaches, cancer is one of the leading causes of death worldwide. Therefore, new therapies, like immunotherapy, are being developed to cure cancer. All immunotherapies have in common that they need targets to recognize malignant cells. Both the malignant and the benign immunopeptidome have to be examined, to define these new targets. We herein present a large immunopeptidome dataset of benign tissues containing multiple tissue types from different individuals. Moreover, we introduce the HLA Ligand Atlas, a web-interface we developed to accompany the data. It provides user-friendly access to the data, a fast, interactive search option which can be used to search for tissue specific HLA-peptides, and provides common statistics to the user.

Using the large dataset of benign samples, we were able to define general properties of the immunopeptidome. First, we showed that a short time storage of the samples at 8 °C does not alter the immunopeptidome in terms of the number of found peptides and their quality. Next, we performed quality control, in which we found an altered immunopeptidome in the samples of stomach tissue, which might be caused by pepsin in the samples. In addition, we analyzed both the inter- and the intra-individual variability of the immunopeptidome on protein and peptide level. This analysis revealed that sample variability was better explained by HLA type than by tissue-specific peptide presentation. Finally, the large dataset of benign samples allows us to describe properties like the length distribution of different HLA alleles and the nestedness of the peptides in the two HLA classes.

In the last part of this thesis, we show how targets can be defined using immunopeptidome data. In this case, we investigated four different hematological malignancies. We describe entity-dividing lines by using a unsupervised hierarchical clustering of allotype-specific peptides, which showed that entity-specific analysis is recommended. Nevertheless, we found "pan-leukemia"-antigens shared across all four hematological malignancies, which were cancer exclusive.

Zusammenfassung

Trotz verschiedenster therapeutischen Behandlungsmethoden ist Krebs noch immer eine der häufigsten Todesursachen weltweit. Deshalb werden weiterhin neue Therapieansätze, wie zum Beispiel Immunotherapie, entwickelt, um Krebs zu heilen. Zur Entwicklung von Immunotherapien gegen Tumorzellen werden Angriffsziele benötigt, anhand derer Krebszellen erkannt werden können. Zur Bestimmung dieser ist es notwendig sowohl das Immunopeptidom von Krebszellen als auch das von gesundem Gewebe zu kennen. Wir präsentieren einen großen Immunopeptidomdatensatz von gesundem Gewebe, der sowohl verschiedene Organtypen eines Individuums, als auch verschiedene Individuen beinhaltet. Wir haben ein Webinterface - den HLA Ligand Atlas - entwickelt, um einen benutzerfreundlichen Zugriff auf die Daten zu ermöglichen. Dieses Webinterface erlaubt eine schnelle interaktive Suche im Datensatz, wie die Suche nach organspezifischen HLA Peptiden, und stellt zusätzliche Statisken bereit. Des Weiteren erlaubt es die Darstellung der Massenspektrometriespektren in einem interaktivem Spektrumviewer.

Mit Hilfe des großen Datensatz an Normalgewebe konnten wir allgemeine Eigenschaften des Immunpeptidom bestimmen. Zuerst zeigen wir, dass das Immunopeptidom sich sowohl quantitativ als auch qualitativ nicht ändert, wenn die Probe kurzzeitig bei 8 °C gelagert wird. Als nächsten führten wir eine Qualitätskontrolle durch, die ein verändertes Immunopeptidom bei den Proben des Magengewebes aufzeigte, welches möglicherweise durch Pepsin in den Proben verursacht wurde. Zusätzlich untersuchten wir die inter- und intra-individuelle Variabilität des Immunopeptidom auf Protein- und Peptideebene. Die Analyse zeigte hier, dass der HLA-Typ einen größeren Einfluss auf die Variablität hat als die organspezifische Präsentation. Der große Datensatz von Normalgewebe erlaubte uns auch die Beschreibung weiterer Eigenschaften, wie die Peptidlängenverteilung für verschieden HLA Allele und die Beschreibung von Längenvarianten in den zwei HLA Klassen.

Im letzten Teil dieser Doktorarbeit zeigen wir wie neue Angriffsziele mit Hilfe von Immunopeptidomdaten gefunden werden können. In unserem Fall untersuchten wir vier verschieden hämatologische Krebsarten. Durch eine unüberwachte hierarchische Clusteranalyse auf allotypspezifischen Peptide wurden hier klare, entitätsspezifische Cluster identifiziert. Dieser Befund spricht für die Notwendigkeit einer entitätsspezifischen Anaylse solcher Datenätze. Nichtsdestotrotz konnten wir auf allen vier hämatologischen Krebsarten „Pan-leukemia“ Antigene finden, die krebsexklusiv sind.

General Remarks

- In accordance with the standard scientific protocol, I will use the personal pronoun ***we*** to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Chapter 1

Introduction

Cancer is one of the leading causes of death worldwide and despite many available therapies, 8.2 million cancer-related deaths have been reported in 2012[43]. The treatment of cancer with classical therapies yields only limited success and new therapies are getting developed with the focus shifting to immunotherapies[99]. The main concept behind immunotherapies is the ability of the immune system to recognize and eliminate cancer cells. The recognition is based on peptides bound to the human leukocyte antigen (HLA) (antigens) and the set of all peptides bound to HLA is defined as the immunopeptidome[3]. There are two kinds of antigens which are important for cancer immunotherapy. The first are tumor-specific antigens (TSAs), which have never been found on benign tissue and can, for example, contain cancer-specific mutations. The second ones are tumor-associated antigens (TAAs). These have also been found on benign tissue but are in general cancer-related, for example strongly over expressed[52]. To define TSAs and TAAs both the cancer and the benign immunopeptidome have to be examined. The former has been analyzed extensively by immunoprecipitation followed by tandem mass spectrometry[19,36,66,70,135]. Nevertheless, to find TSAs and TAAs analyzing tumor samples is not enough and the immunopeptidome from benign tissues is needed as background or negative dataset to avoid side effects like cross-reactivity. Despite its significance, no large benign antigen dataset is publicly available. This thesis tries to fill this gap of knowledge, by presenting a large benign dataset of antigens collected from 85 samples.

Analyzing such a large number of samples can be difficult, considering that each sample can result in more than 5,000 peptides. Therefore, the field of immunoinformatics developed tools like binding prediction (e.g., netMHCpan[56]) and HLA typing (e.g., OptiType[122]), to allow for such datasets to be analyzed. Furthermore, databases containing immune system related data, like lists of HLA peptides and T-cell epitopes (e.g., IMGT[76], IEDB[130], SystemMHC Atlas[114]), have been published, which are are of particular importance for the comparison and development of possible new targets.

1.1 The HLA Ligand Atlas: Providing public access to a benign immunopeptidome

The development of new cancer immunotherapies and the discovery of new possible TAAs and TSAs needs numerous immunopeptidome samples. The resulting large amount of data has to be processed and analyzed. Considering that one single HLA immunopeptidome analysis from one sample can result in a list of more than 5,000 peptides, with many features per peptide[15], databases are needed to store and access these large amounts of data. In the context of immunology, the three largest and/or oldest databases are IMGT (the international ImMunoGeneTics information system)[76] containing information about HLA sequences, antibodies, and T-cell receptors, IEDB[130] (epitopes and epitope-MHC/BCR complexes), and SYFPEITHI[97] (MHC ligands and T-cell epitopes). The latter two contain experimentally derived HLA ligands and in addition, IEDB also non-binding peptides. However, both are mostly focused on immunologically important peptides such as tumor antigens and do not provide a complete benign dataset. All three databases allow researchers to access the data via a user-friendly web interface. For a simple search in the data (e.g., a peptide sequence), this way of accessing data is preferable since it allows the researcher to gather information without further knowledge of how the data is stored or internally accessed.

In DNA and RNA sequencing, generating large datasets and providing public access to them has become standard (e.g., TCGA and Gene Expression Omnibus[37]). Nowadays large datasets can also be obtained and are published using mass spectrometry (MS) driven methods, like proteomics. In 2014, Wilhelm et al. published an MS-based draft of the human proteome[143] and in 2015 Uhlen et al. published the human protein atlas, a tissue-based map of the human proteome[126]. Both publications provide a quick and user-friendly way to access large datasets of MS data via a web interface. Furthermore, Shao et al. recently published the SysteMHC Atlas[114], which gathers immunopeptidome datasets published in the PRoteomics IDEntifications (PRIDE) database and reprocesses the raw files with a standardized pipeline. The SysteMHC Atlas allows searching for peptides and proteins contained in the samples and presents a spectrum library.

Inspired by these three publications, we developed a database for immunopeptidome data. This database contains a large dataset of 85 benign samples from six different individuals. These samples have been obtained along with standard autopsies and from each individual multiple different tissue types have been collected. The data was stored in a MySQL database. In addition, we developed a user-friendly web interface using the python based web framework Pyramid, as well as state-of-the-art web-design tools, like Bootstrap, DataTables, and jQuery. This web interface does not only allow searching for peptides, proteins, and HLA alleles but also contains statistics like tissue-specific peptides and an MS spectrum viewer. The resulting web page is called "HLA Ligand Atlas" and is publicly available to stimulate further HLA immunopeptidome research. We hope that the HLA Ligand Atlas enables researchers to find new antigens and develop new cancer immunotherapies.

1.2 Analysis of the benign tissue immunopeptidome

Despite the possibility of new scientific insights, neither a detailed description nor an extensive analysis of the benign immunopeptidome has been done so far. In general, the focus of previous publications was either on malignant samples[15,62,67,135] or cell lines[71]. Because of that, we performed a comprehensive analysis of the benign immunopeptidome data included in the HLA Ligand Atlas. In contrast to most other studies, our samples were obtained from autopsies, which allowed us to take more than one type of tissue sample, but had the drawback of a possible change of the immunopeptidome during the time between death and autopsy. Therefore, we performed a time-series experiment, which indicated no change of the immunopeptidome over a time of 72h. Furthermore, we performed multiple quality control steps, to ensure that only high-quality data is contained in the HLA Ligand Atlas. This quality control step revealed disturbed HLA peptide motifs for our two stomach samples, indicating the influence of pepsin in our samples of stomach tissue.

The raw data of the mass spectrometry experiments can be analyzed using different identification algorithms. The most common automated database search engines are Mascot[94], SEQUEST[40], X!Tandem[31], Andromeda[30] or Comet[38,39]. We performed a small benchmark of the two algorithms Sequest HT and Comet, to evaluate which one is better suited to analyzing immunopeptidome data. In addition, we tested the influence of Percolator[58] on the number of identified peptides and their quality. This benchmark shows a superiority of Comet over Sequest HT, especially in combination with Percolator.

Current studies of the tumor immunopeptidome contain at most two different tissue samples of one individual (e.g., malignant and adjacent benign) and therefore the tissue-variability of the immunopeptidome inside an individual is not described. Furthermore, although former studies analyzed tumors from different individuals, an inter-individual analysis of the variation in the benign immunopeptidome is not published. To fill this research gap, we analyzed both the inter- and intra-individual variability of the immunopeptidome on protein and peptide level. This analysis revealed that sample variability was better explained by HLA type than by tissue-specific peptide presentation.

In previous studies, mono-allelic cell lines have been used to determine the immunopeptidome of specific HLA alleles. At the same time, properties of the peptides presented by single HLA allele have been described[1]. The large dataset of benign samples allows us to describe properties like the length distribution of different HLA alleles, which was similar to the ones described by Abeline et al.[1] For the development of possible therapeutic vaccines, HLA class I peptides that are contained as a substring in HLA class II peptides are of special interest[64,113].

We therefore also analyzed the general frequency of this phenomenon and calculated the number of peptide length variations within an HLA class. In addition, we computed the protein overlap between HLA class I and II and the number of peptide length variations inside an HLA class. The latter analyses showed that, on average, 8% of HLA class I and 59% of the HLA class II peptides do have variations in length.

1.3 A meta-analysis of the HLA peptidome composition in different hematological malignancies

Antigen-specific immune-checkpoint blockade inhibition has led to major breakthroughs in the treatment of solid malignancies[22,29,48,81,86,100]. However, the effect of this treatment have been found to correlate with the mutational load[102,117]. This might explain this treatment's limited effectiveness for treating hematological malignancies (HM) (excluding Hodgkin lymphoma)[9,10], which are characterized by a low mutational load. HM can be treated by stem cell transplantation[21,101,141], donor lymphocyte infusion[107,108,111] or chimeric antigen receptor (CAR) T cells[44,80,95], but all these methods often show off-target toxicity such as graft-versus-host disease. Hence, to develop new treatment strategies against hematological cancer new targets have to be identified.

To describe the antigen landscape of hematological cancer, we performed a meta-analysis of four different hematological malignancies, which have been described individually by our group beforehand[19,66,118,135]. The dataset consists of samples from acute myeloid leukemia (AML)[19], chronic myeloid leukemia (CML)[118], chronic lymphocytic leukemia (CLL)[66], and multiple myeloma (MM)[135].

In this meta-analysis we first conducted an unsupervised hierarchical clustering of the source proteins of the immunopeptidome, to gain an overview of the antigen landscape of the four different HM. However, instead of distinct clustering for the different HM we found a clustering along non-entity specific common antigens, which was caused by the different HLA typings of the samples. Therefore, we repeated the analysis, but this time with an HLA allotype specific immunopeptidome. This dataset was created by assigning the peptides to their HLA allele using netMHCpan-3.0[89]. In this analysis the clustering of the HLA-A*02:01 peptides showed entity-specific dividing clusters. Next, we subtracted a large in-house immunopeptidome dataset of benign samples and did an allotype-specific overlap analysis, which identified a small panel of naturally presented 'pan-leukemia' antigens. However, these new targets have not yet been evaluated for immunogenicity or tumor-specific cytotoxicity.

Chapter 2

Biological background

This chapter introduces the biological background of this thesis and summarizes the immunological basis of the presented research. First, an overview of the mechanisms of the immune system is given (Section 2.1), followed by a detailed review of the HLA ligand processing pathway (Section 2.2). These sections are based on Janeway's Immunobiology[87]. Furthermore, this chapter contains a short summary of hematological cancer types and possible treatments.

2.1 The immune system

The immune system helps the organism to fight pathogens like viruses, bacteria, and helminths. Another function of the immune system is the surveillance of endogenous cells and the controlled apoptosis of mutated and dysplastic cells. It can be divided into two parts: the innate immune system and the adaptive immune system.

2.1.1 The innate immune system

If the pathogen crosses physical barriers like the skin, the innate immune system protects the organism in a generic way. First, it uses antimicrobial enzymes and peptides, and plasma proteins known as the complement system. Next, the innate immune cells recognize pathogen-associated molecular patterns (PAMPs) and kill pathogens by phagocytosis. If the pathogen withstands the mechanisms of the innate immune system, the adaptive immune system is needed. It targets the pathogen specifically using antigen-specific lymphocytes and provides long-lasting specific immunity. This thesis focuses on HLA immunopeptidome, which is part of the adaptive immune system. Therefore, we will only describe the adaptive immune system in detail.

2.1.2 The adaptive immune system

The adaptive immune system includes a humoral- and cellular-mediated immune response, which will be both described in the next two sections.

The humoral-immune response

The humoral-immune response uses B cells, which recognize special B-cell antigens using the B-cell receptors (BCRs). If an antigen binds specifically to a BCR, the B cell secretes antibodies against the antigen. These antibodies can bind to the antigen presented on the cell and marks the pathogen for ingestion or elimination by phagocytes. This mechanism is called antibody opsonization. In the next step, the antibody-dependent cell-mediated cytotoxicity (ADCC) eliminates the pathogen. This mechanism uses effector cells, like natural killer cells, macrophages, neutrophils, and eosinophils, which recognize the antibody bound to the cell or pathogen. Finally, these cells release enzymes, like granenzymes, which lead to the neutralization of the cell or pathogen.

The cellular-immune response

The cellular-mediated immune response is based on T cells and therefore also known as T-cell mediated immune response. This thesis focuses especially on the cellular mediated immune response and the T-cell HLA interactions. Because of that we will only describe the cellular-mediated immune response in detail.

The cellular-mediated immune response depends on the presentation of antigens by HLA, which can be recognized by T cells. In contrast to the innate immune system, most of the effector T cells act not on the pathogen itself but on other host cells. The development of T cells starts with immature T cells, which arise in the bone marrow. All T cells undergo first a positive selection in the thymus. In the positive selection the cells are removed if their HLA does not bind properly, caused for example by misfolding of the HLA complex. In the negative selection, T cells are removed which present peptides inherent to the organisms healthy tissue. These peptides are called self-antigens and T cells presenting them are discarded to avoid autoimmunity. After maturation in the thymus naive T cells travel through the lymph and blood until they encounter their antigen which triggers their activation and further differentiation into memory T cells.

T cells recognize their antigen using the T-cell receptor (TCR). The antigen is a peptide presented by the human leukocyte antigen. There are two different types of HLA molecules: class I and class II. Class I presents intracellular antigens, which also cover antigens from pathogens such as viruses. Class II presents extracellular antigens, which can originate from extracellular pathogens like bacteria after phagocytosis. The class I HLA peptide complex activates CD8$^+$ T cells. CD8 is a transmembrane protein on the T cell and binds as co-receptor to HLA class I. In contrast, the class II HLA peptide complex activates T cells with the co-receptor CD4. Hence, these are called CD4$^+$ T cells (Figure 2.1).

After the encounter with an HLA class I complex, the CD8$^+$ T cells can differentiate into cytoxic killer T cells (CTL), which kill cells presenting the same antigen using cytotoxic effector molecules like perforin, granzymes, granulysin and Fas ligands. This mechanism is for example used to kill virus-infected cells that present antigens originating from viral proteins. CD4$^+$ T cells can differentiate into different types of T-helper cells after encountering their HLA class II antigen. There are two types of T-helper cells T_H1 and T_H2. T_H1 cells activate macrophages that present antigens recognized by their TCR. These macrophages then destroy intracellular microorganisms. Furthermore, T_H1 cells can stimulate B cells to produce IgG antibodies against the pathogen. Another property of T_H1 cells is the ability to release cytokines like interferon gamma, which activate CTLs and stimulate their differentiation. T_H2 cells stimulate B cells to differentiate and to produce non-IgG antibodies.

Before the T cells can recognize the antigens presented by HLA, these antigens have to be processed and presented on a cell, which is described in the following section.

2.2 The HLA ligand processing pathways

2.2.1 HLA class I

The HLA class I processing pathway begins with the digestion of intracellular proteins by the proteasome (Figure 2.1. The proteasome is a large intracellular protease consisting of a 20S catalytic core and two 19S regulatory caps. All three parts consist of multiple subunits. Proteins are tagged for digestion with markers such as ubiqutin, which are recognized by the 19S caps. The tagged proteins are digested into peptides, which are released into the cytosol. From the cytosol, the peptides are transferred by the transporter associated with antigen processing (TAP) into the endoplasmic reticulum (ER). In the ER, the peptides then bind to the HLA class I molecule and form the peptide:HLA complex. This binding is associated with the peptide-loading complex, which consists of four main components: calreticulin, tapasin, ERp57, and TAP itself. In the last step, the peptide:HLA complex is transported to the cell surface, where it presents the endogenous antigens to CD8$^+$ T cells. HLA class I presents mostly peptides of length 8 to 12 amino acids.

2.2.2 HLA class II

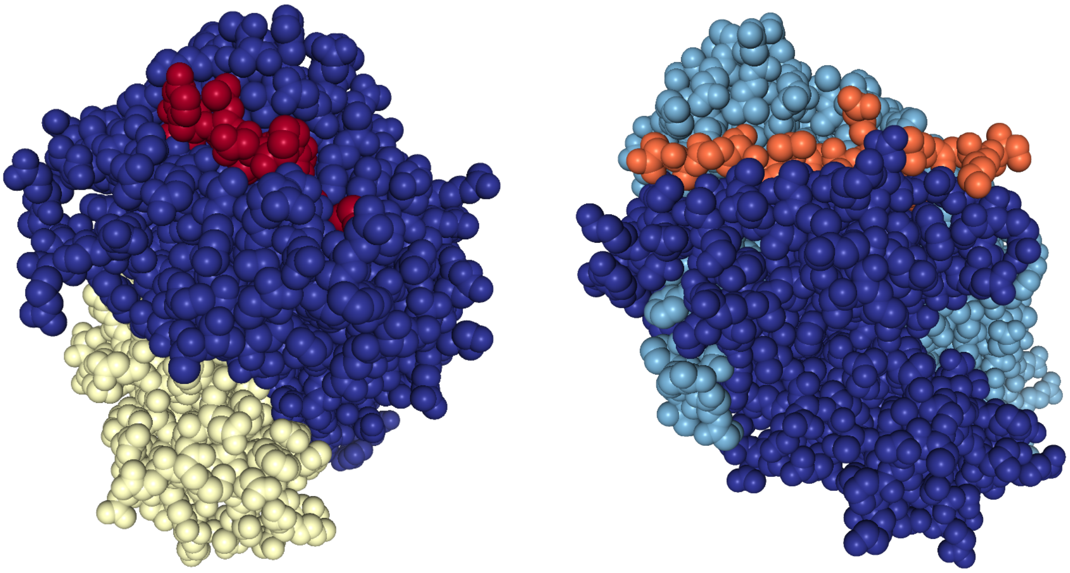
The HLA class II processing pathway begins with the uptake of extracellular proteins into intracellular vesicles. These endosomal vesicles contain proteases, which are activated as soon as the pH value decreases. The proteases cut the proteins into peptides and the vesicle containing the peptides then fuses with vesicles enclosing HLA class II molecules. In the joint vesicle, the peptides bind to HLA and form the class II peptide:HLA complex. In the next step, the vesicle is transported to the cell surface, where the HLA class II complex is presented and can be recognized by CD4$^+$ T cells. In contrast to HLA class I, class II presents mostly peptides of length 12 to 25 amino acids.

2.2.3 HLA binding motifs

The human leukocyte antigen is polygenic. This implies that each individual has several different HLA class I and class II genes. Additionally, HLA genes are the most polymorphic genes known in the human genome. This means that there are multiple alleles of each gene in the population. In detail, every person has at least three different HLA class I (considering HLA-A, -B, -C) and three HLA class II genes (considering HLA-DP, -DQ, -DR). Furthermore, more than 16,000 different HLA alleles are known and due to the price-reduction in sequencing technologies more and more alleles are discovered every year (Figure 2.2). The polygenesis and polymorphisms are important for the adapted immune system, since the presentation of a protein depends on the HLA molecule and the sequences of the digested peptides. This means that a peptide can only bind to an HLA molecule if it fits into its binding cleft. This fit can be described by peptide binding motifs, which have to be defined for each HLA allele individually. This binding motif defines which amino acids are allowed or preferred at which position in a peptide. Furthermore, each allele has a small set of anchor positions, which are more important and less variable than the others and contribute more to the binding (Figure 2.3).

The binding restrictions of HLA class I and class II differs heavily, which is caused by the difference in the 3D structure of the HLA proteins. Whereas the binding cleft of HLA class I can be described as a bathtub form, the HLA class II binding cleft looks more like a hot dog (Figure 2.4). These two different forms lead to both a variety in the length of the binding peptides (class I: 8-12, class II: 8-25) and the general binding motif. The strict length of the HLA class I peptides is caused by the fitting of the peptide into the binding cleft, which does not allow long dangling ends. In contrast, the form of the HLA class II binding cleft allows long dangling ends and the peptide can shift inside it. Because of that the binding motif of class II peptides is not described by a matrix of a length 8-25, but by a binding core, which is again considered to be nine amino acids long.

Due to the large variety of the binding motifs, the prediction of the HLA binding affinity of a peptide is not a simple task and a variety of HLA binding prediction software tools has been developed. Chapter 3.3 provides an comprehensive overview of HLA binding prediction software.



(a) HLA class I (5N1Y)

(b) HLA class II (5LAX)

Figure 2.4: 3D structures of HLA class I A*02 (blue and white) with the peptide MVWG-PDPLYV (red) and HLA class II DRB1*04:01 (dark and light blue) with alpha-enolase peptide 26-40 (orange). Structure obtained from the Protein Data Bank (PDB)[20] (5N1Y[27], 5LAX[47]).

2.3 HLA peptide extraction

The analysis of the immunopeptidome requires the sample preparation and the purification of HLA peptides (Figure 2.5). We here describe in short the main steps necessary to extract HLA peptides. First, the cells of the sample are lyzed, resulting in a tissue lysate. Second, the lysate is purified and only solubilized proteins remain. Third, the immunoprecipitation is used to extract peptides bound to HLA. It is based on affinity chromatography, in which antibodies specific for HLA are bound on a column. The immunoprecipitation, used for this thesis, utilizes three different antibodies that are either specific for HLA class I (e.g., W6/32) or class II (e.g., Tue39/L243). The lysate runs through the columns coated with the HLA specific antibody and only HLA molecules or peptide-HLA complexes are bound to it. Once the HLA molecules/complexes are bound to the column, they are eluted with acid, resulting in a mixture of HLA molecules and HLA peptides. Last, the peptides are separated using ultrafiltration resulting in the isolated peptides, which can then be analyzed in the MS. A more detailed description of the immunoprecipitation is given by Kowalewski et al.[69].

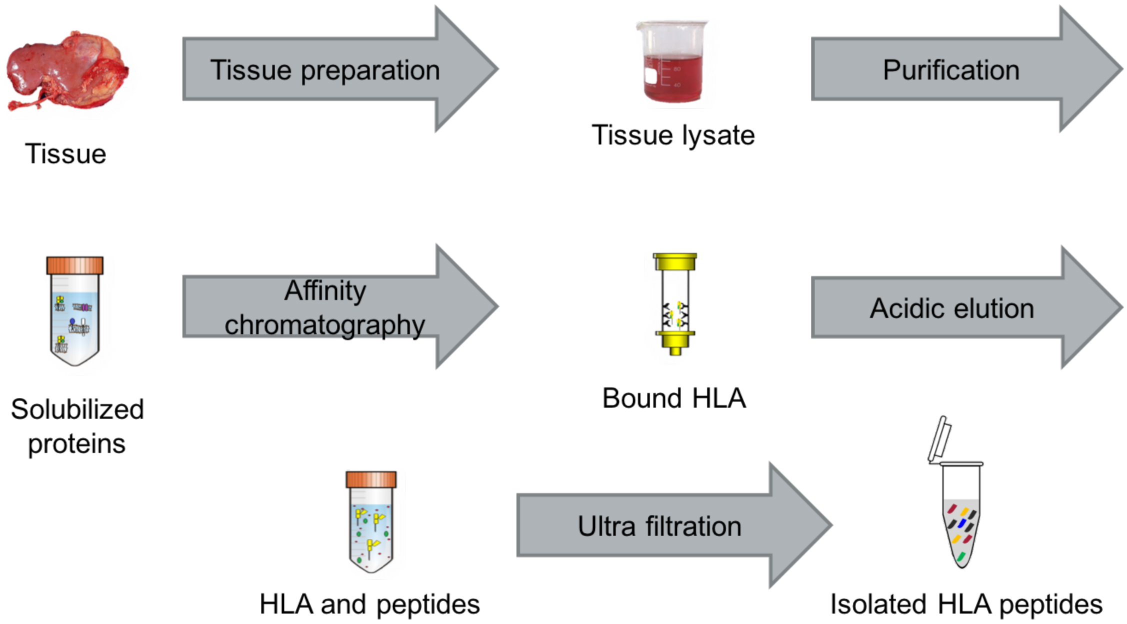


Figure 2.5: Simplified workflow of the isolation of HLA ligands. First, the cells of the sample are lyzed, resulting in a tissue lysate. Second, the lysate is purified and only solubilized proteins remain. Third, the immunoprecipitation is used to extract peptides bound to HLA. It is based on affinity chromatography, in which antibodies specific for HLA are bound on a column and the lysate runs through the columns and peptide-HLA complexes are bound to it. Next, they are eluted with acid, resulting in a mixture of HLA molecules and HLA peptides. Last, the peptides are separated using ultrafiltration resulting in the isolated peptides

2.4 Hematologic malignancies

In Chapter 6, we use data from four different types of hematologic malignancies. To outline the differences of these malignancies, we will shortly describe each of them. Furthermore, a short discussion of possible therapies is presented for each malignancy.

2.4.1 Acute myeloid leukemia

Acute myeloid leukemia (AML) is the result of multiple genetic alterations in hematopoietic precursor cells[132]. These genetic alterations result in abnormal growth and differentiation in the cells, causing large amounts of pathological and immature cells in the bone marrow and peripheral blood. The main problem is that these cells, despite the ability to divide and proliferate, cannot differentiate into mature hematopoietic cells and therefore accumulate in the patient. The overall-survival and prognosis of patients with AML depends significantly on the age, starting with a 60% five-year survival rate for patients younger than 30 to a decreasing rate of 23% for patients between 55 and 64[55]. The age and the overall condition is also a limiting factor for the therapy choices. The primary treatment modality is a combination chemotherapy (cytarbine and anthracyline) and stem cell transplant.

2.4.2 Chronic myeloid leukemia

Chronic myeloid leukemia (CML) is in most cases caused by the Philadelphia chromosome, which is a mutated Chromosome 22. The Philadelphia chromosome leads to a unique gene product (BCR-ABL1), a permanently active tyrosine kinase[41]. Because of its central role in CML, BCR-ABL1 is often used as treatment target[129]. CML develops in three disease phases: the chronic phase, the accelerated phase, and the blast crisis[28]. Furthermore, CML can develop into an acute leukemia. The treatment of CML depends on the disease phase, availability of a donor for hematopoietic cell transplantation, patient age, and the response to treatment with tyrosine kinase inhibitors. Based on the mentioned factors, there are three main treatment options for CML: hematopoietic stem cell transplant, tyrosine kinase inhibitors, and palliative therapy with cytotoxic agents[11].

2.4.3 Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) can be characterized by a progressive accumulation of functionally incompetent mature B cells and is the most common leukemic disorder in the Western hemisphere[23]. This accumulation is induced by the inability of B cells to undergo apoptosis[84]. CLL is frequently associated with the following genetic lesions: deletion of 17p or TP53 mutation, a deletion of 11q, trisomy 12, or a deletion of 13q[34]. Due to its heterogeneity, the treatment of CLL varies from "watchful waiting" to chemotherapy[26]. However, CLL cannot be cured by current treatment options. Therefore, only a treatment of the symptoms of the disease is attempted.

2.4.4 Multiple myeloma

In multiple myeloma (MM), plasma cells undergo neoplastic proliferation. These proliferating plasma cells produce monoclonal immunoglobulin and can be found in the bone marrow, which often causes extensive skeletal destruction[74]. MM is frequently caused by cytogenetic changes in the immunoglobulin heavy-chain locus on chromosome 14q32 and one of five other chromosomes, 11q13, 4p16.3, 6p21, 16q23, and 20q11[73,112]. The treatmant of MM often involves autologous stem-cell transplantation. If the patient is not eligible for transplantation, for example because of his age, lenalidomide in combination with dexamethasone is used for treatment[109].

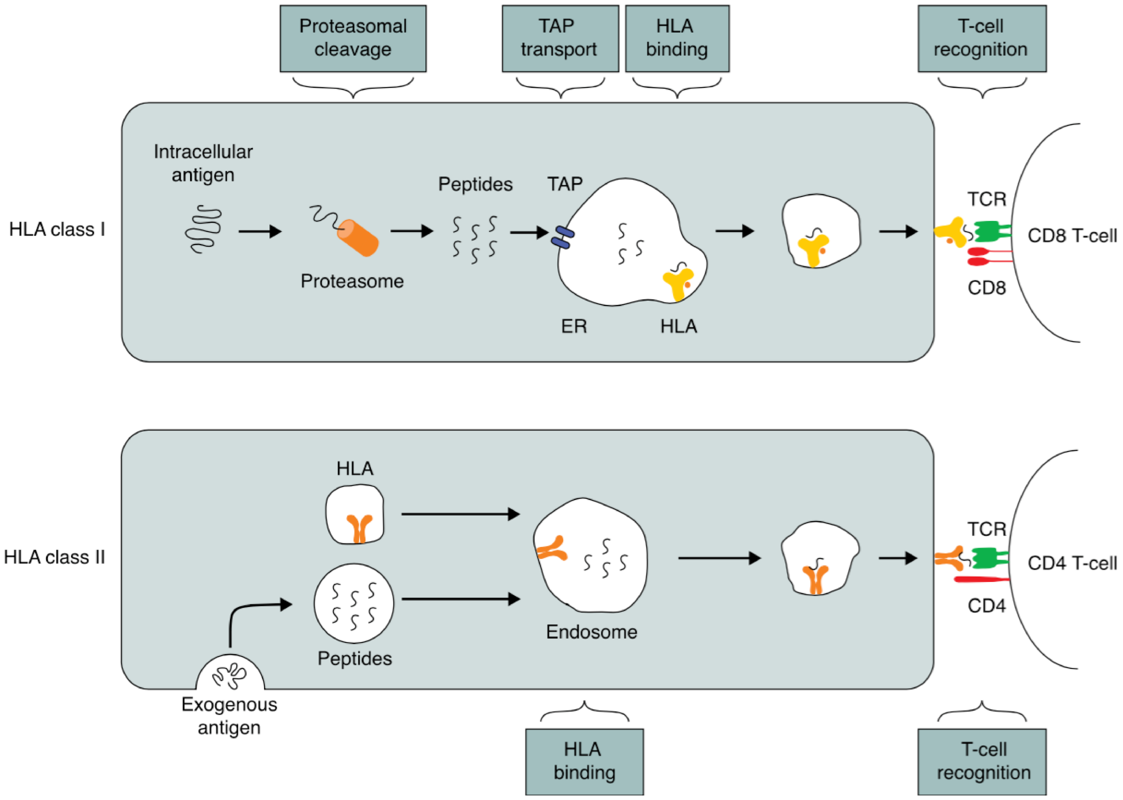


Figure 2.1: HLA ligand processing pathway for class I and class II proteins. The HLA class I pathway presents intracellular antigens, which are digested in the proteasome into small peptides. These peptides are transported into the ER by TAP and there loaded onto HLA class I molecules. Finally, the peptide:HLA complex is transported via vesicle to the cell surface, where they can interact with CD8$^+$ T cells. The HLA class II pathway presents exogenous antigens, which are digested in the endosome into small peptides. These are then loaded onto HLA and the peptide:HLA complex is again transported to the cell surface. The HLA class II complex can be recognized by CD4$^+$ T cells. Boxes on top and bottom highlight processes for which machine-learning based prediction tools exist (for HLA binding prediction see Section 3.3).Reproduced with permission from Backert & Kohlbacher, 2015 [13].

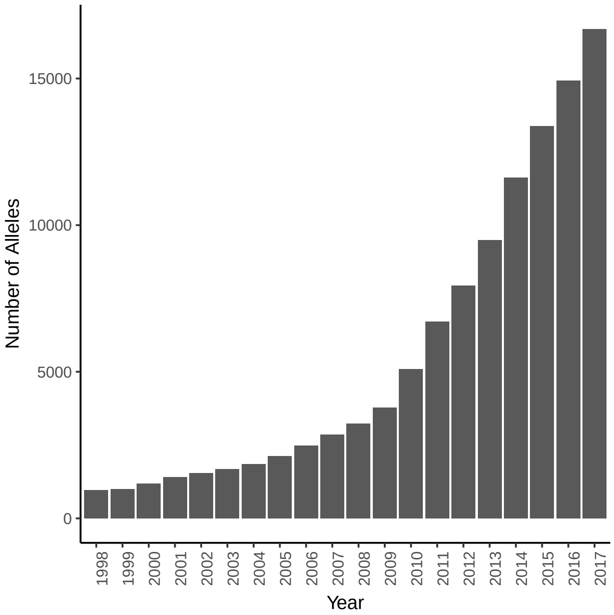


Figure 2.2: Amount of HLA class I and class II alleles contained in the international ImMunoGeneTics information system(IMGT) [104]. Statistics obtained from `http://www.ebi.ac.uk/ipd/imgt/hla/stats.html` (06.19.2017).

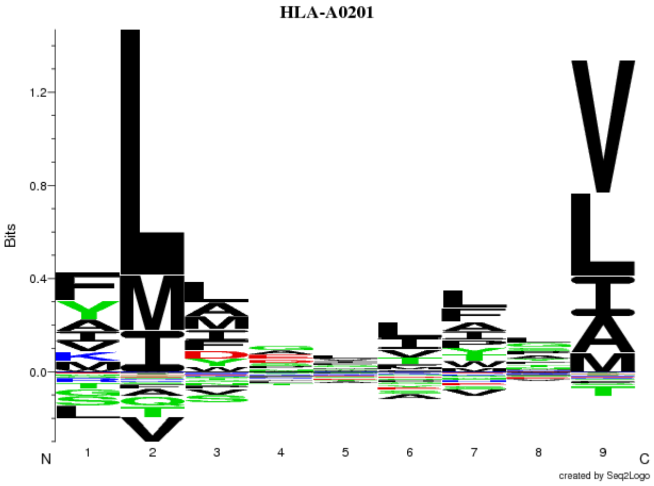


Figure 2.3: HLA binding motif of A*02:01 for nine amino acid long peptides. The size of the letters reflects how often the corresponding amino acid is in the position. Anchor positions are in this case amino acid two and nine. The colors of the letters correspond to different groups of amino acids with certain properties, like polar/ unpolar and hydrophilic/ hydrophobic. Image from the netMHC-4.0 webpage `http://www.cbs.dtu.dk/services/NetMHCpan/logos.php`(06.19.2017) [7].

Chapter 3

Computational background

All data in this thesis was measured from tissue samples using high-performance liquid chromatography (HPLC) coupled to tandem mass spectrometry (LC-MS/MS). This chapter presents the computational methods used to process and analyze the data acquired from the MS.

Measuring samples in LC-MS/MS results in two different types of spectra, the precursor spectra (MS1) and the MS/MS fragment spectra (MS2) (Figure 3.1). Before identification, these spectra have to be peak picked and in the next step, both the precursor spectra and the corresponding fragment spectra allow to identify the peptides and their source proteins. The latter process is called peptide identification. After the identification, the peptides are assigned to an HLA using HLA binding prediction methods, which are based on machine learning (ML). The methods used in all these steps are described in this chapter.

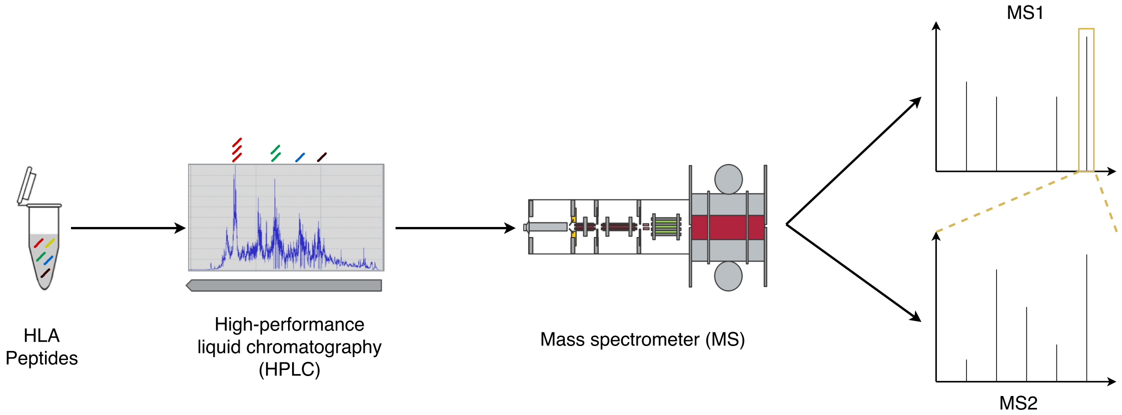


Figure 3.1: High-performance liquid chromatography coupled tandem mass spectrometry workflow. HLA peptides are separated by the HPLC and then measured in the mass spectrometer. The MS reports two kinds of the spectra: the precursor spectra (MS1) measured in the first MS and the fragment spectra (MS2) acquired in the second MS.

3.1 Peptide identification

Peptide identification assigns the MS2 fragment spectra measured in the MS to one or multiple possible peptides and reports a score or probability for each spectrum-peptide assignment. The peptide itself can originate from one or multiple source proteins. This allows to assign also one or multiple possible protein to an spectrum. At the very beginning of the development of mass spectrometry, spectra were assigned manually to peptides by experts. Since new mass spectrometers yield thousands of spectra in each run, a manual assignment is no longer feasible. Nowadays, the identification is done by automated database search engines such as Mascot[94], SEQUEST[40], X!Tandem[31], Andromeda[30] or Comet[38,39]. In this thesis, the three peptide identification algorithms Mascot, Sequest, and Comet were used. Therefore, they will be described below.

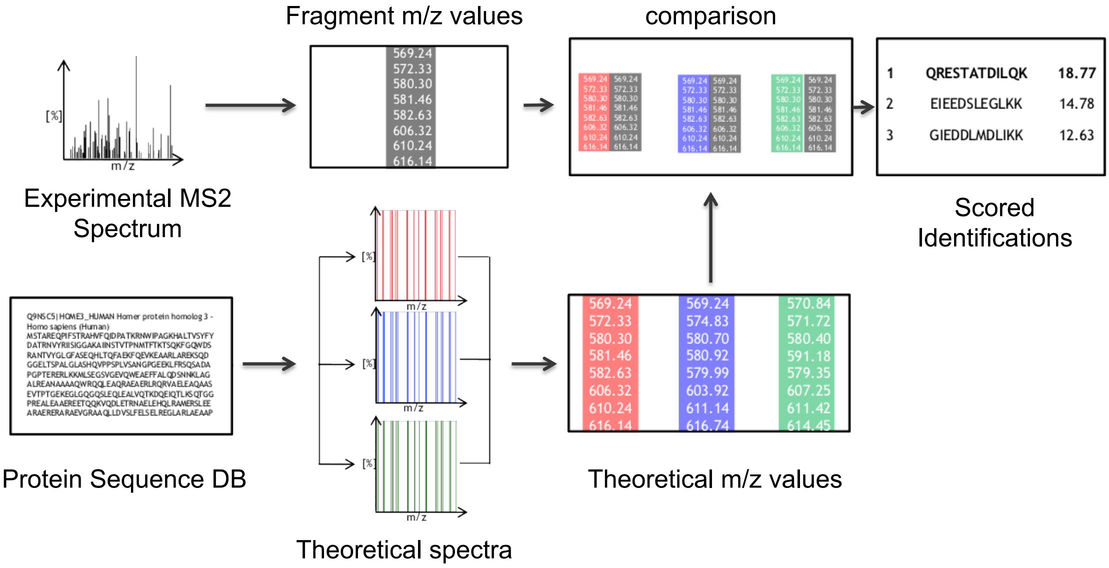


Figure 3.2: Simplified peptide identification workflow. First, the experimental MS2 spectra are recorded and theoretical spectra are computed based on a protein database. In the next step, the theoretical and experimental m/z values of the MS2 spectra are compared and finally, the identifications are reported including the sequences and software dependent scores. Reproduced from lecture 17 Slide 8 from the Bioinformatics 2 lecture from Oliver Kohlbacher. Accessible at `https://abi.inf.uni-tuebingen.de/Teaching/Old/ss-2014/BI2/slides-and-handouts/BI2_SS14_17_ProtID.pdf`.

3.1.1 Mascot

Mascot was one of the first database-based identification tools and is sold as a commercial software by Matrix Science[94]. It uses a probability-based scoring system and supports three different search types: peptide mass fingerprint, sequence query, and MS/MS ion search. Peptide mass fingerprints are the result of the digestion of a protein by an enzyme and the

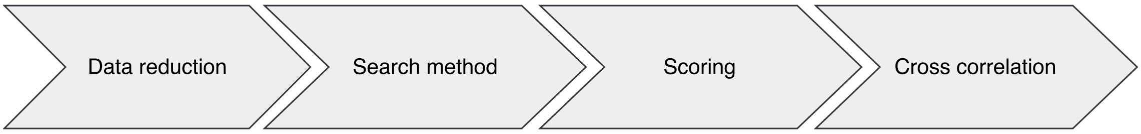


Figure 3.3: Steps of the SEQUEST workflow. First, tandem mass spectrometry data is reduced. Second, the database containing the theoretical spectra is searched. Third, the found hits are scored. Last, cross-correlation analysis of the top 500 identified amino acid sequences.

sequence query combines this mass data with AA sequence data or physicochemical data. However, in our case only the MS/MS ion query is relevant. It allows to search a sequence database in FASTA format and the search itself can be executed in parallel. As MS data input, Mascot needs peak lists with centroided mass values and intensity values.

The probability-based scoring of Mascot allows to calculate the probability of an observed match between the experimental data and each sequence in the database. The match with the lowest probability is then assigned and reported. If a multi-testing correction is performed, searching large databases with millions of sequences requires very small p values to conclude that a match is significant. Therefore, Mascot reports the $-10log_{10}(P)$ value as a more readable score to the user[94].

Mascot is server based and provides a free web interface for all three search types `http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS`. However, the search with the Matrix Science server is limited to 1,200 MS/MS spectra, which is not feasible for high-throughput mass spectrometry containing thousands of spectra. Therefore, we used an in-house Mascot server without limitation in spectra count[94]. Instead of an web interface this sever was accessed using Proteome Discoverer by Thermo Fisher.

3.1.2 SEQUEST and Comet

In contrast to Mascot, SEQUEST and Comet are based on the correlations between the theoretical and experimental spectra. This concept was first described in 1994 by Eng et al. and has later been made available as SEQUEST[40]. SEQUEST is now available as an academic version and commercial version distributed by Thermo Fisher. The original algorithm described in 1994 consists out of four steps (Figure 3.3)[40]. The first one is the data reduction of the tandem MS data. In this step first, the fragment ion mass-to-charge ratios are converted into nearest integer values and a $10\,u$ window is removed around the precursor ion. This helps to remove matches of predicted fragment ions to the mass-to-charge ratios of precursors. To eliminate noise and reduce the run time, SEQUEST only considers the 200 most abundant ions from the spectrum. In the second step, the database is searched by scanning through all protein sequences for matching linear combinations of amino acids with the mass of the peptide. This

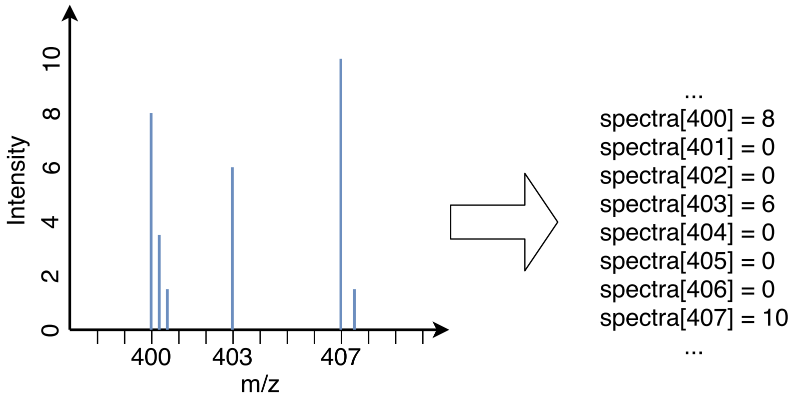


Figure 3.4: Representation of mass spectrometry data in SEQUEST as array format. The data is saved in a dictionary `spectra`, where m/z value of the spectra are keys and their intensity the values. Figure based on Eng et al.[38]

is done by summing up the masses of the amino acids of a subsequence and considering them as possible matches if they fit with a predefined mass tolerance. After searching the database for the potential hits, the final score (S_p) is calculated as

$$S_p = \frac{(\sum i_m) n_i (1+\beta)(1+p)}{n_t} \tag{3.1}$$

where n_i is the number of predicted ions matched, i_m their abundances, n_t the number of all predicted ions for this precursor, and β and p are scoring parameters. The scoring is based on multiple rules. The first rule describes how the predicted fragment ion masses are compared to the measured ones and for each match within a tolerance of 1 u, the abundance is summed up. To give preference to consecutive matches, SEQUEST increments β for each consecutive match. The second rule characterizes how the scoring parameter p is increased if an immonium ion for the amino acids His, Tyr, Trp, Met, and Phe is in the sequence, if not it is decreased. After calculating the scoring, the fourth and last step of the SEQUEST algorithm is conducted, the cross-correlation analysis. This analysis compares the top 500 identified amino acid sequences from the search results with the experimental data. First, the computational spectrum is reconstructed such that all mass-to-charge values of the b- and y-ions are represented by a magnitude of 50 and their surrounding area of 1 u by 25. To consider neutral losses of ammonia, water, and carbon monoxide, their magnitude is set to 10. This reconstruction tries to adapt the appearance of the predicted spectra to the one of the experimental spectra. In the next step, the experimental spectra are modified by removing the precursor mass and dividing the spectrum into 10 equal bins. The spectra in each bin are normalized to a magnitude of 50. Finally, the cross correlation between the theoretical spectrum x_i and the experimental spectrum y_i can

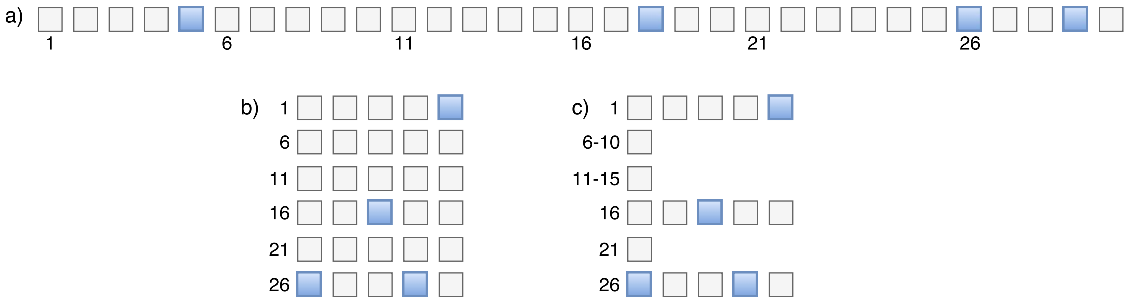


Figure 3.5: Representation of mass spectrometry data in Comet as a sparse matrix. a) Linear representation of the spectrum array. Blue boxes represent bins with spectra. b) Matrix representation of the spectrum array. c) Sparse matrix format. In rows which do not contain any values, only the first bin is assigned a null point, which allows freeing the memory of the rest of the bins in the row. Figure based on Eng et al.[38]

be calculated as

$$C_{xy} = \int_{-\infty}^{+\infty} x(t)y(t+\tau)\mathrm{d}t \tag{3.2}$$

where $x(t)$ and $y(t)$ are the continuous signals of the spectra x and y, τ is the displacement value and describes how much the signal is offset by the translation. Since x_i and y_i represent discrete input signals, the following form of the cross correlation can be used:

$$R_\tau = \sum_{i=0}^{n-1} x[i]y[i+\tau] \tag{3.3}$$

Finally, SEQUEST performs a Fourier transformation of the two spectra, ny multiplying the Fourier-transformed first spectrum with the complex conjugate of the second spectrum and then performs the inverse transformation. The last step is normalizing the resulting value to 1.

Comet is a further development of the described algorithm. To deal with new mass spectrometers with higher accuracies and larger datasets, several changes to the SEQUEST scoring method were necessary. First, the Fourier transformation of the spectra can be avoided, which is essential for the calculation of the cross-correlation and results in a shorter run time. Next, SEQUEST stores the spectra as an array, where the m/z values are keys and the intensities are the values (Figure 3.4). Since the storage consumption grows linearly, this representation is impractical as soon as the bin size for the keys increase. To allow smaller bin sizes, Comet uses a sparse matrix representation of the spectra. The aim is to avoid empty bins, which reserve large blocks of memory. This can be solved by grouping the bins and assigning `Null` to the first one in the group if all bins are empty in the group (Figure 3.5). In combination, the avoided Fourier transformation and the new data storage format allow Comet to identify and score peptides in spectra recorded with high-resolution mass spectrometers in a shorter time and with less memory usage than Sequest.

3.2 Peptide identification statistics

The following section describes how to assess the statistical significance of identified peptides and is based on a review by Käll et al.[58]. The identification tools described above only report how well the spectra fit the theoretical spectra. However, the identified peptide-spectrum matches (PSMs) may contain false positive (FP) identifications. These FPs occur often due to the complexity and erroneousness of the data, for example missing ion peaks. Raising the threshold score for accepting an identification removes many FP identifications, but may also remove many true positives (TP). Therefore, a score assessing the statistical significance (p-value) is needed. In this case, the null hypothesis is that the peptide was not measured by the MS. The most common method to calculate this p-value is the decoy database approach. In this method, the spectra are also searched against a database of decoy sequences. These decoys can be created by reversing[85] or shuffling the sequences of the target database[65]. Another method is to generate random sequences using the same amino acid frequencies from the target database. To avoid false negatives, it must be assured that the decoy sequences are not contained in the target database.

The p-value can be calculated for each PSM by determining the percentage of decoy PSMs that receive the same or higher score as the candidate PSM. Calculating the p-value for each PSM results in thousands of tests and increases the possibility of accepting a PSM by chance due to multiple hypothesis testing[18]. Multiple-testing correction and false-discovery rate (FDR) estimation are needed to solve this problem. The FDR describes the number of expected percentage of PSMs that are incorrect for a certain score threshold. In other words, we compute the ratio between the number of decoy PSMs and target PSMs above the threshold. More advanced FDR estimation methods also try to incorporate the number of false positive target PSMs.

In addition to the FDR, we can also compute the q-value. The q-values is defined as the minimal FDR threshold that accepts a certain PSM[120]. The q-value represents the significance of a single PSM, whereas the FDR represents the significance of a set of peptides. As described above, there are multiple ways to describe the statistical significance of mass spectrometry identifications. In the next section, we describe the semi-supervised machine learning algorithm Percolator[57,59,123] for calculating further and enhanced statistical properties of PSMs.

3.2.1 Percolator

Percolator was first introduced by Käll et al. in 2007[57] to improve the rate of confident peptide identification in tandem mass spectrometry. Percolator is based on a semi-supervised machine learning algorithm that can learn and use the differences between decoy and target identifications to identify true PSMs.

Percolator uses multiple different scores reported by the search algorithms as the input vector for the machine learning algorithm. Instead of looking at each score independently, the algorithm combines the scores to gain further information about the spectrum. In addition to the basic scores such as XCorr (SEQUEST and Comet correlation score), Percolator incorporates other features that describe the PSM. The most important features are ΔC_n and ΔC_n^L, which describe the difference between the current and the second/fifth XCorr, the mass and ΔM, the fraction of the b- and y-ions, the number of peptides in the database in the associated m/z range, the length of the peptide, the charge state, the number of PSMs, which are the best matching for this peptide, the number of PSMs matching into this protein, and the number of peptides for this protein. A full list of the features and a description can be found in the Supplementary Table 1 in Käll et al[57]. This list also describes enzymatic features. However these are not used for immunopeptidome data since no enzymatic cleavage is performed.

The Percolator algorithm can be divided into three phases. In the first phase the PSMs are computed for the target and decoy database and a feature vector for each PSM is computed. The second phase iterates a fixed number over the following three steps. One: generate a positive training set, which consists of a high-confidence target PSM. Two: train a linear SVM using the positive training set and the decoy hits. Three: re-rank all PSMs with the trained classifier. The third phase re-ranks the target and decoy PSMs using the final SVM. Finally, the FDR for all target PSMs is estimated as

$$E\{FDR(t)\} = \frac{\pi_0 \frac{m_f}{m_d} |\{d_i > t; i = 1, ..., m_d\}|}{|\{f_i > t; i = 1, ..., m_f\}} \tag{3.4}$$

where f_i are the scores of the target PSMs, d_i are the scores of the decoy PSMs, t is a given threshold, and π_0 is the estimated proportion of incorrect target PSMs. In addition, the q-value for a PSM with a score t can be calculated as

$$q(t) = \min_{t' \leq t} E\{FDR(t')\} \tag{3.5}$$

In 2016, version 3.0 of the Percolator algorithm was published[123], which is optimized to be used on larger datasets and allows protein interference for mass spectrometry data. However, the protein interference performed by Percolator is applicable to proteomics data and not to immunopeptidome data, because of the different chemical and biological processes, by which HLA ligands are generated in the cell.

3.3 HLA binding prediction

In silico prediction methods are available for each step of the HLA class I antigen processing pathway, which includes proteasomal cleavage, TAP transport, HLA binding, and T-cell recognition (Figure 2.1). However, the quality of the predictions varies greatly between the different parts of the pathway and is not sufficient to predict the immunopeptidome *in silico*. HLA binding prediction alone performs reasonably well. Therefore, we can use it to assign peptides to HLA types and calculate run properties like the content of binders that can be used as an additional quality metric. The next paragraph shortly describes different methods for the HLA binding prediction.

The first developed methods for HLA binding prediction where based on position-specific scoring matrices (PSSMs) (e.g., SYFPEITHI[97], RANKPEP[98], or BIMAS[92]). More recently, methods based on machine learning have been developed. The most common ones use either support vector machines (SVMs) (e.g., SVMHC[35], SVRMHC[137]) or artificial neural networks (ANNs) predictors (e.g., netMHC[78]). All these methods have advantages and disadvantages, which have been discussed in detail by Backert & Kohlbacher[13].

All described machine learning based prediction methods are supervised methods. They aim to learn a function that maps a given input (peptides and HLA allele) to its output[13]. These functions or predictors have to be trained on a dataset for which both input and output values are known. The output can be either a classification (e.g., binder vs. non-binder) or regression (e.g., binding affinity). Once the predictor is trained, it can map the input with unknown output to its corresponding output, which is also called prediction. In the next step, the predictor has to be evaluated on a test data set, which was not used in the training step. This is often done using k-fold cross-validation. In this validation method, the test data is divided into k disjoint subsets and training is performed on $k-1$ of these folds (Figure 3.6). Next, the resulting predictors are evaluated on the left-out dataset. This evaluation can for example be the computation of the receiver operating characteristic (ROC) curve for each of the sets, which then can be combined to a average ROC curve. After the evaluation of the predictor, the predictor is again trained, but this time on the complete dataset.

To develop an HLA binding predictor, large datasets of binding data for each HLA allotype have to be available. Since this data is commonly only available for the most frequent HLA alleles, allele-specific HLA-binding predictors do not cover all 13,000 known HLA alleles[105]. To solve this problem, so called pan-specific HLA-binding predictors have been developed, which allow the prediction of any HLA allele with known protein sequence. The most popular pan-specific predictors are netMHCpan[90] for HLA class I and netMHCIIpan[60] for HLA class II.

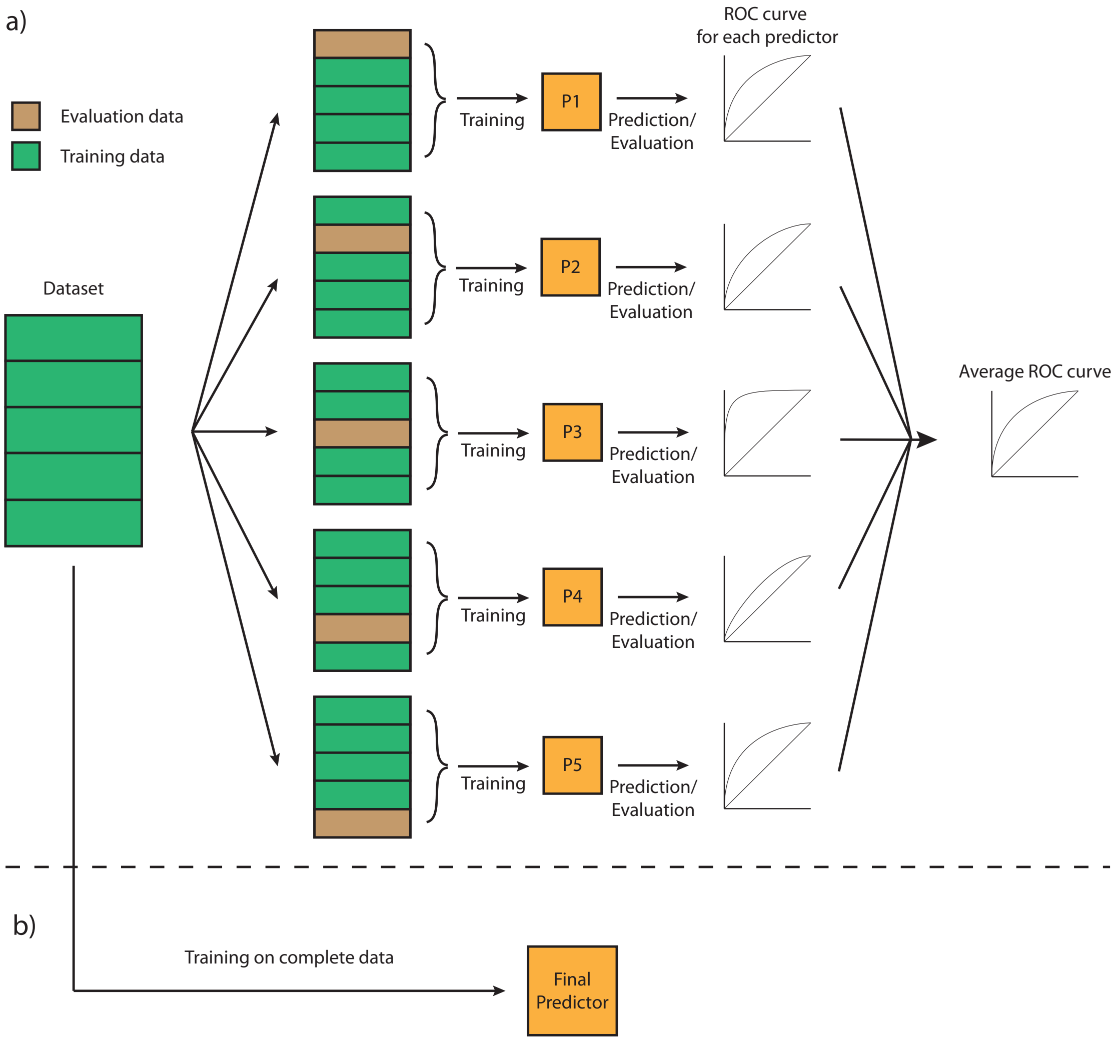


Figure 3.6: Generating predictions from data. a Evaluation of the predictor using cross-validation: first the data-set is split into k-folds ($k = 5$). Next, five predictors are trained on four folds and validated on the one left out. Evaluation can be, for example, a receiver operating characteristic (ROC) curve analysis. Finally, an average ROC curve is generated. b Training of the final predictor: after evaluation, the final predictor is trained on the complete data-set. Reproduced with permission from Backert & Kohlbacher, 2015[13].

Both predictors are ANN based and were trained mostly on data originating from HLA binding assays. To overcome the lack of data for the most HLA alleles, both methods try to measure the distance between alleles without training data (unknown alleles) and alleles with training data (known alleles). Based on this distance, closely related known alleles are used to predict the binding for the alleles without training data. In the corresponding publications[60,90], it has been shown that for known alleles these pan-specific methods perform as well as allele-specific predictors and achieve a reasonable accuracy for unknown alleles.

In this thesis, we used netMHCpan and netMHCIIpan for all HLA binding predictions, because we consider many HLA allotypes, including allotypes with few or no binding data and no allele specific predictors.

3.4 Gibbs Clustering

As mentioned above, each individual has multiple HLA alleles, which present peptides with different binding motifs and sequence properties. Therefore, the measured immunopeptidome is a mix of peptides presented by different HLA alleles. Furthermore, these peptides have lengths varying between 8-12 amino acids for HLA class I and 8-25 amino acids for class II. The Gibbs clustering method by Andreatta et al. helps to deconvolve this mix of peptides by aligning and clustering them[5,6,91]. It is based on Gibbs sampling, which is a Markov Chain Monte Carlo (MCMC) algorithm and named after the physicist Josiah Willard Gibbs[46]. It allows to get a sequence of observations, which resemble a multivariate probability distribution.

In the first step the peptide sequences have to be aligned. However, the binding motif of HLA is short and unspecific, which makes the alignment difficult[91]. The Gibbs clustering method developed by Andreatta et al.[6] solves this problem by clustering the peptides and calculating an alignment score based on the Kullback-Leibler distance (KLD). Next, a log-odds (LO) weight matrix describes the amino acid preference in each position of the alignment. The LO weight matrix for an amino acid A at position j is calculated using

$$LO_{A,j} = \frac{n}{n+\sigma} \log \frac{p'_{A,j}}{q_A} \tag{3.6}$$

where n is the number of peptides in the alignment, σ a weight for the cluster size, $p'_{A,j}$ the pseudo-count corrected frequency, and q_A the background frequency. The weight σ penalizes small, highly conserved clusters, resulting in larger and more general groups. Finally, a peptide x can be scored by summing up the LO value for each amino acid in each position.

A general problem in clustering is to determine the number of clusters. The goal is to find the number of clusters that maximizes intra-cluster fitness while minimizing the similarity between clusters. The Gibbs clustering solves this problem by calculating the relationship

$$S*_i = S_i - \lambda \max_{\substack{1 \leq n \leq g \\ n \neq i}} (S_n, 0) \tag{3.7}$$

where S_i is the score of a given peptide to *cluster i*, max$(S_n, 0)$ calculates the closest cluster to *cluster i*, and λ is the weight for the inter-cluster similarity. The Gibbs clustering algorithm then uses this equation to determine the clusters as follows. At the beginning, all peptides are distributed randomly in g clusters, where g is a number of clusters set by the user. Then the Gibbs clustering tries to align and cluster the peptides by moving them. In general, a move is accepted with the probability

$$P = \min\left[1, e^{\delta E / T}\right] \tag{3.8}$$

where δE is the energy change and T the temperature. The Gibbs clustering uses three different kinds of moves. The first is the single sequence move, which tries to move a peptide x from group G_0 to G_d. The energy for this move is calculated using $\delta E = S*_d - S*_0$. The second move is the simple shift, which tries to move a peptide X in between a group by applying a random shift of the alignment core of x. The energy is the score of x in the group after the shift minus the energy before. The last move is the phase shift, which tries to move the entire alignment of a group by a random number of positions. The energy is then calculated by subtracting the score before and after the shift.

The obtained peptides often contain contaminant peptides, which do not fit in any cluster. Because of that, the Gibbs clustering algorithm allows to include a trash cluster. This cluster is treated like a normal cluster with the exception that is not included in the overall scoring.

The described algorithm tries to find the best global solution. However, since it a is a heuristic method, it can get stuck in local optima. The Gibbs clustering tries to solve this problem by using multiple restarts with different initial seeds.

3.5 Databases and database design

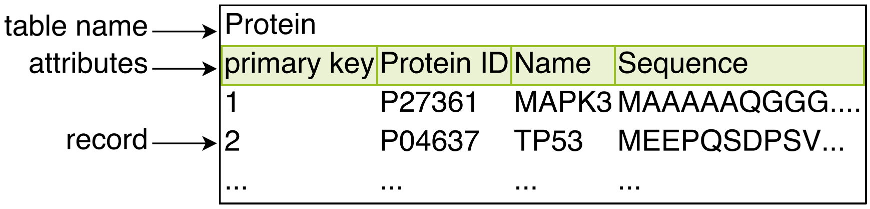
This section is based on the book "SQL- & NoSQL-Datenbanken" from Meier and Kaufmann[82] and the book "Foundations of Databases" by Abiteboul et al[2].

Today's analysis technologies, like Next Generation Sequencing (NGS) and mass spectrometry, produce large amounts of data. To store this data and to allow fast access, databases and database management systems (DBMS) are needed. The DBMS is the bridge between the user and the physical storage of the data. Most DBMS fulfill the following primary functionalities: secondary storage management (store data which does not fit into the main storage), persistence (the data should survive the termination of a database application), concurrency model (support simultaneous access), human-machine interface (allow access to the database via simple queries), distribution, compilation, and optimization (translate requests into executable programs)[2]. Furthermore, a DBMS supports actions like the definition, the creation, the querying, the update, and other administrative actions of the database[82].

There are many DBMS developers. The most used relational DBMS softwares are Oracle, MySQL (open source), Microsoft SQL Server, PostgreSQL (open source), IBM DB2, Microsoft Access, and SQLite (open source). For this thesis, only the three (MySQL, PostgreSQL, and SQLite) open source softwares are of interest. SQLite is the simplest one. It is not server-based, but is stored in a single file instead. Performance-wise, it is much slower than MySQL and PostgreSQL. Comparing MySQL and PostgreSQL is much more difficult, as they are both open source and quite similar. Two advantages of MySQL over PostgreSQL are its superior performance and a well supported connectivity to web servers. An advantage of PostgreSQL over MySQL, however, is that it is stricter if incorrect or meaningless values are inserted. For example, MySQL allows `0000-00-00` as a date, whereas PostgreSQL does not.

3.5.1 Relational databases and normal forms

In this thesis, a relational database is used. Relational databases are based on the rather simple concepts of relations or tables to represent the data. This simplicity is a major advantage of relational databases. Each table or relation in a relational database has a name (e.g., protein). The columns denote attributes (e.g., sequence) of each record. In our case, a record represents a protein with many attributes to describe it (Figure 3.7)[2].



Protein			
primary key	Protein ID	Name	Sequence
1	P27361	MAPK3	MAAAAAQGGG....
2	P04637	TP53	MEEPQSDPSV...
...	...	...	...

Figure 3.7: Example for a table in a relational database. The table represents proteins in the database with multiple attributes. Each row corresponds to one protein.

The first step in designing a relational database concept is to create an entity-relationship (ER) model. The ER model helps to design an abstract model of the data. An entity in an ER model is an object, which can be uniquely defined and exists independent of other things. In our example, a protein would be an entity and all different proteins can be described with the properties of the entity. In addition to the entities, an ER model describes the relationships between them. In our example a peptide originates from a protein and is connected via a 'has' relationship (Figure 3.8). A relationship can be described by association types.

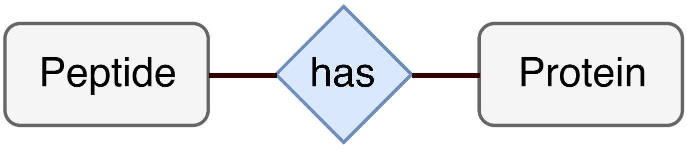
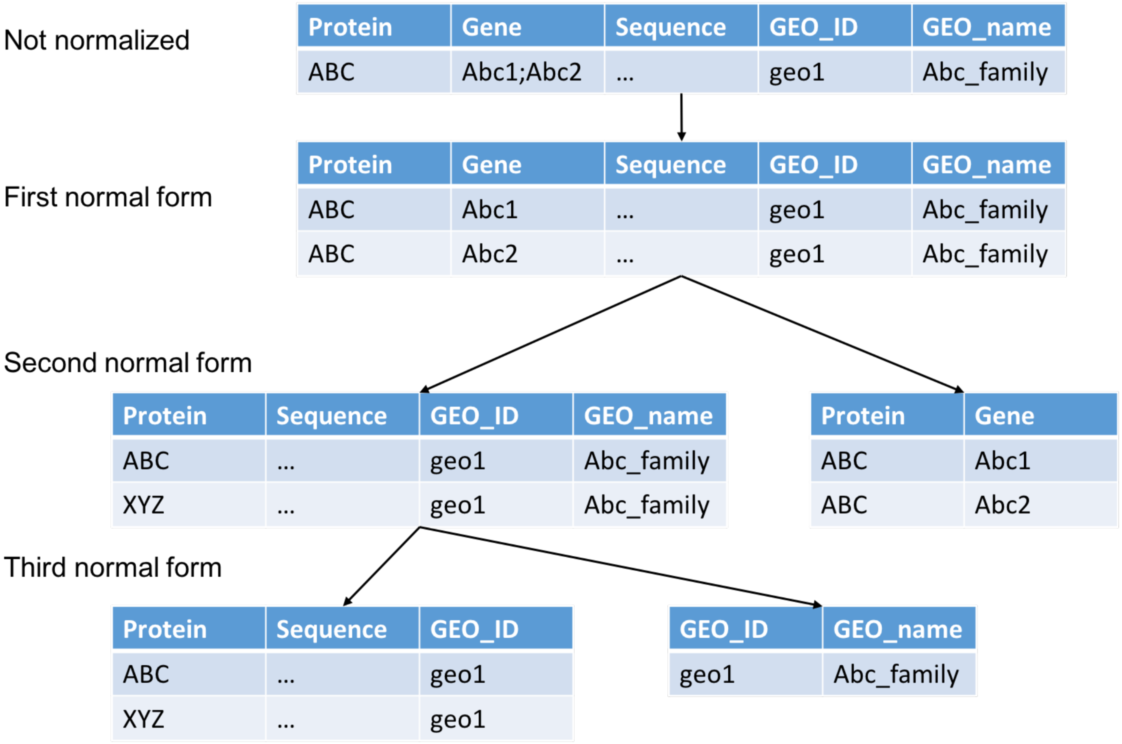


Figure 3.8: Example of a entity-relationship (ER) model. The two entities Protein and Peptide are connected by a relationship.

In a relational database, there can be three association types: one-to-many, many-to-many, and one-to-one. In a one-to-many relationship, one row in a table can be connected to multiple rows in another table (e.g., one protein to many peptides). The many-to-many relationship is used if one row in a table can be connected to multiple rows in another table, but also vice versa (one peptide can be found in multiple samples but each sample also contains multiple peptides). To construct this relationship, an intermediate table called "junction table" is needed, which contains the primary keys of the two connected tables. A primary key of a table is a property which is unique and defined for each row. In most cases, it is a consecutive integer, which is incremented if a new entry is made. A primary key can be used to connect two tables by adding it as a foreign key to a related table. In addition, foreign keys can also be another attribute from another table. Finally, the one-to-one relationship is less frequently used because this relation can be simplified represented by merging both tables into one using the connecting key.

While designing a relational database, many false or missing abstraction can be made. This so-called missing normalization can lead to redundancy, which again can cause anomalies if updates, insertions, or deletions are performed, which is illustrated by the following example: when a peptide and protein table are merged, each row in the merged table contains a peptide, its protein, and the properties of both (e.g., peptide sequence and gene name). If the gene name of a specific protein has to be updated later on, it must be updated in all rows, to prevent an update anomaly.

Normal forms have been defined to avoid such anomalies (Figure 3.9) and we will give a short overview of the first three normal forms. The first normal form enforces atomicity of the values, meaning that they can not be separated any further. In our case of a protein table, a protein may have multiple genes. These could be stored as a concatenated string. However, this would violate the atomicity criterion. As a simple solution each row could be repeated with only one gene. A relationship is in second normal form if it is in first normal form and if each non-key property is functionally independent of each key. In our example, the sequence is only dependent on the protein and not the genes. Therefore, we create a new table to represent the protein-gene relationship. The third normal form requires the second normal form and that every non-key property is not transitively dependent on any key of the table. In our case, the GEO name is a non-key property, which is transitively dependent on the key GEO ID. We can create a new table containing the GEO ID and its name to ensure the third normal form. Using these three normal forms removes most of the potential anomalies, however, there are multiple more normal forms which try to remove more specific and rare anomalies. These are described in detail in the book "SQL- & NoSQL-Datenbanken" by Meier and Kaufmann[82].



Not normalized

Protein	Gene	Sequence	GEO_ID	GEO_name
ABC	Abc1;Abc2	...	geo1	Abc_family

First normal form

Protein	Gene	Sequence	GEO_ID	GEO_name
ABC	Abc1	...	geo1	Abc_family
ABC	Abc2	...	geo1	Abc_family

Second normal form

Protein	Sequence	GEO_ID	GEO_name
ABC	...	geo1	Abc_family
XYZ	...	geo1	Abc_family

Protein	Gene
ABC	Abc1
ABC	Abc2

Third normal form

Protein	Sequence	GEO_ID
ABC	...	geo1
XYZ	...	geo1

GEO_ID	GEO_name
geo1	Abc_family

Figure 3.9: Example of the first three normal forms. The first normal form enforces atomicity. The second normal form ensures the first normal form and that each non-key property is functional independent from each key. The third normal form requires the second normal and that every non-key property is not transitively dependent on any key of the table.

3.5.2 Queries and Views

To access and update the data in a table, queries have to be written. The ANSI (American National Standards Institute) and the ISO (International Organization for Standardization) have defined a language called Structured Query Language (SQL), which is used by many different DBMS (e.g, MySQL and SQLite). SQL itself is descriptive, which means that the user describes what he wants to retrieve and does not have to write the actual code that the DBMS uses to calculate the result[82]. A simple query to retrieve the `Name` of all proteins with the `Protein_ID` equal to "P27361", would be

```
SELECT Name FROM Protein WHERE Protein_ID == "P27361"
```

The `SELECT` defines the attribute (`Name`) to query. The `FROM` describes from which table (`Protein`), and the `WHERE` clause defines criteria, which can be used to filter the result.

To describe more complex structures, like peptide-protein relationships, associations are needed (Table 3.1). These associations are defined by shared attributes. In our example, a peptide table could have the attribute `Protein_ID` used as a foreign key, which would connect the peptide table with the protein table. A query to select attributes from both tables would be:

```
SELECT Protein.Name, Peptide.Sequence FROM Protein
INNER JOIN Peptide ON Protein.Protein_ID == Peptide.Protein_Protein_ID
WHERE Protein_ID == "P27361"
```

The `INNER JOIN` describes the relationship between the peptide and the protein table and connects the two tables using the `Protein_ID`.

Table 3.1: Table to describe peptides in the database. The foreign key `Protein_ID` can be used to express the association to the protein table.

Peptide		
Primary Key	Sequence	Protein ID
1	EALAHPYL	P27361
2	AAANFRRL	P27453
...	...	...

A design following the normal forms leads to queries with many joins. Writing these large queries can be inconvenient and often is redundant. Therefore, most DBMS allow creating views. They are virtual or logical tables defined by a select statement. A view may for example contain all peptides found in a specific sample. Instead of having to write the full query with all joins, the view allows a selection in a table-wise fashion (e.g. `SELECT * FROM sample_has_peptides`). This could be also achieved by creating a materialized table with the content of the select. However, if the database is updated, the table would have to be updated as well. In contrast, views do not store the data but select them directly from the original tables. Therefore, they do not need to be updated.

Chapter 4

The HLA Ligand Atlas

4.1 Introduction and motivation

The analysis of HLA ligands helps to understand the immune system and the gained knowledge can be used to find new treatment targets[69]. These targets can then be used to develop new therapies against diseases like cancer. Unfortunately, the identification of such new targets needs a large number of samples for which the immunopeptidome has to be obtained, analyzed, and stored. Because one HLA immunopeptidome experiment for one sample can result in the identification of up to 5,000 different peptides with many parameters such as identification software specific scores or the number of Peptide-Spectrum Matches (PSMs), the storage and the analysis can be laborious[15]. Traditionally, scientists analyze and search this immunopeptidome data using Excel or other easy-to-use spreadsheet programs. However, this method is not feasible if, as in this presented study, a large number of different samples is measured. Therefore, we developed a user-friendly web interface, which allows fast and simple access to the provided data, to support biologists and biochemists in their analysis of such large datasets. This interface allows wet-lab scientist to search and perform frequent meta-analyses on the contained data using only their web browser and thus provides access to the immunopeptidome dataset for scientists regardless of computer science skills.

The data is stored in a database, in our case, a MySQL database, to allow fast queries of the data. The standard way to access such a database is via SQL queries. However, these queries require knowledge of the underlying database schemata and their usage can be difficult for the average wet-lab scientist. Databases, such as the human protein atlas[126] or ProteomicsDB[143], demonstrate that such large databases can be searched by everyone via a web interface. Therefore, we implemented the HLA Ligand Atlas, a web interface to the collected immunopeptidome data.

Other websites and databases like the SysteMHC Atlas[114], The Cancer Immunome Atlas (TCIA)[24], The Cancer Genome atlas (TCGA `https://cancergenome.nih.gov/`), and the International Cancer Genome Consortium (ICGC https://icgc.org/) already provide access to cancer specific data. TCGA and ICGC present information about the genomic, transcriptomic and epigenomic changes in cancer. This information can be obtained either as already analyzed data, like somatic mutations and overexpression of genes, or raw data, like FASTQ or BAM files. TCIA contains information on immune-related gene sets, cellular composition of immune infiltrates, HLA types, neoantigens and cancer-germline antigens, as well as tumor heterogeneity. These information are obtained using computational genomic methods and are based on data provided by TCGA, Van Allen et al.[127], and Hugo et al.[54]. Similar to our HLA Ligand Atlas, the SysteMHC Atlas provides access to peptides presented by MHC/HLA. It allows to query proteins and peptides, and to filter these results for the binding top allele and MHC class. In addition, it allows to download spectral libraries for peptides sorted by their top binding allele. SysteMHC Atlas contains 16 published human immunopeptidomics projects/datasets and 7 unpublished datasets. All 23 datasets are reprocessed using the raw data and a standardized pipeline (Comet[39], X!Tandem[31], PeptideProphet[79], and iProphet[116]). The aim of the SysteMHC Atlas is to gather and provide information and data across multiple heterogeneous projects. In contrast, the HLA Ligand Atlas is focused on one homogeneous project, in which all samples are prepared using the same protocol and are analyzed on one mass spectrometer. Furthermore, the HLA Ligand Atlas provides additional meta data, like the tissue or HLA type for each sample. This information can be queried, which allows answering more detailed questions on the immunopeptidome. On the web interface level, the SysteMHC Atlas only provides basic queries for proteins and peptides, whereas we developed an interface which provides more query options and more statistics.

This chapter first describes the design and architecture of the database and the web interface. In the next part, its implementation is described, followed by a description of the data available trough the database and how the data was processed. The last section summarizes the results and provides an outlook on the topic.

4.2 Design and architecture

The development of the HLA Ligand Atlas had two main goals. The first was to develop a fast database to store the immunopeptidome data. The second was to provicde a user-friendly web interface to access the data. The first goal was achieved with a MySQL database, the concepts of which are explained at the beginning of this chapter. In the following section, the second goal, the design and architecture of the web interface is illustrated.

4.2.1 Database concept and design

The first step of database design is to identify the data structure. This includes the structure of the raw data collected by the mass spectrometer (e.g. spectra, identified peptides and their scores) and the metadata of the underlying experiment (e.g., source, tissue, and the individual's HLA type). The curation of metadata can be very tedious and requires close communication with wet-lab scientists. Furthermore, the experimental design has to be understood to develop such a database concept.

The smallest unit in our experiment is the assignment of a spectrum to a peptide (Figure 4.1). For each spectrum, potential matching peptides (PSMs) are identified and need to be stored in the database. We call the table storing all PSSMs `spectrum_hit`. It will be the table with the most entries, caused by the thousands of spectra per MS run. These result in a table with millions of rows. Aside from the peptide information, each PSM has one or multiple assigned proteins. Technically, this information could be stored in the `spectrum_hit` as an additional column. However, this would have two major drawbacks. First, one protein will contain multiple peptides, which means the protein information will be stored redundantly in the `spectrum_hit` table. This reduces efficiency and increases data usage. Second, since assigned peptides are often contained in multiple closely related proteins, especially if the protein database contains splice variants or homologous proteins, we would either have to add an extra row (spectrum 1: protein 1, spectrum 1: protein 2,...) for each spectrum-protein pair or would have to store the protein information using string concatenation (spectrum 1: protein 1, protein 2,...).Doing so would also increase data usage. Moreover, especially the string concatenation is impractical to query.

Instead of storing the protein in an extra column of the `spectrum_hit`, we use the protein to form the second table in our database (`protein`). The number of rows is determined by the number of proteins contained in the reference database (UniProt: 20,000). After obtaining this table, we model the peptide-protein many-to-many relationship. This relationship can describe both the fact that multiple peptides will be found originating from one protein, and that one peptide can map onto multiple proteins. This many-to-many mapping table is called `spectrum_protein_map`, which contains foreign keys to the `spectrum_hit` and the `protein` table.

Next, we need to assign the spectra contained in `spectrum_hit` to their corresponding MS runs. Therefore, we describe our third main table as `ms_run`, which contains the information about the MS experiment. The relationship between the `spectrum_hit` and the `ms_run` is a many-to-one relationship: each row in `spectrum_hit` is linked via a foreign key to its row in `ms_run`. The `ms_run` table will be small compared to the `spectrum_hit` and will only contain a few thousand rows for the final dataset.

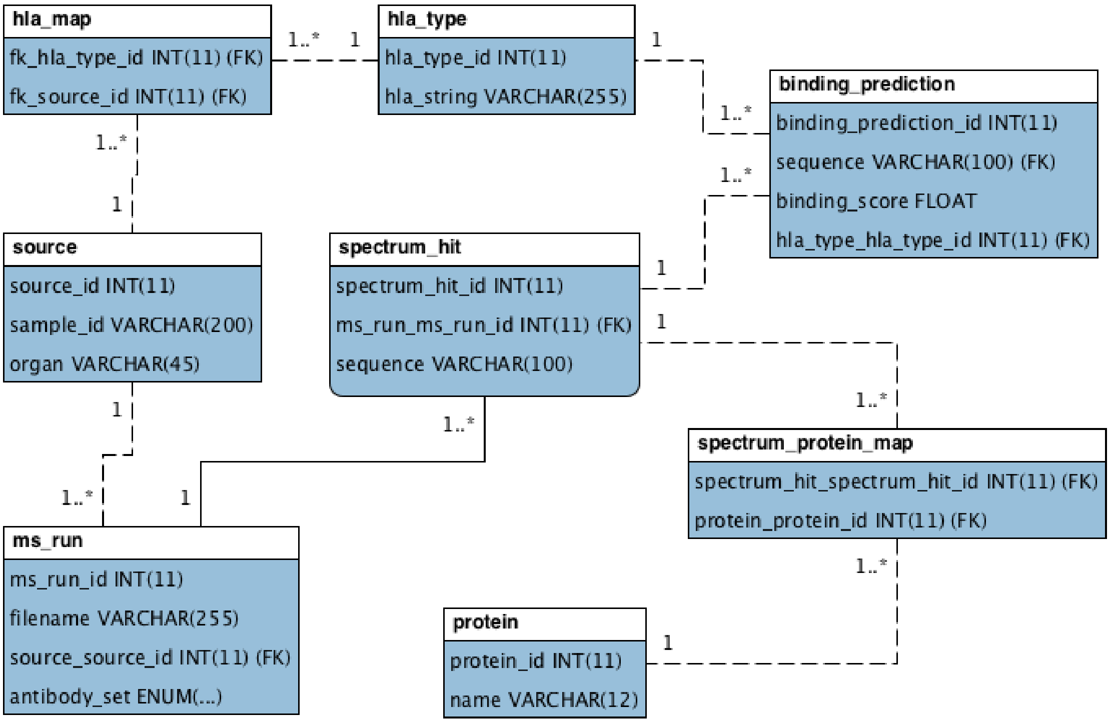


Figure 4.1: Entity Relationship Model of the database of the most important tables. It shows the relationships between spectrum hit (PSM), protein, MS run, source, HLA type, and binding prediction. The `spectrum_hit` table is the smallest unit in the database, containing the PSMs. It is connected to the `protein` table via many-to-many relationship, which is designed as mapping table `spectrum_protein_map`. Each PSM is recorded in an MS run, which is represented as the `ms_run` table. The `spectrum_hit` table has a foreign key to this table. The MS runs are then, again via a foreign key, connected to the table `source`. The HLA type is represented as `hla_type`. A mapping table, called `hla_map`, is used, to represent the many-to-many relationship between the source and the HLA type. Finally, the `binding_prediction` table contains the predicted binding scores and is linked to HLA allele in `hla_type` and the sequence in `spectrum_hit`.

Each of our MS runs belongs to a source. Each source has multiple MS runs and is represented as a separate table (`source`). The relationship is again a one-to-many relationship, solved by a foreign key in the `ms_run` table. Each source belongs to a donor meaning that each sample/tissue of a patient is represented as an individual source. However, the number of sources is small an will be between 100-300 in the final dataset.

One the most important parameters of a source is its HLA type. The number of HLA alleles for HLA class I are at most 6 (2 HLA-A, 2 HLA-B, 2 HLA-C). For HLA class II this is much more complex, caused by the pairing of the α and β chains. We created a table `hla_type` modeling each HLA allele. In addition, we created the mapping table `hla_map`, to represent the many-to-many relationship between the HLA allele and the source. The reason is the same as for the peptide-protein relationship: avoiding redundancy and allowing fast and simple queries.

The confidence of the PSMs and the rate of possible false positive identifications are the most relevant quality metrics for immunopeptidome databases. To assess these metrics, we calculate the binding affinities of all identified peptides to the corresponding HLA alleles with netMHCpan/netMHCIIpan and try to verify the identifications. These scores we stored in the table `binding_prediction`. The prediction is a combination of the sequence, stored in `spectrum_hit` and the HLA allele stored in `hla_type`. Therefore, we have two foreign keys to these two tables. The `binding_prediction` table could be even larger than the `spectrum_hit` table if all sources would have different HLA types. Since multiple tissues are analyzed from each individual and their large overlap in peptide sequences the `binding_prediction` table is only the second largest table with below one million rows.

Supplementary Figure C.1 shows an overview of the complete database concept, which contains many more tables including tables with precomputed statistics to allow a faster access to queries often requested trough the web interface. The implementation the database concept will be explained in Section 4.3.

4.2.2 Web interface design

The idea of the HLA Ligand Atlas is to enable non-computer scientists to query and analyze the large amount of data contained. The access to the data should be simple, intuitive, and fast. We thought of different scenarios or question the user might want to try or ask, to achieve an intuitive and simple web interface. These scenarios will be discussed in this section.

The very first question we came up with, is: "*I have a peptide sequence. Is it contained in your database?*" The simplest way to answer this question trough a web-interface would be a search field that allows to search for a peptide sequence. If the peptide is contained in the database, the result should show further information like which tissue it was found on and to which HLA it might bind. Furthermore, the information of the source protein should be provided and since we use MS data the individual spectra for the peptide should be displayed. These spectra would allow the expert user to decide manually if he trusts the spectrum and its assigned peptide. The implementation could be similar to the spectra viewer on `https://www.proteomicsdb.org/`[143]. Next, the question could ask for a protein in the database, which could again be queried using a search field. The result should provide all peptides that were found and are originate from the respective source protein.

The peptide and protein questions are very basic and result in only one possible hit in the database. An implementation of only these two questions could be in line with the System MHC Atlas by Shao et al.[114] However, we wanted to provide more features to the user, containing statistics and larger queries. Therefore, we thought about additional possible requests.

For example, the user might also ask for all peptides found on one tissue or HLA allele to be displayed. This question could be again solved trough a search field, but the user might not know which tissues or HLA alleles are contained in the database. To avoid wild guessing from the user, we planned to have a list of all analyzed tissues and HLA alleles. If the tissue or HLA allele is in the database, the user might want to know how many alleles and tissues are contained and how many peptides could be identified for them. In addition, he might also ask for minor statistics of the presented peptides, like tissue-specific peptides or the length distribution of the peptides assigned to the HLA allele.

In the end, the user might ask for information about the sample preparation or the analysis pipeline. Furthermore, and because it is required by law an imprint should be provided. Users may also require the possibility to download the MySQL database or a link to the RAW files on PRIDE.

All these different questions and their answers have to be connected and interactively accessible. A central navigation menu and a central search field would fulfill these requirements. The concept of the structure of the menu can be found in Figure 4.2.

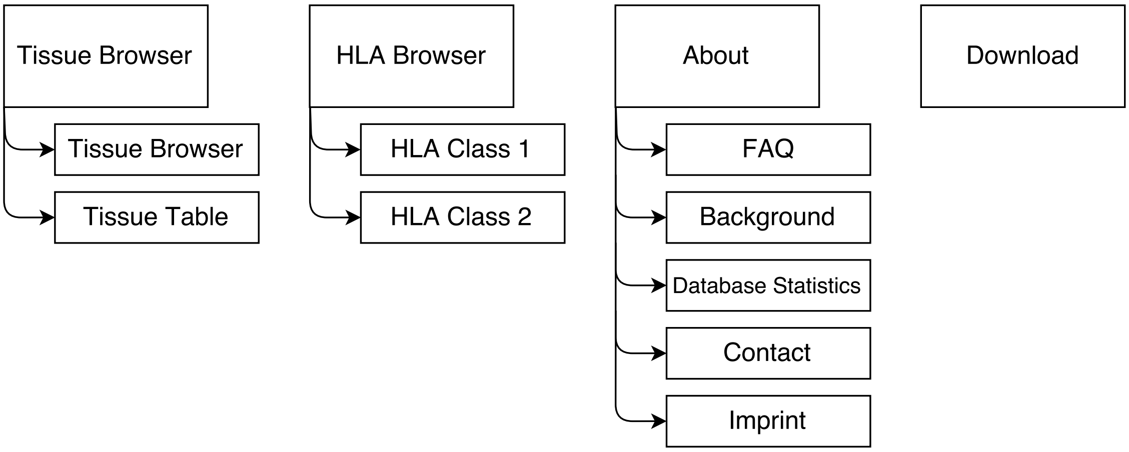


Figure 4.2: The structure of the main menu of the HLA Ligand Atlas. The main menu is structured into four drop-down menus. The first one allows accessing the *Tissue Browser* and the *Tissue Table*. The second one navigates to the *HLA Browser*, which is separated into HLA class I and II. The third drop-down menu features *FAQ*, *Background*, *Database Statistics*, *Contact*, and the *Imprint*. The last menu item leads to the download page.

4.3 Implementation

The following chapter will discuss the implementation of the database and the web interface. First, we will present the used software. Second, we will shortly describe how the database was implemented and optimized to allow fast access to the data.

4.3.1 Used software

An overview of the software described in this chapter is shown in Figure 4.3. Additionally, the used software including its versions is listed in Table 4.1.

After the collection of spectra by the MS and identification of the corresponding peptides (Chapter 4), the data has to be stored in a database to allow fast access. In our case, we use a MySQL database. The main reasons for this choice are that MySQL is a mature, open-source software with stable releases and that the data that has to be stored is well suited for a relational database. This means, that the data can be organized in formally-described tables (Section 3.5). An alternative to MySQL could be SQLite. It is easier to handle, since no database server is needed (everything is stored in one file), but performance-wise it is much slower than MySQL and not suited for large datasets, such as those contained in the HLA Ligand Atlas. Another alternative could be a NoSQL solution like MongoDB. In contrast to MySQL, MongoDB is a non-relational database which is beneficial if the data is not structured. Due to the relational data and its superior performance, we decided to use MySQL. After the development of an appropriate database design, the data was imported and the web interface was implemented. We used the Pyramid web framework as the basis for the development. Pyramid is based on Python and has been developed as a part of the Pylons Project. Like MySQL, it is open source software. There are many other web frameworks, which allow developing websites. The simplest solution would have been to develop out interface without any web framework instead using plain HTML. However, this is very impractical. We chose Pyramid for two reasons: due to its good connectivity to the MySQL database and, due to it being based on Python. Compared to other programming languages, Python is easy to learn and is nowadays one of most popular programming language. Besides Pyramid, there are several of other big Python-based web frameworks like Django, Flask, and web2py. Especially Django is very similar to Pyramid, and our decision to chose Pyramid over Django was made at random.

We use SQLAlchemy to connect the web interface and the MySQL database. SQLAlchemy is a Python library which represents MySQL tables as Python objects. Using these objects, it is much simpler to retrieve data from the database than with plain MySQL. Furthermore, SQLAlchemy validates queries before submitting them to the server, which can detect simple problems (e.g., selecting unknown columns) and more complicated ones (e.g., incorrect joins). It also tries to optimize the query before execution and provides a sophisticated SQL injection protection.

The HLA Ligand Atlas is written in HyperText Markup Language (HTML). In addition, we used multiple JavaScript (JS) libraries to enhance the view and usability. The most used library is JQuery, a small, feature-rich JS library that simplifies working with HTML objects and JS itself.

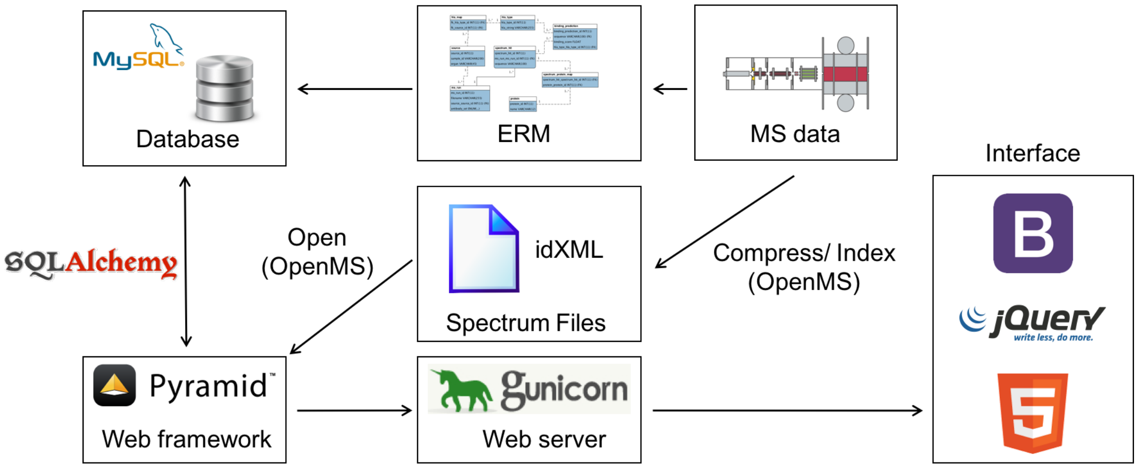


Figure 4.3: Overview of the software used to implement the HLA Ligand Atlas. First, the data from mass spectrometer is modeled with an ERM and then stored in a MySQL database. The website was developed using the web framework Pyramid. The communication between the database and the web framework is achieved by SQLAlchemy. The interface was implemented using Bootstrap, JQuery, HTML and other JS libraries. The website was run on a Python-based web server called Gunicorn. The MS data is compressed in parallel and indexed as idXML files using OpenMS and later parsed and sent to the server using OpenMS and the JSON file format. Database image from `http://cliparts.co/clipart/2829444`; File image from `https://openclipart.org/detail/38899/new-file`.

The main design was implemented using the Bootstrap framework, which claims to be the most popular HTML, JS, and Cascading Style Sheets (CSS) framework for mobile first projects. This means that the HLA Ligand Atlas is developed not only for desktop PCs with normal screen sizes but also for mobile devices with very small screens, too. Although the target user will probably use a desktop PC, the website is displayed correctly on smaller devices like mobile phones. Bootstrap enables the programmer to develop web interfaces, which look state-of-the-art, but are very easy to design. All interactive tables are implemented using the DataTables JS library. DataTables supports various data formats and we used JavaScript Object Notation (JSON) to hand the data from the Pyramid framework to the web interface. JSON is human readable and language independent, which allows passing JSON objects between Python and JS. DataTables provides options to adjust the tables for many purposes. For example, DataTables allows pagination, multi-column ordering, searching inside the table, and downloading the table in multiple formats. Additionally, there is a Bootstrap theme for DataTables, which was used to achieve a uniform style for the website. The diagrams were implemented using HighCharts (Figure 4.10), which allows adding interactive charts using JS. It is available under a free non-commercial license. The communication between the python framework and HighCharts is again based on JSON. Although we used HighCharts only to create bar charts, it

Table 4.1: List of software used for the development of the HLA Ligand Atlas, including the context of usage and their version.

Software	Usage	Version
MySQL	Database	5.7
SQLAlchemy	Server database communication	1.0.9
Pyramid	Web framework	1.5.7
Gunicorn	Webserver	19.7
HTML	Webdesign	5
CSS	Webdesign	-
JavaScript	Webdesign and functionality	-
JQuery	Webdesign and functionality	2.1.4
Bootstrap	Webdesign	3.3.5
DataTables	Interactive tables	1.10.9
HighCharts	Interactive charts	4.1.9
OpenMS	idXML filtering and indexing	2.1.0
Lorikeet	Spectrum viewer	-

supports much more diagram types, like area charts, line charts, pie charts, and many other chart types.

For experts in mass spectrometry, we included a spectrum viewer in the web interface. It is based on the JS plugin Lorikeet by John Chilton `https://github.com/jmchilton/lorikeet`. Lorikeet uses spectrum data in JSON format. The Python library of OpenMS is used, to read the spectrum data stored in idXML files.

A web server is needed to run and serve the website. We use Gunicorn (Green Unicorn), a Python-based WSGI HTTP server, which runs the Python-based Pyramid web framework and claims to be light on server resources and fairly fast. An alternative was an Apache server. However, Apache would need to run WSGI programs an extra module called mod_wsgi. Therefore, we decided to use Gunicorn.

The code of the HLA Ligand Atlas is freely accessible at `https://github.com/linusb/ligandomat-2.0`. It can be used under the BSD 3-Clause License, which means it free for private and commercial use. Furthermore, it is allowed to modify and distribute the code, but there is no guarantee for liability and no warranties are provided. The only condition for using the code is that the license and the copyright notice is always included.

Database implementation and optimization

The main tables of the database were implemented based on the model described in subsection 4.2.1 (complete ERM Supplementary Figure C.1) and the required parameters were added to each table. Each row in the `spectrum_hit` now stores further parameters besides the sequence, like the retention time, the mass-to-charge ratio (m/z value), the charge of the peptide, the injection time, the mass of the peptide, the mzML ID (for the spectrum viewer), the precursor area, the search engine score, the *e*-, and the *q*-value. The `protein` table is comprised of the protein name, the gene name, the protein sequence, and the protein description. The `ms_run` table contains the antibody and its mass, as well as the original filename, the date, the sample mass and volume, and many flags for the processing of the data. The `source` table includes many parameters to describe the metadata, like the organ/ tissue, the organism, the immunoprecipitation date, and the organism/ patient ID (`ZHXX`). The `hla_type` table now the describe an HLA allele with the name and the number of digits (2: A*02, 4: A*02:01). Finally, the `binding_prediction` table contains the binding score, the binding rank, the method (netMHCpan or netMHCII and version), and a boolean indicating whether the peptide binds to the HLA or not (rank < 2). The data types for each column of the tables were determined by a trade off between flexibility (e.g., text) and smallest possible size (e.g., tiny integer).

One of the biggest issues in the development of the HLA Ligand Atlas was to achieve a reasonable performance for the required queries. Since it is a web interface, the user may only be willing to wait a few seconds for the page to load. Therefore, all queries on one page should take less than 5 seconds. Achieving this loading time was not possible when using only the main tables described in subsection 4.2.1. Therefore, many different optimizations were implemented to reduce loading time. In this section, we will give some examples of the used optimization techniques.

In the MySQL database, each matched spectrum of each MS run is represented by an entry in the `spectrum_hit` table. A spectrum hit is the smallest unit of an MS experiment with all spectrum information. However, most of the times users search for the peptides in a run, which means that all spectra for each peptide have to be accumulated. This can be done in MySQL using `group by` statements in combination with `aggregate` functions like `Count` and `Sum` (Algorithm D.1). When a user searches for the peptides in a run, the MySQL server has to calculate the query containing the `group by` and the `aggregate` functions. In general, MySQL caches the result of frequent queries in temporary tables, but not if `group by` and `aggregate` functions are combined. To overcome this issue we used a materialized view, called `peptide_run` (Supplementary Figure C.1). For each run and peptide it contains the accumulated information and has to be recomputed only when the database is updated. This precomputed table reduces the time needed to find all peptides in a run from minutes to milliseconds (Table 4.2).

The statistics shown on the different pages of the HLA Ligand Atlas create an additional problem with performance. It takes bewtween multiple seconds and minutes to query and compute these statistics. Because all pages should load rapidly, the statistics cannot be calculated each time a user accesses one of these pages. Therefore, all the statistics are precomputed and stored in additional tables. Especially the calculation of tissue-specific and HLA-specific

Table 4.2: Benchmark of the query for all peptides and their properties in one or multiple MS runs. The old query is shown in Supplementary Algorithm D.1. The new query accesses all columns in the new materialized view `peptide_run`.

Number of MS runs	Old [s]	New [s]
all (377)	277.254	0.005
300	207.550	0.008
200	106.467	0.002
100	52.280	0.002
50	8.839	0.002
1	0.068	0.003

peptides is time-consuming and therefore stored in the tables `tissue_specific_peptides` and `tissue_hla_specific_peptides` (Supplementary Figure C.1).

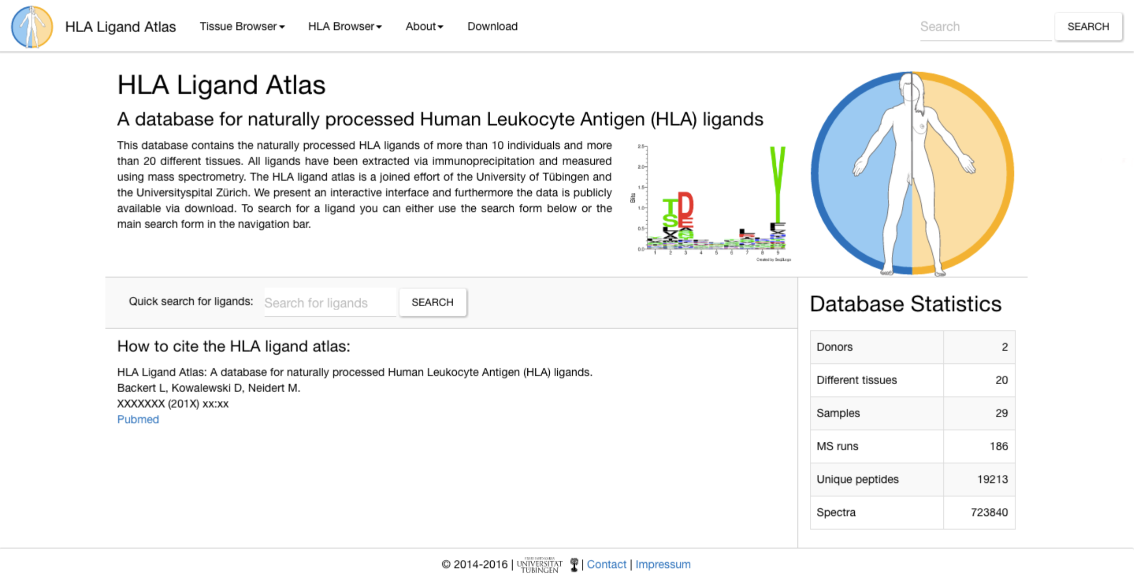
The implementation of the spectrum viewer resulted in the problem that large amounts of data had to be stored. Considering that each peptide has multiple spectra and each spectrum contains multiple peaks, this quickly results in millions of data points and multiple gigabytes of data. Therefore, we stored the spectrum/peaks data in IdXML files, which are saved and indexed for each run separately, to avoid having to store this large amount of data in the MySQL database. The IdXML file format is XML based and optimized to store MS identifications. The indexing of these files allows a very a fast access of individual spectra. Furthermore, we added a column in the `spectrum_hit` table containing the index of the spectrum in the file to access these indexes. These IdXML files can be very large (up to 1 GB) and have to be stored for each MS run, resulting in large storage costs. Therefore, we preprocessed the IdXML files and removed all un-assigned spectra, resulting in files approximately a tenth of their original size. The files are stored locally on the server hard drive. If the spectrum viewer is opened, the corresponding file is opened using OpenMS and the spectra information are extracted using the index of the file. The read information is then converted to JSON and sent to the server. However, this implementation assumes that most of the users do not access the spectra since it would result in a very high hard drive usage and a bad overall performance. If this assumption is not met and there are more requests of the spectrum viewer than expected, there are two possible solutions. One would be to move the data to a fast storage (e.g., solid state discs). A second possibility would be to use a second MySQL server with a spectral database to reduce hard drive usage. However, this would require the development of an additional database scheme, which was not reasonable for a minor feature of the HLA Ligand Atlas.

Web interface implementation

Concurrently to the implementation of the database concept, we implemented the design of the web interface. We programmed all features described in subsection 4.2.2, and their implementation is described in the following section.

Homepage

The first page shown by the browser is called the homepage. Its content and design are most crucial, because the user decides based on a first glance whether he will browse further. We decided to provide a short overview text about the database and basic statistics on the homepage (Figure 4.4). These two elements are important for first-time users, as they provide an overview of the web page and furthermore show the amount of data contained. In addition, we added a quick search bar, to allow for fast data access. This element allows searching the database for peptides by simply typing their sequence into the search field. Furthermore, the main navigation menu can be accessed trough the header of the web interface. It will be described in the next paragraph.



HLA Ligand Atlas

A database for naturally processed Human Leukocyte Antigen (HLA) ligands

This database contains the naturally processed HLA ligands of more than 10 individuals and more than 20 different tissues. All ligands have been extracted via immunoprecipitation and measured using mass spectrometry. The HLA ligand atlas is a joined effort of the University of Tübingen and the Universityspital Zürich. We present an interactive interface and furthermore the data is publicly available via download. To search for a ligand you can either use the search form below or the main search form in the navigation bar.

How to cite the HLA ligand atlas:

HLA Ligand Atlas: A database for naturally processed Human Leukocyte Antigen (HLA) ligands.
Backert L, Kowalewski D, Neidert M.
XXXXXXX (201X) xx:xx
Pubmed

Database Statistics

Donors	2
Different tissues	20
Samples	29
MS runs	186
Unique peptides	19213
Spectra	723840

© 2014-2016 | UNIVERSITÄT TÜBINGEN | Contact | Impressum

Figure 4.4: Homepage of the HLA Ligand Atlas. The homepage contains an information text about the website, basic statistics of the database, a quick search field that can be used to query the database for peptide sequences, and the navigation bar to browse through the content of the database.

Navigation menu

The navigation menu allows to browse the web page in a straightforward way. It is shown on all pages of the web interface and leads the user to the most important navigation pages and content (structure, see Figure 4.2). The first element directs to the Tissue Browser and the second element to the HLA Browser. Both pages are main navigation pages. The Tissue Browser allows looking at the data with a focus on different tissues types. The HLA Browser allows access via HLA types. Both are explained in the next paragraph in detail. The About section leads to Frequently Asked Question (FAQ), the Background, a short information page about the experimental methods used, the Contact information and finally the imprint. The last element, the Database download, allows accessing all the data stored in the HLA Ligand Altas in spreadsheet format and as a MySQL dump. In addition to the navigation menu, the header of the web page redirects the user to the home page by clicking on the HLA Ligand Atlas logo. Furthermore, we implemented a search field in the upper right corner, which will be explained in detail later.

Tissue Browser and HLA Browser

The tissue browser is an exploratory navigation menu and it enables users to quickly see the available tissues and navigate to tissue-specific pages within the HLA Atlas. Each tissue is shown as pictogram (Figure 4.5). Because the HLA Ligand Atlas contains more tissues than an illustration as individual pictograms would allow, we developed two features to allow a faster and reasonable view on all kinds of tissues. First, we changed the color of gender-specific (in this case female). This shall allow the user to find gender-specific tissues faster and to differentiate them from shared tissues. Second, we grouped related tissues into tissue group pictograms. For example, we grouped different kinds of neurological tissues under the brain pictogram. When a user clicks the brain pictogram, a drop-down menu opens and shows all neurological tissues, i.e. brain, cerebellum, small brain, myelon and bone marrow (Figure 4.6). Now, when the user clicks on a tissue name, he is forwarded to the page of the selected tissue.

The pictographic symbols for the tissues are an intuitive way to discover the different kinds of tissues in the HLA Ligand Atlas. However, experienced users may want to have an exhaustive view of all included tissues at a glance. Therefore, we additionally provided the Tissue Table. It is an interactive table that lists all tissues and contains the number of samples per tissue. This table is, like all interactive tables in the HLA Ligand Atlas, sortable and searchable.

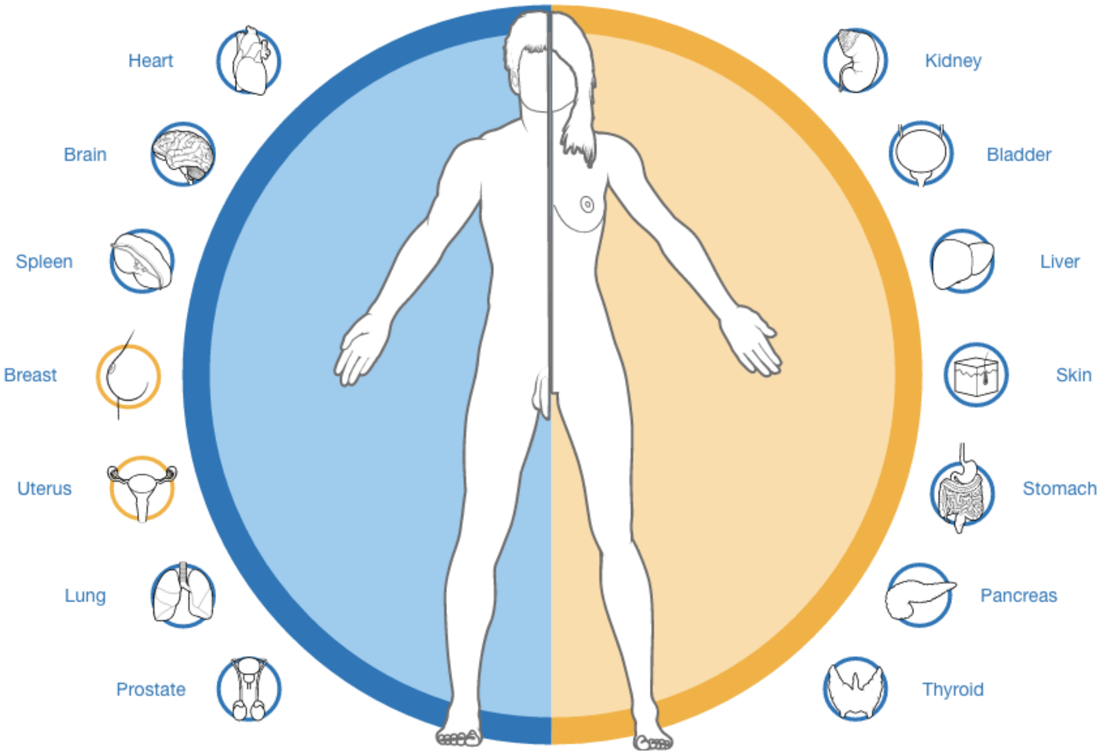


Figure 4.5: Tissue Browser of the HLA Ligand Atlas. The Tissue Browser shows all tissue groups included in the HLA Ligand Atlas as intuitive pictograms. When a user clicks on the pictograms, a drop-down menu opens containing links to all sub-tissues in the group (Figure 4.6). Female gender-specific tissues are color coded in orange, to allow a differentiation from tissues that are not gender-specific.

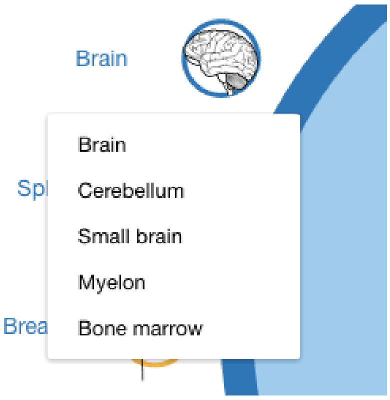
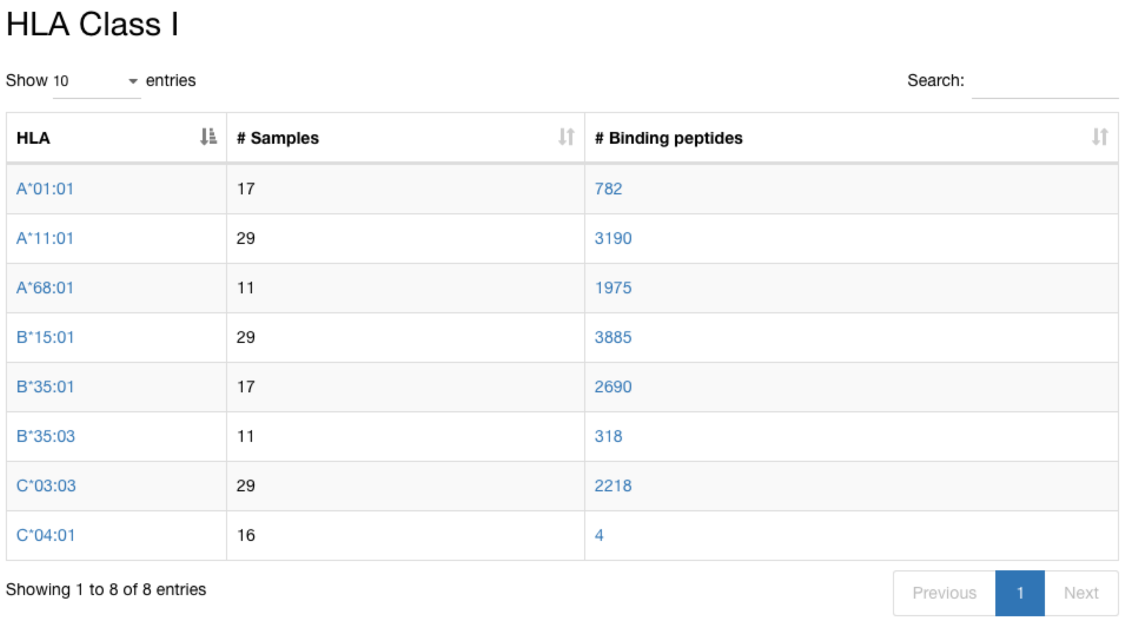


Figure 4.6: The drop-down menu of the Tissue Browser (Figure 4.5). The drop-down menu allows the access to sub-tissues that where grouped into one pictogram. In this example, the brain pictogram encompasses all neurological tissues, like brain, cerebellum, small brain, myelon, and bone marrow. When the user selects one of the tissues in the menu he is forwarded to the corresponding tissue page.



HLA Class I

Show 10 entries

Search:

HLA	# Samples	# Binding peptides
A*01:01	17	782
A*11:01	29	3190
A*68:01	11	1975
B*15:01	29	3885
B*35:01	17	2690
B*35:03	11	318
C*03:03	29	2218
C*04:01	16	4

Showing 1 to 8 of 8 entries

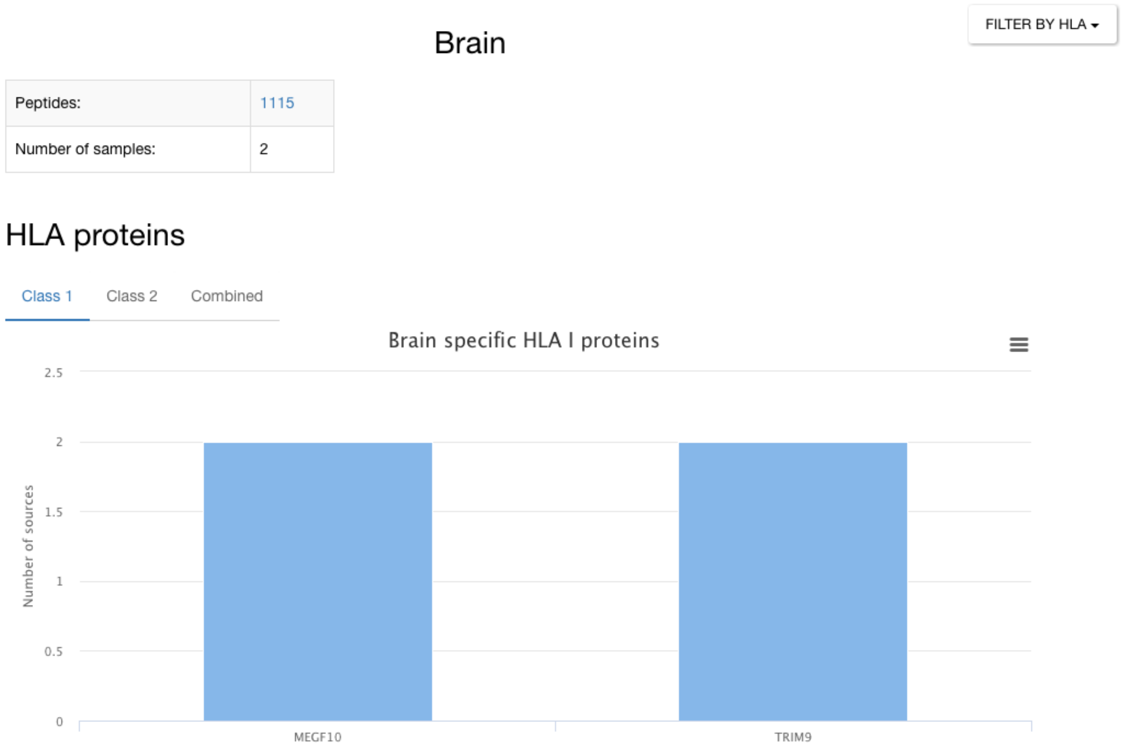
Figure 4.7: HLA Browser for HLA class I. The HLA Browser lists all HLA class I types that are available in the HLA Ligand Atlas. For each HLA type, the number of sources and the number of assigned peptides is shown. The table is sortable and can be searched using the search field in the upper right corner.

The second major navigation tool is the HLA Browser, which allows the user to get an overview of all HLA types available in the HLA Ligand Atlas. This information is visualized in an interactive table with further details such as the number of samples and the number of binding peptides for each HLA type (Figure 4.7). When the user clicks an HLA type, he is forwarded to the individual page of the selected HLA type. We divided the two HLA classes into two separate tables to provide a concise overview of each class.

Tissue page

Through the Tissue Table or the Tissue Browser, the user arrives at the tissue page (Figure 4.8). This page presents all gathered information about the selected tissue and answers the most common questions asked by biologists, such as how many samples were collected for this type of tissue or: How many and which peptides were identified on this tissue type. A short table that contains this basic information answers both questions. Additionally, the user can click on the number of peptides to access all identified peptides. This triggers a database search for all peptides identified on the tissue, which is presented as an interactive table (Figure 4.9).

After the basic overview of the tissue, we provide a more comprehensive data analysis for each tissue. By comparing all tissue types, we identified all proteins and peptides, which were found only on the selected tissue. These peptides are presented in an interactive plot, with



Brain

Peptides:	1115
Number of samples:	2

HLA proteins

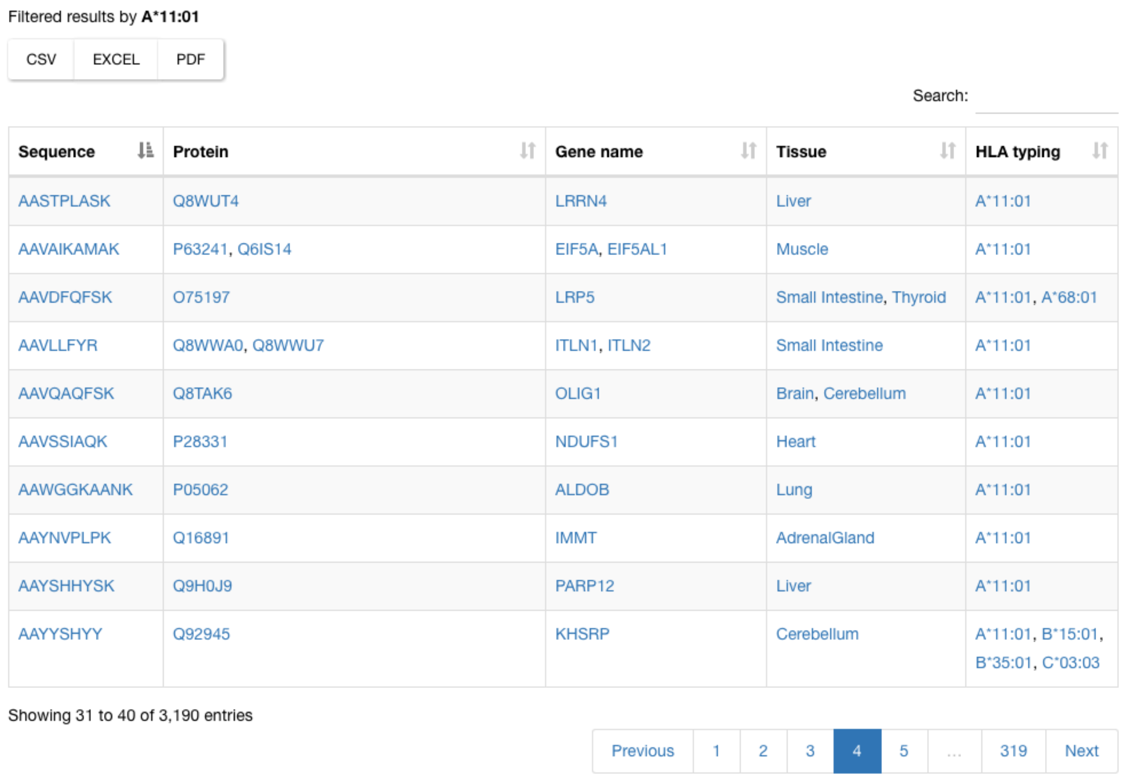
Figure 4.8: The tissue page for the brain. It contains statistics on the number of found peptides and the number of samples. Furthermore, an interactive table on the comprehensive analysis of tissue-specific proteins is shown. In this case, only two brain-specific HLA class I proteins have been found. The interactive plots allow to access further information of the proteins by clicking on the bars, which redirects to the corresponding protein page. A drop-down menu in the upper right corner allows the user to filter the tissue for a specific HLA type.

additional features such as clickable protein bars. When the user clicks the bar of peptides he is forwarded to the specific protein page. The plot and the underlying data can be downloaded in various formats (Figure 4.8).

Finally, we added a filter option to allow a combined analysis of tissue and HLA. Selecting this filter leads to a page with a similar presentation of information, where the data is filtered based on a combined search for the tissue and the selected HLA type (e.g., Brain and A*01:01).

HLA page

The HLA page gives an overview of each HLA type (Figure 4.10). A table shows the number of binding peptides and the number of sources (upper left). Furthermore, an interactive diagram (right) gives a more comprehensive view on which tissues are contained in the database for the selected HLA type. The pictograms represent the same tissue groups as used in the Tissue



Filtered results by **A*11:01**

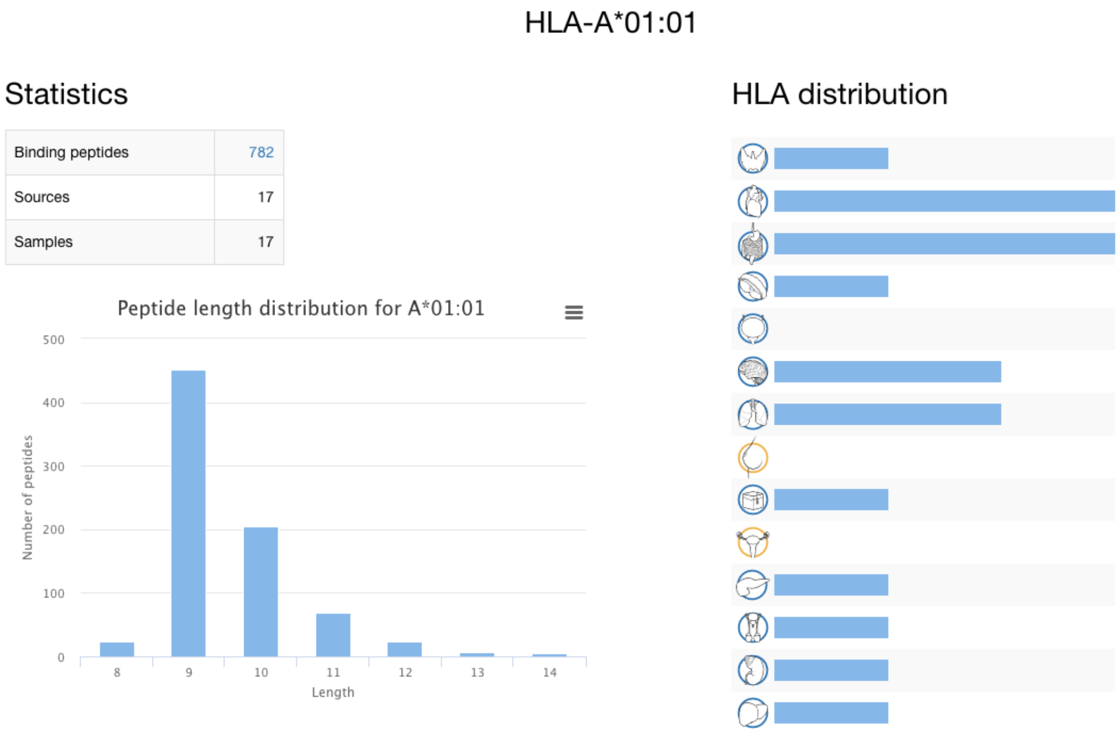
Sequence	Protein	Gene name	Tissue	HLA typing
AASTPLASK	Q8WUT4	LRRN4	Liver	A*11:01
AAVAIKAMAK	P63241, Q6IS14	EIF5A, EIF5AL1	Muscle	A*11:01
AAVDFQFSK	O75197	LRP5	Small Intestine, Thyroid	A*11:01, A*68:01
AAVLLFYR	Q8WWA0, Q8WWU7	ITLN1, ITLN2	Small Intestine	A*11:01
AAVQAQFSK	Q8TAK6	OLIG1	Brain, Cerebellum	A*11:01
AAVSSIAQK	P28331	NDUFS1	Heart	A*11:01
AAWGGKAANK	P05062	ALDOB	Lung	A*11:01
AAYNVPLPK	Q16891	IMMT	AdrenalGland	A*11:01
AAYSHHYSK	Q9H0J9	PARP12	Liver	A*11:01
AAYYSHYY	Q92945	KHSRP	Cerebellum	A*11:01, B*15:01, B*35:01, C*03:03

Showing 31 to 40 of 3,190 entries

Figure 4.9: Interactive result table for a query of the database. The database was searched for all peptides presented by HLA-A*11:01. Besides the peptide sequence, it contains information about the source protein, the gene name, the tissue it was found on, and the assigned HLA type. The table is sortable and allows pagination. Furthermore, all data in the table can be downloaded as Excel, CSV, and PDF file. For a secondary search inside the table, a search box is provided.

Browser. Clicking on a bar opens in a drop-down list that contains a detailed view of the tissues in the category and the corresponding number of available samples for the HLA type. Accessing one of the individual tissues refers to a view with combined information on the HLA type and the tissue, as described in the paragraph above.

Already at the very beginning of the discovery and research of the HLA molecule, binding motifs for individual HLA types were described[97]. These binding motifs can be best displayed as sequence logos. Therefore we provide sequence logos that were generated with Seq2logo[125] for each HLA allele and peptide length (Figure 4.11). The peptide length distribution for each allele is also included in the HLA page as a diagram (lower left in Figure 4.10). Clicking the peptide length bar searches the database for peptides of the HLA type with the specific length.



HLA-A*01:01

Statistics

Binding peptides	782
Sources	17
Samples	17

HLA distribution

Figure 4.10: Overview of the HLA page for A*01:01. The HLA page contains a small statistics table, which contains the number of binding peptides and samples with the HLA type in the database. Furthermore, the number of samples from each tissue for the HLA type is shown as a bar plot (right). To access all sub-tissues of a tissue group the user can click the bar to open a dropdown menu with a more detailed list. The peptide length distribution is shown in the lower left. The bars can be clicked to query the database for all A*01:01 peptides of a specific length. The bar plot and the data behind it can be downloaded using the drop-down menu in the upper right corner of the plot. The HLA page also contains a plot of the peptide binding motifs (Figure 4.11).

Protein page

The protein page provides information on the protein such as the gene name and the UniProt accession ID and shows the complete protein sequence with all peptides found in the protein underlined (Figure 4.12). This view can present proteins of any length and can underline up to sixteen overlapping peptides. Therefore, length variants of peptides, which are a common case in immunopeptidome data, can be found by just looking at the protein sequence. Additionally, peptides found in the protein are also shown in a tabular format. This table shows if a peptide was also found in other proteins, which is a very important information for HLA studies. A more detailed table that shows similar information about the peptides found in the protein similar to the table is shown in Figure 4.9.

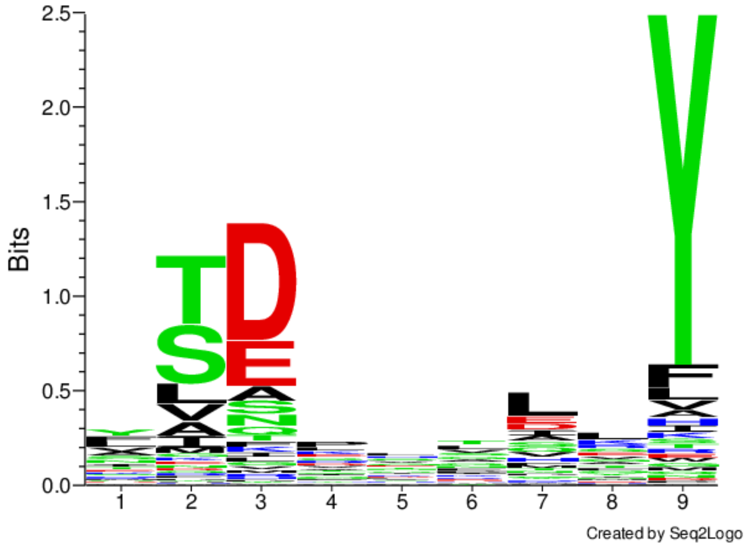
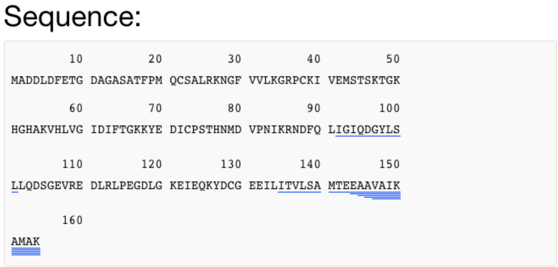


Figure 4.11: Peptide binding motif view for HLA A*01:01. Binding motifs are calculated for each peptide length and can be accessed via the tabs. The sequence logos were created using Seq2Logo[125]. The height of the letters is equal to the information they contain at each amino acid position. The colors of the letters symbolize amino acid properties.



Sequence:

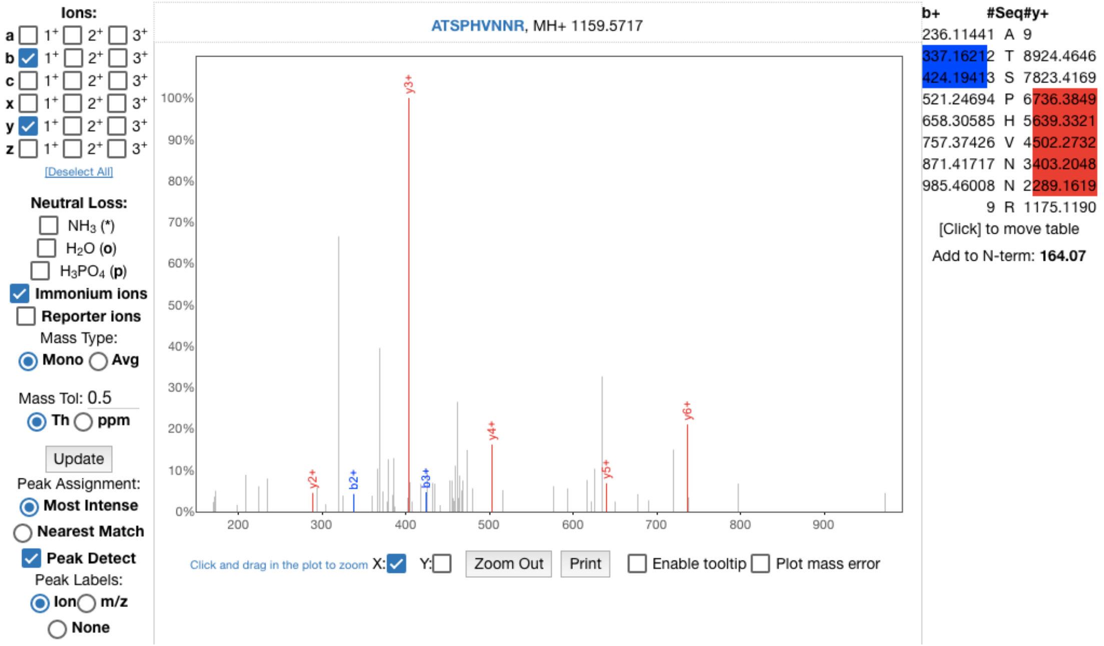
```
        10         20         30         40         50
MADDLDFETG DAGASATFPM QCSALRKNGF VVLKGRPCKI VEMSTSKTGK

        60         70         80         90        100
HGHAKVHLVG IDIFTGKKYE DICPSTHNMD VPNIKRNDFQ LIGIQDGYLS

       110        120        130        140        150
LLQDSGEVRE DLRLPEGDLG KEIEQKYDCG EEILITVLSA MTEEAAVAIK

       160
AMAK
```

Figure 4.12: The sequence of a protein as shown on the protein page. Blue lines mark peptides and their position in the protein. This illustration is suitable for proteins of any length and can mark up to sixteen overlapping peptides. Starting from sequence position 135 a cluster of overlapping peptides that contains multiple length variants of one core peptide is shown.

Peptide and spectrum page

The smallest units of information in the database are the peptides and the corresponding spectra. The peptide page shows from which proteins the peptide originates and its position in the source protein. For each peptide, we predicted the binding affinity and the rank to identify its source HLA. In addition, it also shows on which tissues the peptide was found in a similar fashion as the tissue distribution shown on the HLA page (Figure 4.10).

The peptide page also provides access to the most crucial and for some biologist the most important information about the peptide: the individual spectra of the peptide. When the user requests the spectra, a spectrum viewer opens and shows the spectrum for the peptide (Figure 4.13). This viewer allows marking different ions, like *y*-, *b*-, *a*-, and *z*-ions and assigns their charge in the view. The spectrum viewer can also mark neutral losses (NH_3, H_2O, and H_3PO_4), immonium ions, and reporter ions. The mass tolerance can be set to either *Th* or *ppm* and the peak can be assigned to either the most intense peak or the nearest match. For offline comparison, a static view of the plot can be downloaded or printed via the spectrum viewer.



Ions:
a □ 1^+ □ 2^+ □ 3^+
b ☑ 1^+ □ 2^+ □ 3^+
c □ 1^+ □ 2^+ □ 3^+
x □ 1^+ □ 2^+ □ 3^+
y ☑ 1^+ □ 2^+ □ 3^+
z □ 1^+ □ 2^+ □ 3^+
[Deselect All]

Neutral Loss:
□ NH_3 (*)
□ H_2O (o)
□ H_3PO_4 (p)
☑ Immonium ions
□ Reporter ions
Mass Type:
● Mono ○ Avg

Mass Tol: 0.5
● Th ○ ppm

Update
Peak Assignment:
● Most Intense
○ Nearest Match
☑ Peak Detect
Peak Labels:
● Ion ○ m/z
○ None

Click and drag in the plot to zoom X: ☑ Y: □ Zoom Out Print □ Enable tooltip □ Plot mass error

b+	#	Seq	#	y+
236.1144	1	A	9	
337.1621	2	T	8	924.4646
424.1941	3	S	7	823.4169
521.2469	4	P	6	736.3849
658.3058	5	H	5	639.3321
757.3742	6	V	4	502.2732
871.4171	7	N	3	403.2048
985.4600	8	N	2	289.1619
	9	R	1	175.1190

[Click] to move table

Add to N-term: **164.07**

Figure 4.13: Interactive spectrum view of a peptide. The spectrum viewer allows to interactively show the spectra of the peptide. It can mark different ions types, like *y*-, *b*-, *a*-, *z*-ions. Furthermore, it assigns their charge in the view. Furthermore, she spectrum viewer can also mark neutral losses (NH_3, H_2O, and H_3PO_4), immonium and reporter ions. The mass tolerance can be set to either Th or *ppm* and the peak can be assigned to either the most intense peak or the nearest match. The whole plot can be downloaded and printed.

Interactive global search

As mentioned in the description of the navigation menu, we implemented a search field in the header, which is always available. This search allows querying the database for almost all stored categories. However, no category has to be selected to search for any item. This is possible because the database is queried for all categories in parallel and every hit will be reported for each category. The most common search item is the peptide sequence. To search for peptides, you either type in the full peptide sequence or a subsequence. This allows searches for fragments like "SYF", which will report all sequences containing "SYF". To look up a protein, you can either type the UniProt identifier or the gene name. Furthermore, you can search for tissue types (e.g., "brain") and HLA types (e.g., "A*01:01"). The search will report any hit in any of the mentioned categories, which allows a fast and powerful query of the database without knowledge of the structure of the data or the available categories.

4.4 Discussion

The HLA Ligand Atlas presents the largest publicly available benign immunopeptidome dataset. At the time of writing, it contained the immunopeptidome of 6 individuals. We provide the raw data and allow the biologists to access the data via an interactive interface. This interface helps the wet-lab scientist to search the data using their web browser and answers many different immunological questions. These question can vary from a simple peptide sequence search to the identification of tissue-specific peptides. In addition, the interface provides interactive plots and tables that provide a personalized view of the data and can be downloaded for use in presentations. If new questions raised by the users require additional plots and tables, the flexible framework can easily be extended using the tools at hand.

Besides the intuitive interface, the focus was on a fast response time for the queries, which necessitated database optimization. Finally, the web page will be publicly available and all raw data will be provided for downloading and offline presentation and analysis. At the moment, the database contains 6 individuals. However, this is only a first release and the aim is to extend the database with more individuals to cover more tissues and HLA types. Furthermore, the contained data is only from benign samples, but the future aim is to add data from malignant samples, which are also measured in our lab. However, this data from malignant samples contains various tumor types and stages, which would need an extension of the metadata information stored in the database.

The SysteMHC Atlas by Shao et al.[114] provides access to various MHC peptide datasets. However, these datasets are from different studies, that differ in the methods and instruments used. In contrast, the HLA Ligand Atlas contains data from one study that were obtained using homogenous methods and were analyzed on only one instrument. Furthermore, the System MHC Atlas provides only basic possibilities to query the contained data, whereas the HLA Ligand Atlas further statistics provides beside basic queries, such as overviews and an included spectrum viewer. As soon as the raw data is uploaded on PRIDE and publicly accessible, the System MHC Atlas can access this data and rerun the processing using their pipeline. This will help researchers to compare the dataset with others using the System MHC Atlas website and especially access the then extended spectrum library.

In contrast to The Cancer Immunome Atlas (TCIA)[24], the HLA Ligand Atlas contains HLA peptides, which are detected by mass spectrometry and not only predicted based on sequencing data.

We hope to receive feedback on the statistics that are available trough the website, when it goes public. This feedback will help us to provide answers to frequent questions and to implement a solution for them, which will enhance the website and make it even more useful. Especially, as soon as data from malignant sample is added, there will be new tumor-related questions, which we will try to answer with dedicated statistics and analyses. Many of these statistics and visualizations can be adapted from other websites that encompass proteomics data from both benign and malignant samples, like the ProteinAtlas[126] and ProteomicsDB[143].

The HLA Ligand Atlas is optimized for fast queries and a fast response time. However, we cannot fully anticipate the number of requests that will be received. As of yet, only a small number of users has access to the website, causing only minor traffic. Therefore, we will certainly have to optimize parts of the website for larger traffic, but the potential bottlenecks are not yet predictable. This optimization could be done by either avoiding large queries trough adjusting interface content or by optimizing the underlying MySQL database scheme. Another option could be installing the web server and the MySQL server on dedicated or distributed servers to allow parallelized access and queries.

After the presentation of the HLA Ligand Atlas, the next chapter will provide an overview of the contained data, the processing pipeline, and an extended statistical analysis of the data in the HLA Ligand Atlas.

Chapter 5

Analysis of the benign tissue immunopeptidome

5.1 Introduction

In the last decade many immunopeptidome analyses and datasets have been published[15,62,67,71,135]. However, most of these are either focused on malignant tissue[15,62,67,135] or cell lines[71]. If these analyses contain normal or healthy tissue, in most of the cases, they are obtained from the same patient as the tumor sample (adjacent benign tissue). Although the immunopeptidome should be different to the malignant tissue, it might be affected by the disease or even might partly contain malignant tissue. This creates a problem when these benign samples are used as a negative data set to calculate the difference between the malignant and the benign tissue, as the malignant contamination or influence can alter the immunopeptidome leading to incorrect conclusions. This is especially problematic, if possible tumor targets are excluded because they were found also on adjacent benign tissue, which was influenced by the nearby tumor.

In the field of proteomics[126,143] and transcriptomics[72], the comparison of different benign tissue types is a frequent analysis and large data sets are publicly available. These data sets allow to find similarities and differences between the different tissues types. Furthermore, if multiple samples of the same individual are obtained, the variability between the tissues within one person can be assessed. However, in the field of immunopeptidome data, there is no such data set. We try to fill this gap with data provided here, which are also accessible in the HLA Ligand Atlas.

Most of the immunopeptidome studies are focused on finding new targets for cancer therapies. Hence, a detailed description of the properties of the immunopeptidome in general is not available. Other non-cancer related studies only try to define the binding motifs of the different HLA alleles[16] and are often focused on cell lines[33]. We therefore try to present a large-scale analysis of the immunopeptidome of different tissue types and individuals. Within this theses, we present an analysis of the intra- and inter-individual variability, the differences between various tissue types of one individual, and the variance between the same tissue type of different individuals. Furthermore, we try to define more general properties of the immunopeptidome, such as the length distribution of HLA ligands or the overlap of class I and II proteins. Another interesting aspect we present is the analysis of different length variations of one peptide and how frequent this event is.

As mentioned above, there are many different studies on the immunopeptidome. However, most of these studies do not publish their raw data, but at most lists of peptides. In contrast, we here will publish all raw data of all MS runs. We hope that with this large dataset at hand, new tumor targets can be discovered and new computational methods can be developed using it as training data set.

5.2 Material and Methods

5.2.1 Sample acquisition and measurement

All samples were obtained from autopsies at the University Hospital Zurich. Between death and autopsy, the bodies were stored in a cold room at 4 °C. After obtaining them, the samples were snap frozen until immunoprecipitation. In the immunoprecipitation, the pan-class I HLA antibody W6/32 and the class II antibodies Tue39 (HLA-DR, HLA-DP, and HLA-DQ) and L243 (HLA-DR) were used. A simplified overview of the immunoprecipitation is shown in Figure 2.5. Afterwards, the samples were measured in the HPLC coupled tandem mass spectrometer. All samples, except for the time-series experiment, were measured on an Orbitrap Lumos. The time-series experiment was conducted on an Orbitrap XL coupled with an HPLC.

The experiments on the Orbitrap Lumos were done using the Nano Trap Collumn C18 $75\mu m$ x $2cm$ and PepMap C18 $50\mu m$ x $250mm$ NV FS column, both by Fisher Scientific, in the HPLC. The column used in combination with the Orbitrap XL were the same. The Orbitrap Lumos recorded all spectra with a fragment mass tolerance of 0.02 Da and precursor mass tolerance of 5 ppm. The Orbitrap XL recored with a fragment mass tolerance of 0.5 Da and precursor mass tolerance of 5 ppm. The spectra of the Orbitrap Lumos were next centroided by Orbitrap Fusion Lumos Tune Application (version 2.1.1565.23) by Thermo Scientific.

5.2.2 Sample processing

Efficient data processing is crucial for the development of a large database like the HLA Ligand Atlas. This section provides an overview of the whole processing workflow, which is also shown in Figure 5.1.

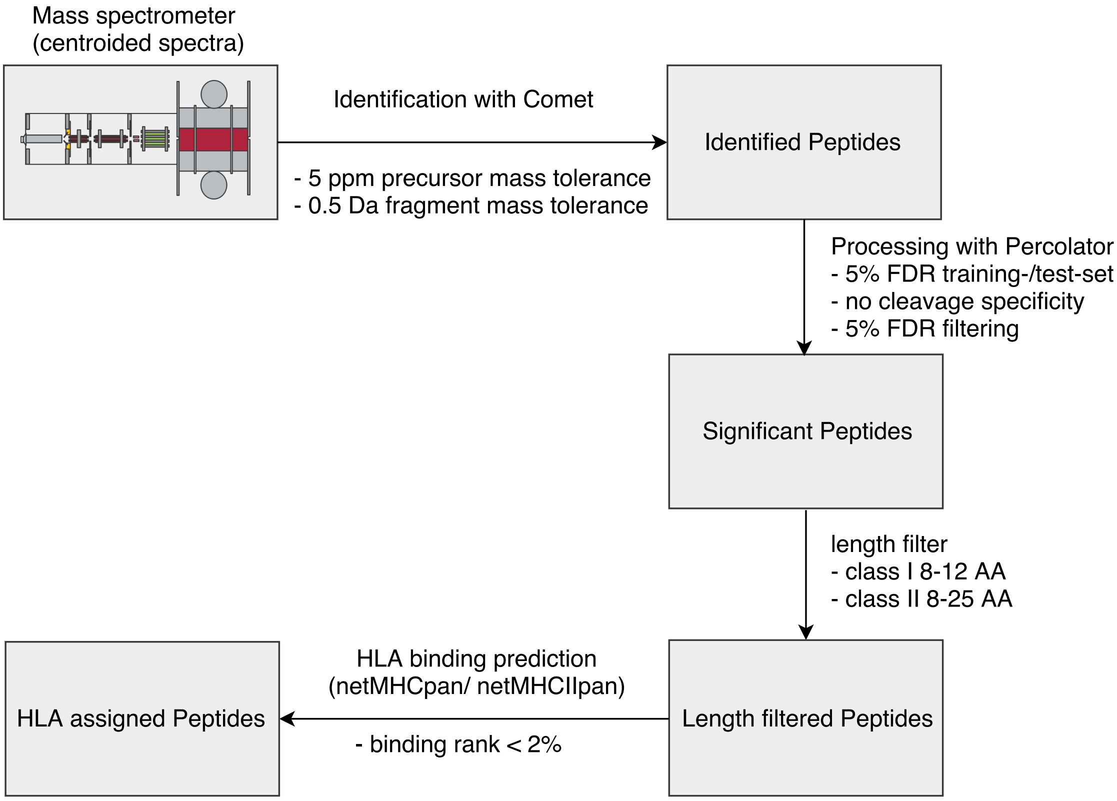


Figure 5.1: Processing workflow of the HLA Ligand Atlas, including filter criteria. After measuring the samples in the mass spectrometer, we used the identification software Comet with 5 ppm precursor mass tolerance and 0.02 Da fragment mass tolerance to identify peptides. Then we processed the identified peptides with Percolator, using a 5% FDR for the test- and training-set, and no cleavage specificity. We used only hits with an FDR below 5%. We filtered the identified peptides by length for their corresponding HLA class (Class I: 8-12 amino acids, Class II: 8-25 amino acids). Then, we predicted binding affinities for the length-filtered peptides with netMHCpan-3.0 or netMHCIIpan-3.1. We considered peptides to be binders if their binding rank was below 2%.

After the data was recorded in the mass spectrometer, it had to be processed to identify the peptides. To process the data set, we used Comet (version 2017.01 rev. 2), with the precursor mass tolerance (5 ppm) and fragment mass tolerance (0.02 Da) described above. Furthermore, we allowed methionine oxidation as variable modification. We used Comet to search the UniProt[124] reviewed human proteome version September 2013 (former SwissProt) for peptide identification. Comet is described in detail in Subsection 3.1.2 and is available at `http://comet-ms.sourceforge.net/`. Next, we processed the results of Comet with Percolator (version 3.1) available at `http://percolator.ms/`. We used a 5% FDR for the training and test set and set no enzyme specificity. Finally, we filtered the results of Comet+Percolator with an FDR of 5% and a length restriction of 8-12 amino acids for HLA class I and of 8-25 amino acids for HLA class II. We used OpenMS (version 2.1), to simplify file handling between the different tools.

To create benchmarks for Comet and Sequest HT, we processed samples with the following combinations: Comet, Comet + Percolator, Sequest HT, and Sequest HT + Percolator (Percolator version 2.05 + tryptic digestion). After the identification of the peptides, we assigned them to their corresponding HLA molecule, based on the individual's HLA type and the binding affinity, which we predicted with netMHCpan-3.0 and netMHCIIpan-3.1 (see Section 3.3). We considered a peptide to be a binder if it has a binding rank below 2%.

5.2.3 Heat map and hierarchical clustering

All distance heat maps were created using the gplots package[140] in R. The distance was calculated as the Jaccard index between each pair:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{5.1}$$

where A and B are sample sets containing either peptides or proteins. The hierarchical clustering algorithm used complete linkage and a Euclidean distance.

5.3 Results

5.3.1 Available data

For the HLA Ligand Atlas, we collected tissue samples from six different individuals. From each individual we obtained multiple different tissues and analyzed the immunopeptidome. Table 5.1 gives an overview of the collected samples. For each sample, we measured the HLA class I and II ligands in five replicates. Three of these were recorded with Data-Dependent Acquisition (DDA), the other two with Data Independent Acquisition (DIA). We integrated only the three DDA replicates into the HLA Ligand Atlas and therefore only these are accessible via the web interface. However, we will make the raw data of all five replicates publicly available trough the PRoteomics IDEntifications (PRIDE) database as soon as the HLA Ligand Atlas is published. In addition to the raw data via PRIDE, we will make the whole database, including all identified peptides, available for download from the HLA Ligand Atlas web page.

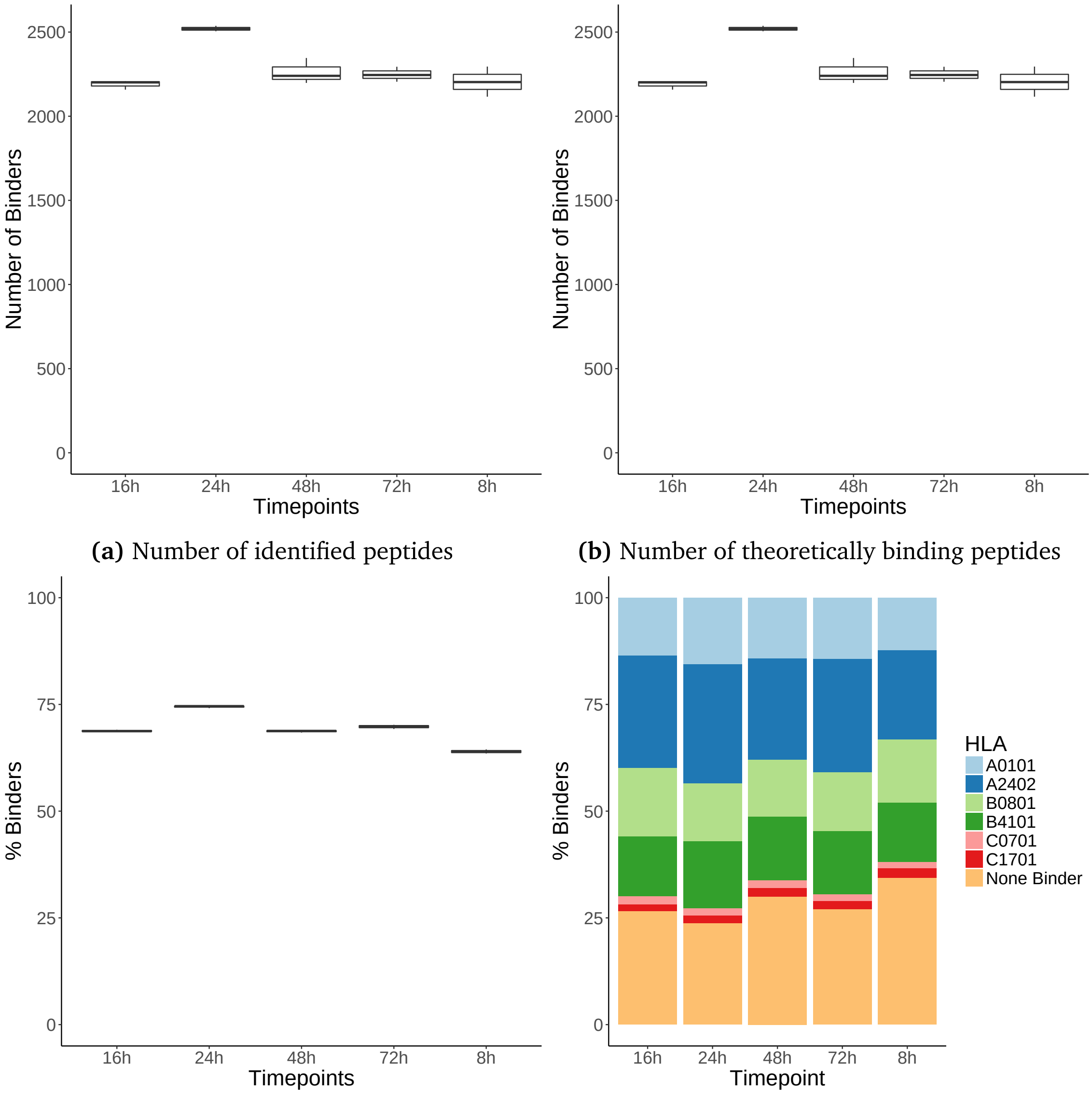
5.3.2 Time-series experiments

The samples for the HLA Ligand Atlas were obtained during autopsies. However, autopsies are often not conducted right after a person's death, but for example, on the next day if an individual dies during the night. Typically, the body is stored at 4 °C until the autopsy is

Table 5.1: Summary of all individuals, tissues, and the corresponding number of samples contained in the HLA Ligand Atlas.

Tissue	ZH02 (f)	ZH05 (m)	ZH06 (f)	ZH08 (m)	ZH09 (m)	ZH13 (m)	Total
Adrenal Gland	0	0	1	1	0	0	2
Aorta	0	0	1	1	0	1	3
Bladder	1	0	1	1	0	1	4
Bone Marrow	0	0	1	1	1	1	4
Brain	1	0	1	1	1	1	5
Cerebellum	1	0	1	1	1	1	5
Colon	0	0	1	1	0	1	3
Esophagus	0	0	1	1	1	1	4
Heart	0	1	1	1	1	1	5
Kidney	0	1	0	1	1	1	4
Liver	0	1	1	1	1	1	5
Lung	1	1	1	1	0	1	5
Lymph node	0	0	1	1	1	1	4
Muscle	0	0	1	1	1	0	3
Myelon	1	0	0	0	0	0	1
Pancreas	0	0	0	0	0	1	1
Skin	0	1	1	1	0	1	4
Small intestine	1	0	1	1	0	1	4
Duodenum	0	0	0	0	0	1	1
Skin	0	0	0	0	1	0	2
Spleen	0	0	1	1	1	1	3
Stomach	0	0	0	1	0	1	2
Testis	0	0	0	0	1	0	2
Thyroid	1	1	1	1	1	1	6
Tongue	0	0	1	1	0	0	2
Trachea	0	0	1	0	0	1	2
Total	7	6	19	20	13	20	85

conducted and the samples can be taken. Before we could collect and measure the samples, we evaluated the time-dependent degradation of the immunopeptidome. To this end, we carried out a time-series analysis as follows. A liver sample was taken at an autopsy in the morning, to minimize the time between death and sample extraction, and stored at 4 °C. After 8, 16, 24, 45, and 72 hours, a small part of the sample was measured using immunoprecipitation and MS. The number of identified peptides was constant over all time points and both HLA classes (Figure 5.2a and 5.3a). Furthermore, the amount of binding peptides predicted with netMHCpan/netMHCIIpan did not change (Figure 5.2b and 5.3b). In conclusion, the purity, which is the defined as the number of binders divided by the total number of identified peptides, was constant over time (Figure 5.2c and 5.3c). For all comparisons not tests were performed, as technical instead of experimental replicates were conducted. In addition, we evaluated whether the distribution of peptides belonging to different HLA types changed, but found no changes in the the composition of either HLA class I or HLA class II (Figure 5.2d and 5.3d).



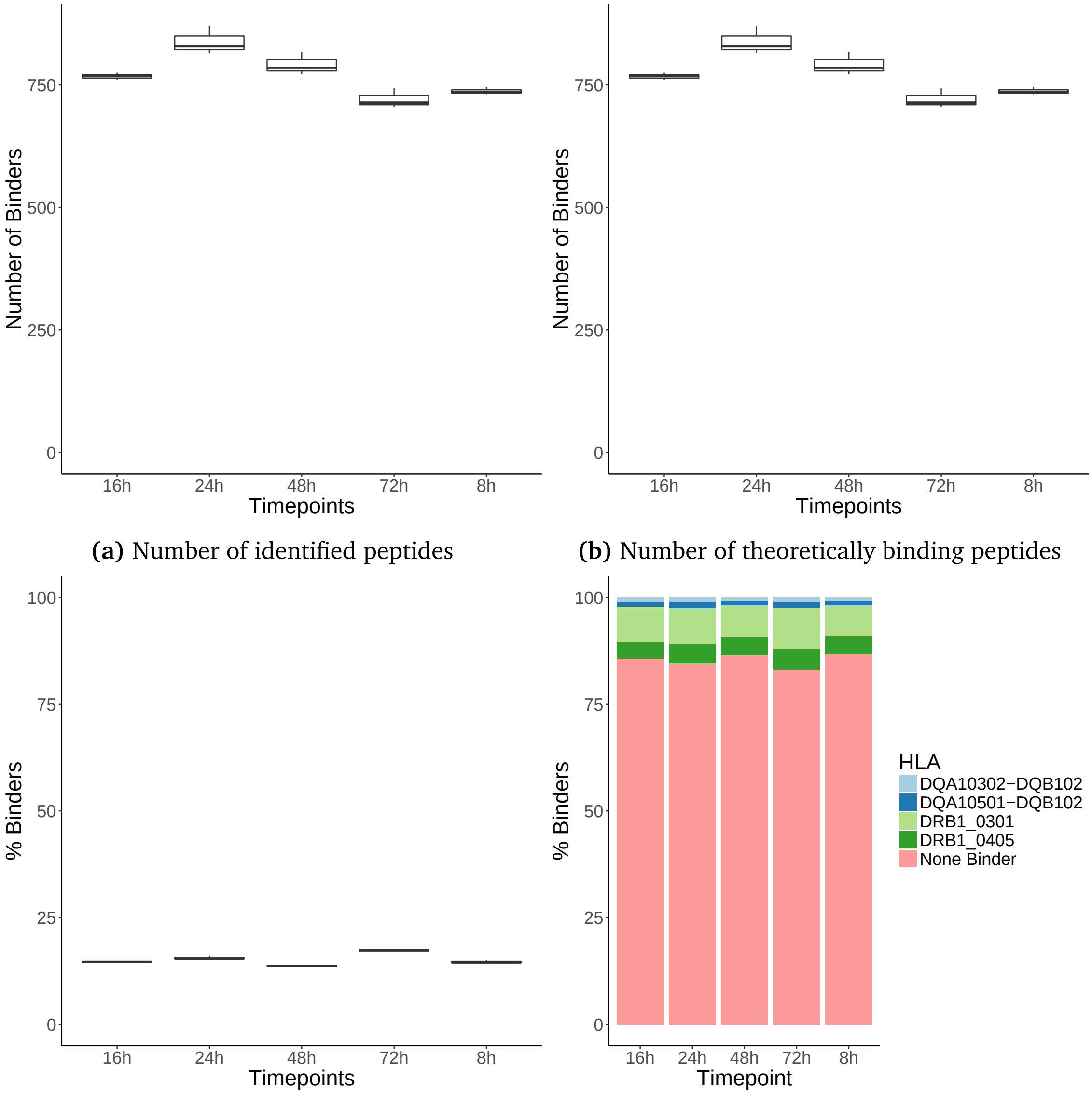
(a) Number of identified peptides

(b) Number of theoretically binding peptides

(c) Purity of runs

(d) HLA allotype distribution

Figure 5.2: Time-dependent degradation of the immunopeptidome for HLA class I. For each time point, three technical replicates (MS runs) were measured. We used netMHCpan-3.0 to assign peptides to their corresponding HLA type.



(a) Number of identified peptides

(b) Number of theoretically binding peptides

(c) Purity of runs

(d) HLA allotype distribution

Figure 5.3: Time-dependent degradation of the immunopeptidome for HLA class II. For each time point, three technical replicates (MS runs) were measured. We used netMHCIIpan-3.1 to assign peptides to their corresponding HLA type.

Based on the results of the time-series experiment, we included only samples that were obtained less than 72 hours after death as we cannot be sure no changes occur after this time frame. The first QC step was part of the sample collection, as we only collected healthy tissues, which in our case means that we included only benign tissue. Furthermore, we did not acquire samples from individuals who died from diseases affecting most of the body (e.g., sepsis).

5.3.3 Quality control

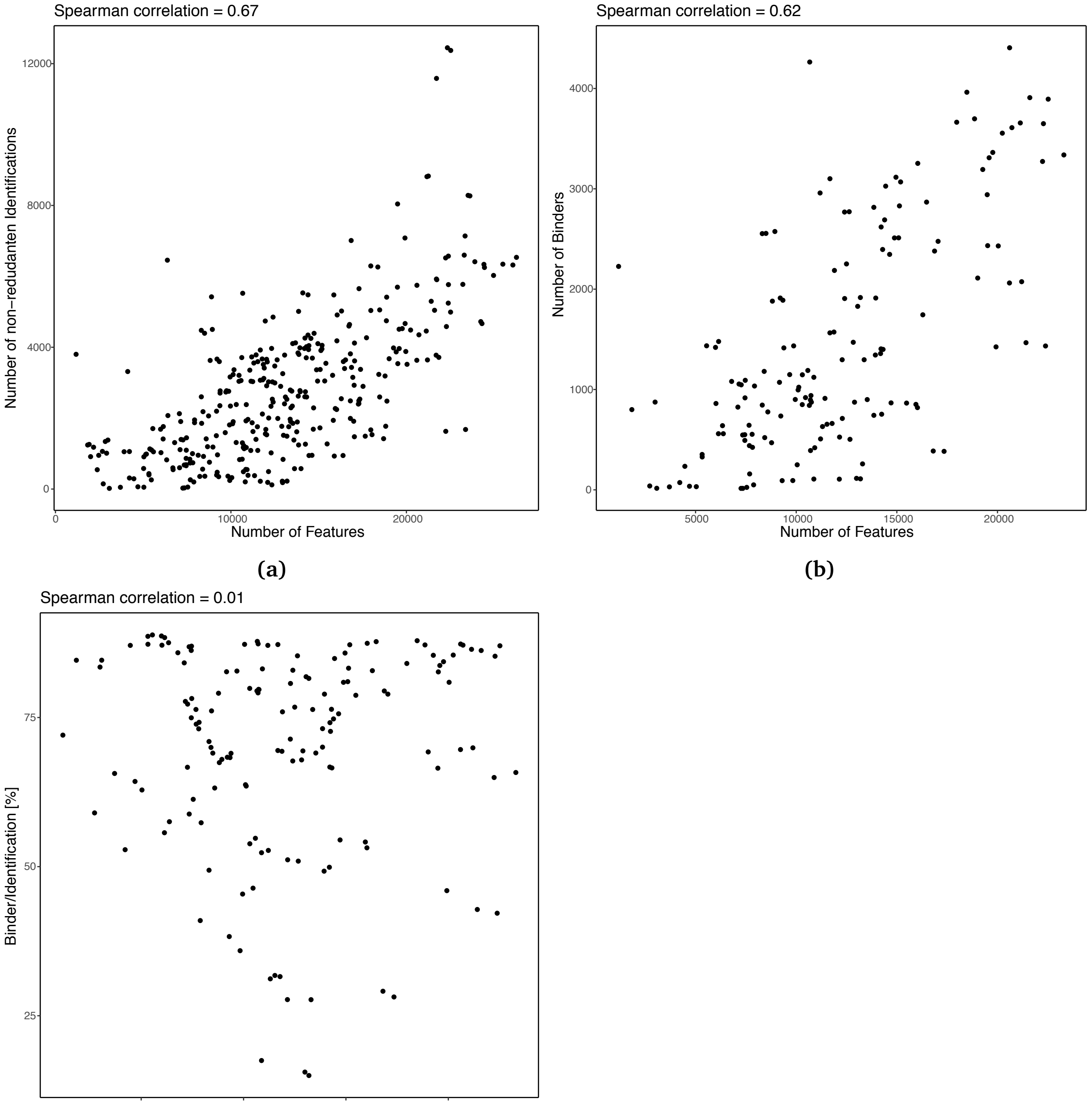
A database is only as good as the quality of the contained data. Therefore, quality control (QC) was crucial for the creation of the HLA Ligand Atlas and its data. We performed multiple QC steps during the data acquisition. First we did an QC on MS run level, comparing different properties like the number of features or identifications using Spearman correlation and hierarchical clustering. The second part is the study QC. It consists of the detailed analysis of the immunopeptidome of the stomach and the definition of contaminants in the immunoprecipitation.

Quality control of MS runs

Comet provides numerous properties that can be used to assess the quality of an MS run, after the mass spectrometry runs were processed. We used these properties to identify differences between the runs. The most valuable metrics for an MS run are the number of identifications and the number of identified features. These two metrics and their ratio can indicate several problems. For example, a high number of features and a low number of identifications may indicate problems during peptide identification or the presence of contaminants in the sample.

We first calculated the Spearman correlation between the number of features and the number of identifications (Figure 5.4a). We found a mediocre correlation ($R = 0.67$). This was expected because more IDs can be identified in runs with more features. However, not all contained features can be associated with a peptide. Mainly caused by the variance in the spectrum quality. We also calculated the Spearman correlation between the number of features and the number of binders and between the number of features and the ratio of binders to identifications (Figure 5.4b and 5.4c). Whereas the number of binders correlates with the number of features, we found no correlation between the number of features and the ratio of binders to identification. Therefore, we concluded that the number of features is associated with the number of binders but not with the ratio of binders to identifications. However, the mediocre Spearman correlation between the number of features and the number of identifications, the number of binders, and the binder-identification ration, indicates that the number of feature only is a weak criteria for the QC of MS runs. Therefore, we used 12 properties to examine the quality of our MS runs (Table 5.2). These features were extracted from the idXML files using the qcExtractor tool included in OpenMS[136]. It generated qcML files containing the necessary information, which could be used to ensure the quality of the MS run. For 101 of the 435 runs, we were not able to extract all 12 features. These were discarded from further QC analyses.

After the extraction of the run properties, we calculated Euclidean pairwise distances for the scaled property values. We used column-based z-score scaling to normalize the range of the different properties. Figure 5.5 illustrates the calculated pairwise distances in a heat map.



(a)

(b)

(c)

Figure 5.4: Spearman correlation between (a) number of features and number of identifications, (b) number of features and number of binders, and (c) number of features and percentage of binders in identifications for HLA class I. We used netMHCpan-3.0 to predict binding.

The heat map shows that most of the runs are very similar (red) and only few runs are different (yellow). To identify these specific runs, we obtained the top 10 runs with the largest mean distance (Table 5.3). The top 10 runs are mostly from samples of stomach tissue, as well as one spleen and two skin tissue samples. All these runs have above average numbers of identification and features. Furthermore, most of their features could be assigned to identifications. Because most of the highly-different runs originate from stomach samples, we analyzed these in detail in the next section.

Table 5.2: Mass spectrometry run properties used to calculate the distance matrix.

Property
number of identifications with charge 1
number of identifications with charge 2
number of identifications with charge 3
number of identifications with charge 4
number of identifications
number of non-redundant identifications
number of protein hits
number of features
total ion current in features
number of peptide identifications without an assigned feature

Table 5.3: The ten mass spectrometry runs with the largest mean Euclidean distance, including the main properties. All results are computed without FDR filtering.

Run	Mean distance	Identifications	Non-redundant identifications	Protein hits	Features
ZH09_Stomach_class2_#2	9.86	20,267	12,449	2,529	22,339
ZH09_Stomach_class2_#3	9.57	20,503	12,374	2,592	22,528
ZH09_Stomach_class2_#1	9.44	19,909	11,585	2,509	21,717
ZH09_Stomach_class1_#1	7.20	15,974	8,043	2,675	19,500
ZH09_Stomach_class1_#2	6.90	16,029	8,814	2,746	21,172
ZH06_Spleen_class2_#2	6.88	13,982	7,140	1,890	23,346
ZH09_Stomach_class1_#3	6.85	15,923	8,830	2,735	21,253
ZH09_Skin_class2_#2	6.54	11,968	6,537	2,078	26,275
ZH09_Skin_class2_#3	6.47	11,804	6,321	1,988	26,071
ZH13_Stomach_class2_#3	6.42	15,013	8,285	1,808	23,518
Mean (all runs)	4.05	5,881	2,784	1,401	12,971
Standard deviation (all runs)	0.97	3,870	2,042	910	5,396

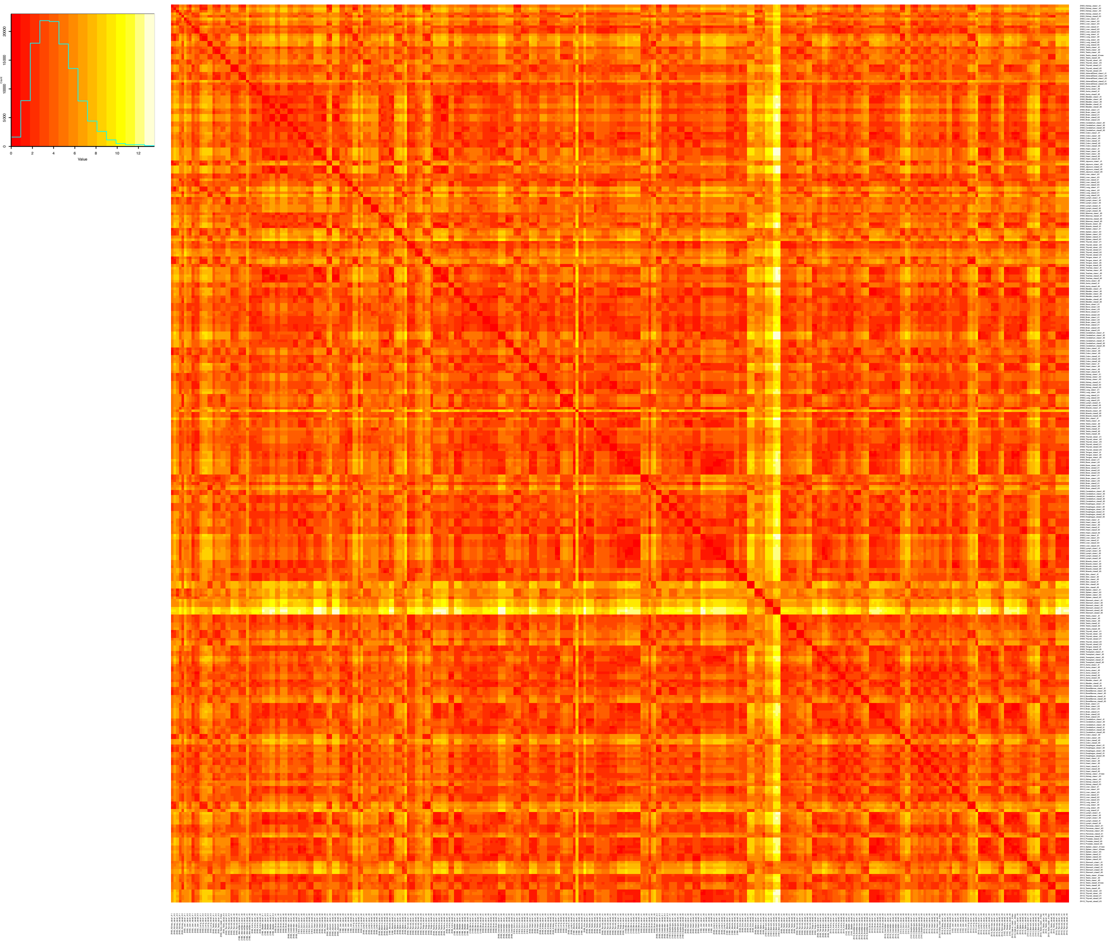


Figure 5.5: Pairwise distance of 346 MS runs in the HLA Ligand Atlas. We calculated the pairwise Euclidean distance with a z-score scaled property vector. The properties are: number of identifications with charge 1 to 4, number of identifications, number of non-redundant identifications, number of protein hits, number of features, total ion current in features, number of peptide identifications without an assigned features. The heat map shows that most of the runs are very similar (red) and only few runs are different (yellow). To identify these specific runs, we obtained the top 10 runs with the largest mean distance (Table 5.3).

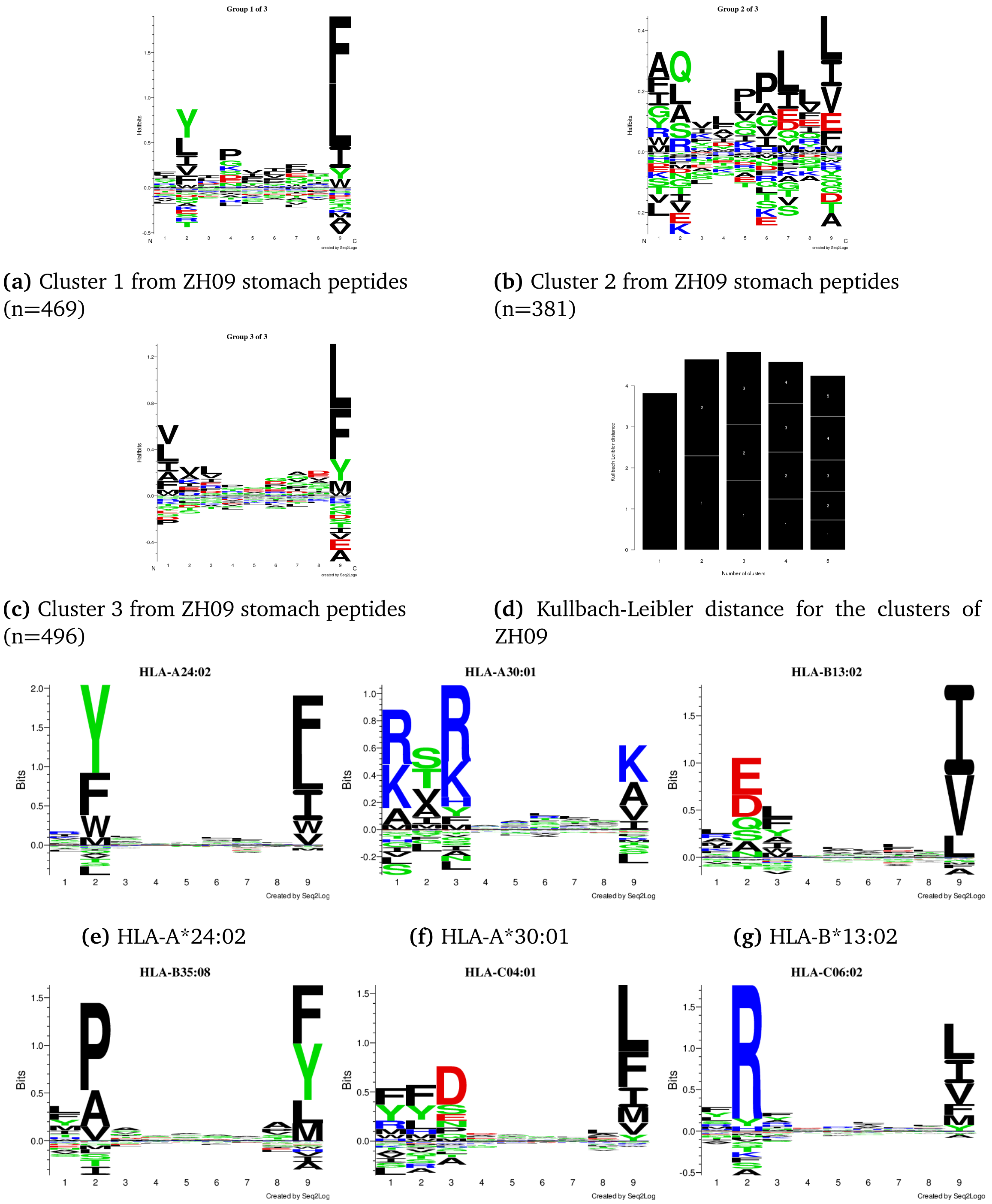
5.3.4 Study quality control

The immunopeptidome of the stomach

In the previous section, we performed a pairwise distance calculations of the MS runs, in which the stomach samples showed a high distance to the other MS runs. Therefore, we will have a closer look into their properties and the identified peptides. The analysis of the HLA class I immunopeptidome of the two HLA class I stomach samples yielded large amounts of peptides (ZH09: 9783, ZH13: 6956). However, the number of peptides that bind to the individual's HLA type (predicted by netMHCpan-3.0) was very low (ZH09: 1916 binders /19.6%, ZH13: 1451 binders /20.0%). Furthermore, the average number of features and the pair-wise distances were larger than for the other samples. Therefore, we analyzed the identified peptides with the Gibbs-clustering algorithm to assign them to their sequence motifs. The optimal number of clusters was two for ZH13 (Figure 5.7) and three for ZH09 (Figure 5.6). Position 9 was the most important in the motifs of all resulting clusters. All clusters had either leucine (L), isoleucine (I) or phenylalanine (F) at position 9. Next, we compared the cluster motifs with the motifs of the individual's HLA type (Figure 5.6 and 5.7). The motifs of ZH09 showed a possible overlap of the motifs of A*24:02 and cluster 1. However, cluster 1 has only a week signal for tyrosine (T) and phenylalanine in position two. Furthermore, B*13:02, B:35:08, C*04:01, and C*06:02 have either leucine or phenylalanine at the anchor at position 9, but none of their first anchors at position 2 can be found in any of the clusters of ZH09. The comparison of the clusters of ZH13 and its HLA type showed a weak overlap at anchor position 9 for HLA A*02:05, A*11:01, B*58:01, C*02:02, and C*07:02. Again this overlap is either leucine or phenylalanine. As for ZH09, no additional anchor at position 2 can be found. The identification of the clusters of ZH09 and ZH13 reveals an overlap of the amino acid frequency at position 9. Both leucine or phenylalanine can be found at this position.

The comparison of the HLA binding motifs and the clusters indicated that the measured peptides originated from sources other than HLA or were processed before, during, or after the immunoprecipitation. After a closer look at the stomach environment, we suspected the cause to be pepsin. Pepsin cleaves phenylalanine, tyrosine, tryptophan, and leucine in either the P1 or P1', where P1 is the C-terminus and P1' the N-terminus of the resulting peptides[61]. However, at pH 1.3 pepsin is more specific and cleaves phenylalanine and leucine in position P1. The cleavage specificity at pH 1.3 matches the sequence motifs of the stomach peptides.

Although we could explain the genesis of our peptides, we could not determine when the digestion occurs. It could either happen in the living individual, after death, or during the immunoprecipitation. However, if the digestion happened in the living individual or after its death, that would mean that pepsin can digest HLA peptides while they are bound to the HLA, which is unlikely. Therefore, we assume that the digestion occurred during immunoprecipitation, although most of the time the immunoprecipitation is conducted at 4 °C and protease inhibitors are added during the preparation of the cell lysate. The used protease inhibitor is the cOmplete Protease Inhibitor by Roche, which inhibits serine, cysteine, and acidic proteases. Although pepsin is a aspartatic protease in cleaves bet in acidic solutions, it might not be specific enough to inhibit pepsin. In addition, the peptides could also be normal proteins digested by pepsin in the organism. If they are present in high concentrations, they might not be filtered out during the immunoprecipitation, which would mean they are not HLA-related.



(a) Cluster 1 from ZH09 stomach peptides (n=469)

(b) Cluster 2 from ZH09 stomach peptides (n=381)

(c) Cluster 3 from ZH09 stomach peptides (n=496)

(d) Kullbach-Leibler distance for the clusters of ZH09

(e) HLA-A*24:02

(f) HLA-A*30:01

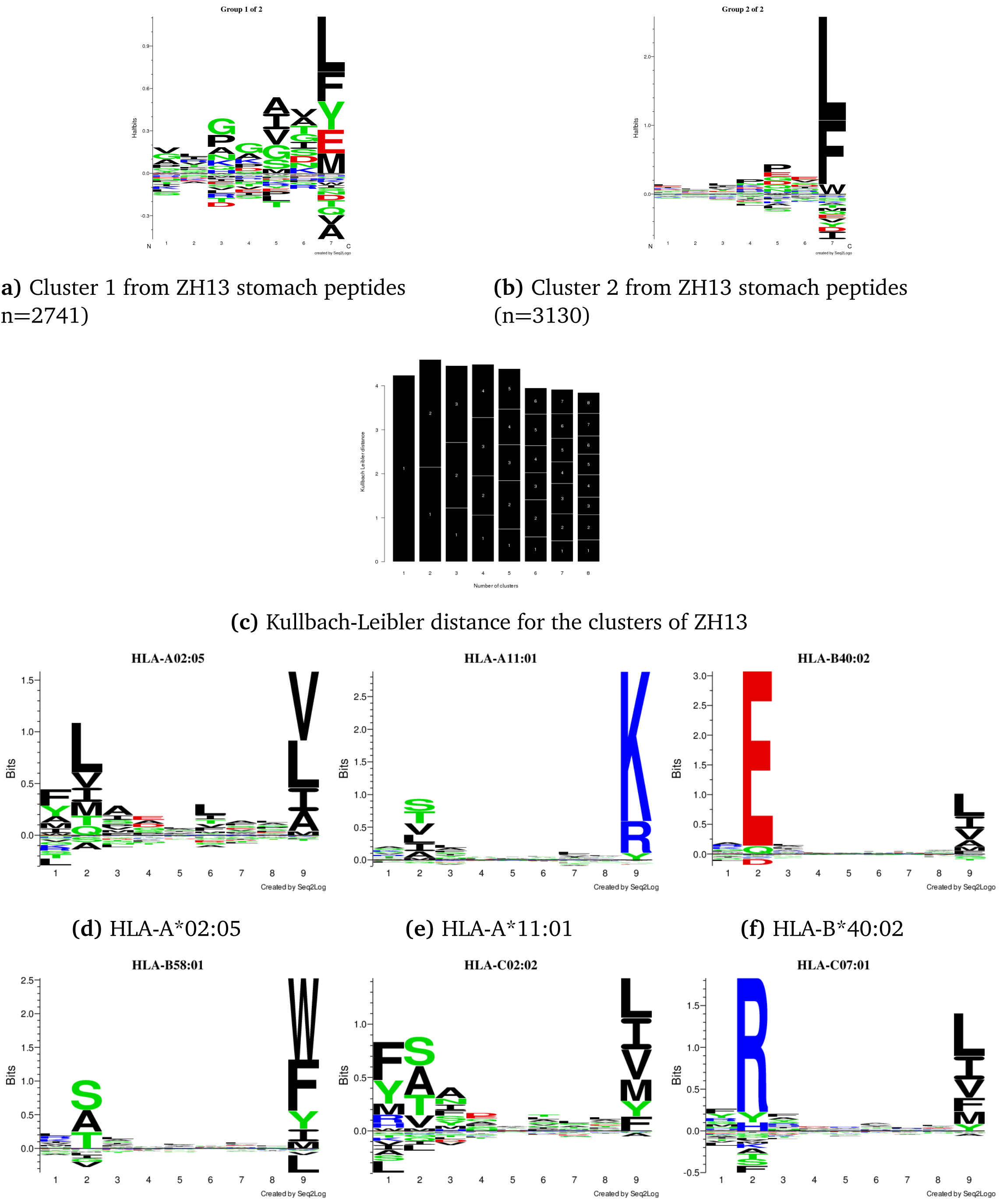
(g) HLA-B*13:02

(h) HLA-B*35:08

(i) HLA-C*04:01

(j) HLA-C*06:02

Figure 5.6: Gibbs clustering of the peptides from the stomach sample ZH09. (a)-(c) show the motifs of the three clusters of the Gibb clustering. (d) shows the Kullbach-Leibler distance for the clusters, resulting in an optimal number of cluster of three. (e)-(j) Sequence motifs of the HLA alleles of ZH09. These public sequence motifs have been obtained from `http://www.cbs.dtu.dk/services/NetMHCpan/logos.php`[89] (accessed 04.05.2017).



(a) Cluster 1 from ZH13 stomach peptides (n=2741)

(b) Cluster 2 from ZH13 stomach peptides (n=3130)

(c) Kullbach-Leibler distance for the clusters of ZH13

(d) HLA-A*02:05

(e) HLA-A*11:01

(f) HLA-B*40:02

(g) HLA-B*58:01

(h) HLA-C*02:02

(i) HLA-C*07:01

Figure 5.7: Gibbs clustering of the peptides from the stomach samples ZH13. (a)-(b) show the motifs of the three clusters of the Gibb clustering. (c) shows the Kullbach-Leibler distance for the clusters, resulting in an optimal number of cluster of two. (d)-(i) Sequence motifs of the HLA alleles of ZH13. These public sequence motifs have been obtained from `http://www.cbs.dtu.dk/services/NetMHCpan/logos.php`[89] (accessed 04.05.2017).

For the HLA Ligand Atlas, we measured only two stomach samples. Further experiments with pepsin specific inhibitors are necessary to find the time point of digestion and to investigate if pepsin cleavage occurs *in vivo* or during immunoprecipitation. Because most of the identified peptides do not match any HLA binding motif, we excluded both stomach samples from the further analysis.

Contaminants in the immunoprecipitation analysis

After removing the two stomach samples from the dataset, we looked for contaminations that might occur in immunoprecipitation samples. Mostly expert knowledge is used to identify contaminants. However, this approach may be biased and is not feasible for large datasets like the HLA Ligand Atlas. Therefore, we tried to distinguish the contaminations with an unbiased approach that uses multiple different features which describe peptide and protein properties.

We used all 68,845 HLA class I peptides to find the best method for the definition of the contaminants. We calculated the receiver operating characteristic (ROC) curve and its area under the curve (AUC) for each, to compare the methods. In addition, the peptides were classified into non-binding and binding peptides with netMHCpan-3.0 to calculate the false positive rate (FPR) and true positive rate (TPR), to calculate the ROC and AUC.

The first approach that we tested used the occurrence of peptides across all 216 HLA class I runs. Because we analyzed six different individuals with a very small HLA type overlap (for the individual's HLA type, see Supplementary Figure D.1 and D.2), we expect that only few peptides would be shared between the individuals. We found 53 peptides that were detected in more than 50% (108) of the runs (Supplementary Table D.3). These peptides had 96 distinct source proteins, the most common ones (occurrence > 5) belonging to the gene families of actin, hemoglobin and histones. Especially actin is a well known contaminant in mass spectrometry and is brought into the experiment by the lab scientist. All three genes are listed in the CRAPome database[83] (`http://www.crapome.org/`) and are also frequent in other mass spectrometry experiments. Since our experiment is not a standard proteomics analysis, we also find peptides belonging to HLA-A,-B, and -C, which is caused by our experimental setup. Next, we benchmarked the peptide cut off. The calculation of the TRP and FPR including the ROC curve and the AUC resulted in an AUC of only 0.41 (Figure 5.8). Therefore, we discarded this approach.

In the second approach used, we changed the focus from peptide to protein level. This also allows the translation to HLA class II, for which no valid peptide binding predictors are available. The first protein feature level was the number of runs in which the protein was found. Again, we observed a poor AUC of 0.47. Next, we tried protein coverage, which is commonly used in MS and proteomics experiments. The protein coverage describes how many amino acids of a protein are found in an experiment normalized, by the length of the protein.

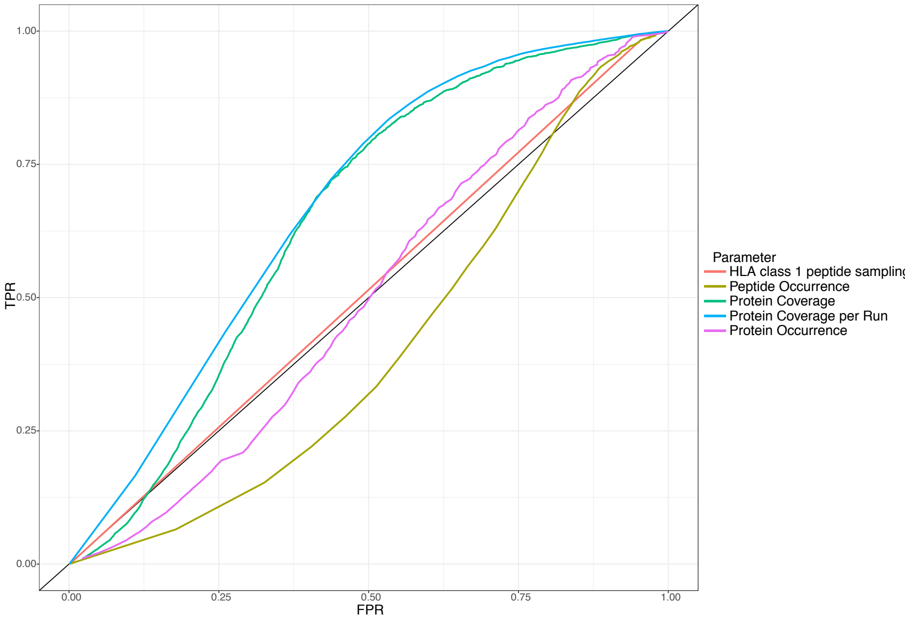


Figure 5.8: Receiver operating characteristic (ROC) curve for different features. The AUCs for the features are: HLA class I peptide sampling density = 0.51, Peptide Occurrence = 0.41, Protein Coverage = 0.65, Protein Coverage per Run = 0.67, Protein Occurrence = 0.47.

This coverage can be either calculated per run or for all experiments combined. We calculated the AUC for both methods and observed an AUC of 0.67 for the coverage per run and 0.65 for all experiments combined. Furthermore, we computed the HLA class I peptide sampling density for each protein as defined by Bassani-Sternberg et al.[17]. This yielded a poor AUC of 0.51. We also tried to combine properties such as protein coverage and protein occurrence, but these combinations did not yield better AUCs. Based on these results, we decided to use the protein coverage to assess contamination. This has the added benefit that it can be applied to any new immunopeptidome experiments.

After having defined a way to assess contamination, we had to determine the coverage threshold that marks proteins as contaminants. Removing FN proteins is crucial in our analysis, because they could be of biological significance. Hence, we allowed only 1% of FNs and set the coverage threshold to >40% (Figure 5.9). We applied the filter and removed a total of 1581 proteins in the 216 runs (136 unique proteins, see Supplementary Table D.4). Many of these contaminants are well-known contaminants in proteomics (e.g., actin), others, such as the different hemoglobin proteins occur in most of our immunoprecipitation experiments and are mainly represented by non-binding peptides. In addition, many of these contaminants are listed in the CRAPome database and are common in MS experiments. However, like in the peptide contaminant definition, we found HLA-A, -B, and -C in our contaminants list, which are specific for immunopeptidome experiments.

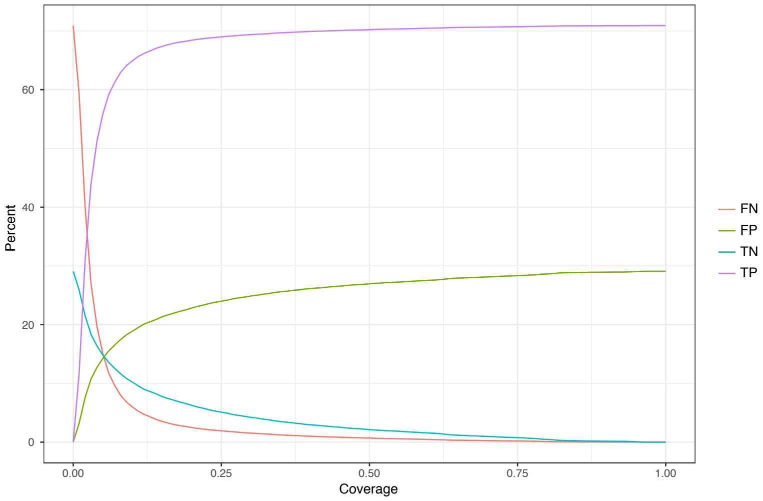


Figure 5.9: Percent of FN, FP, TN, and TP dependend on the protein coverage per run.

Taken together, we flagged 136 proteins as contaminants in our immunopeptidome experiments based on AUCs.

5.3.5 Benchmark of the identification algorithms: Sequest HT vs. Comet + Percolator

The MS raw data can be processed with commercial programs(e.g., the Proteome Discoverer by Thermo Fisher) or with open-source toolboxes (e.g., OpenMS[106]). Both types of software have advantages and disadvantages. Proteome Discoverer is specialized on processing raw files from mass spectrometers of the same distributor, which allows user-friendly analysis. On the other hand, the whole software is a black box, meaning that neither the source code is available nor a full access to the integrated algorithms is possible. In contrast, open-source software like OpenMS allows full access to the algorithm and supports more than one MS manufacturer. However, OpenMS only has a limited user interface, which makes setting up a processing pipeline for the data more difficult. On the other hand, due to the free source code, almost every parameter and algorithm can be optimized for the specifc machine and experimental set-up. Here, we provide a benchmark of four identification pipelines (Figure 5.10). We implemented the first two in the Proteome Discoverer 1.4 and used either Sequest HT or Sequest HT + Percolator (version 2.04)[123]. We used OpenMS to implement the other two, which use either Comet[38,39] or Comet + Percolator (Comet: 2017.01 rev. 2; Percolator: 3.1). Only runs in which Comet identified more than 300 peptides were used in the benchmark (122 HLA class I/ 62 class II runs). Excluding runs with very few peptides removes outliers, which would cause large standard deviation. The whole benchmark was like in all analyses in this thesis performed with an run level FDR of 5%. Our benchmark compares first the number of identified HLA class I and II peptides per sample, and second the fraction of binding HLA class I peptides and the number of identified peptides (binding predicted with netMHCpan-3.0). These two parameters provide a good estimate for assessing which identification pipeline and algorithm performs best.

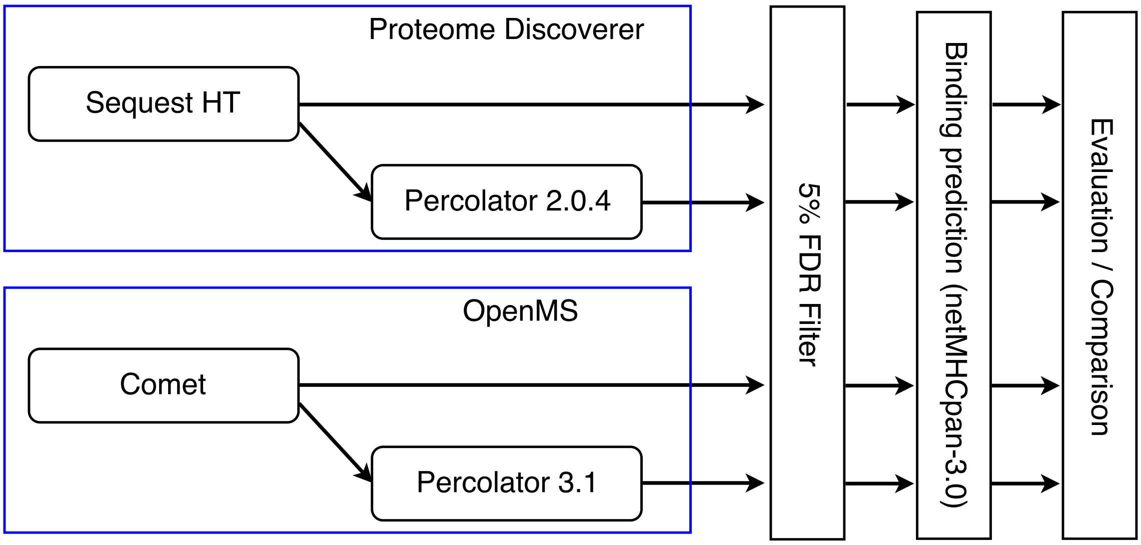


Figure 5.10: Workflow of the benchmark of the identification algorithms. First, Comet and Sequest HT are used for identification. Next, either the data was directly FDR filtered or advanced identification statics calculated using Percolator. After filtering, the binding affinity is predicted using netMHCpan-3.0. Last, the data is evaluated and compared.

Figure 5.11 shows an overview of the number of identified HLA class I and II peptides per sample. The results indicate that the choice of the identification software has a large influence on the number of identifications. The median number of identified peptides is two times larger with Comet than with Sequest HT. Furthermore, we were interested in the potential gain of identified peptides when Percolator is used. Therefore, we used Comet and Sequest HT with and without Percolator. Figure 5.12 shows a median gain of 50% more peptides if Percolator is used. The comparison for HLA class II also shows a gain in identified peptides by Comet. However, the median gain is only 35% without Percolator and 12.5% with Percolator. We did not perform any binding prediction for HLA class II, because no good peptide binding predictors are available.

With Sequest HT we hat to use an old version of Percolator (2.04) because no newer version is available for the Proteome Discoverer. With Comet, we used Percolator 3.1. In the Comet + Percolator setting, we used OpenMS and set the digestion enzyme to *none*. Proteome Discoverer always uses trypsin as the digestion enzyme for Percolator and does not provide an option to change it. Since, our peptides are not tryptic digested this provides the machine learning algorithm a wrong prior knowledge, which could result in a incorrect learning of the

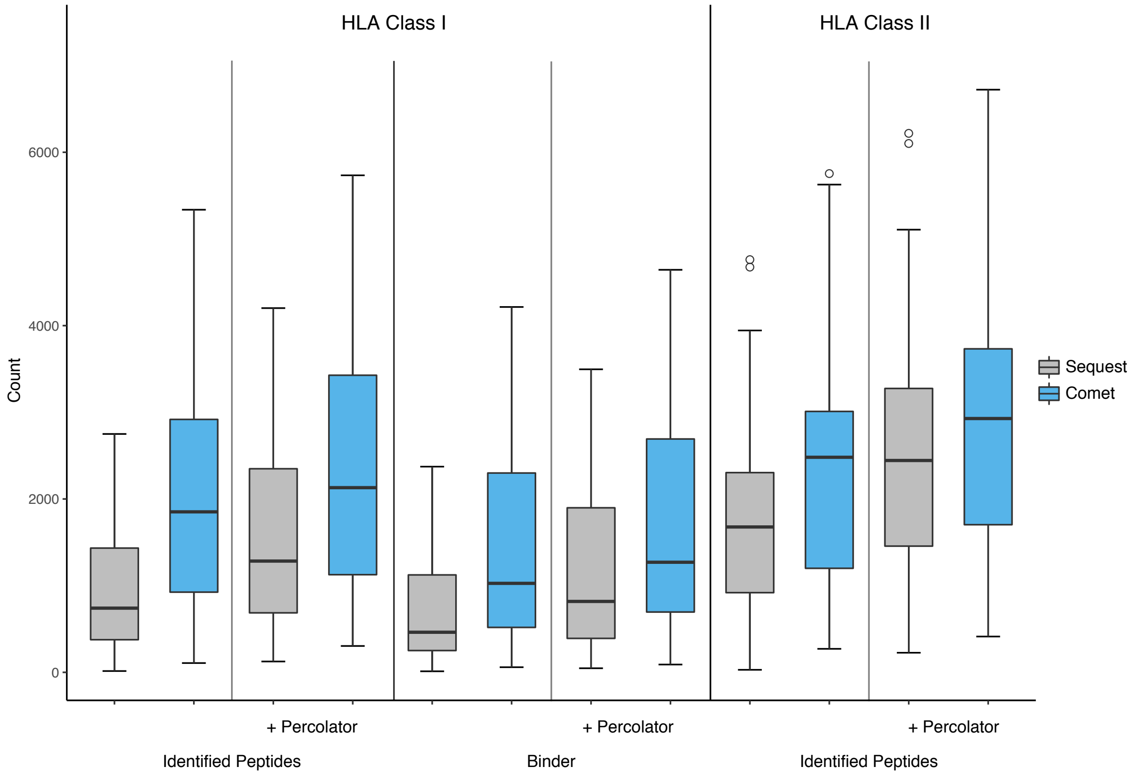


Figure 5.11: Results of the benchmark of Sequest HT, Sequest HT + Percolator, Comet, and Comet + Percolator. The number of identified HLA class I peptides per MS run (122 runs) is shown on the left. The number of binding peptides (binders) is on the right. The binding was predicted with netMHCpan-3.0 and the individuals' HLA type. Due to its bad prediction performance, no prediction for HLA Class II was performed. The line inside the boxplots represents the median, the box borders are the first and third quantile (25th and 75th percentiles), and the whiskers are the largest value no further than 1.5 *x* IQR from the hinge (IQR is the inter-quartile range, or distance between the first and third quartiles)[144].

peptide properties and a wrong significance calculation. Nonetheless, Percolator was robust and assigned more significant peptide identifications even when the wrong enzymatic cleavage was set by the Proteome Discoverer.

In addition to the number of identified peptides, we were also interested in the number of putative binders in our dataset. The number of binders can be used to verify that the found peptides are viable identifications. We calculated the binding affinity with netMHCpan-3.0 and the individual's HLA type. Figure 4.11 shows that we also gain more peptides when we use Comet + Percolator than with Sequest HT + Percolator. Furthermore, we calculated the ratio of binders to identified peptides. The percentage of binders was slightly higher with Sequest HT + Percolator (2.2% on average).

The benchmark between Sequest HT and Comet showed that Comet is superior to Sequest HT. This result was coherent with the benchmark results by Eng et al.[38]. Although both

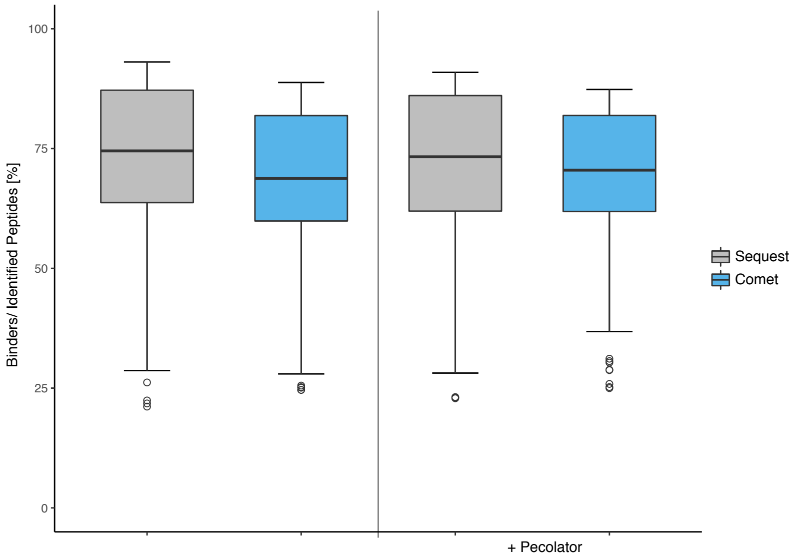


Figure 5.12: Results of the benchmark of Sequest HT, Sequest HT + Percolator, Comet, and Comet + Percolator. The fraction of binding HLA class I peptides and the number of identified peptides of 122 MS runs is shown as quality measurement. The binding was predicted with netMHCpan-3.0 and the individuals' HLA type. The line inside the boxplots represents the median, the box borders are the first and third quantile (25th and 75th percentiles), and the whiskers are the largest value no further than 1.5 x IQR from the hinge (IQR is the inter-quartile range, or distance between the first and third quartiles)[144].

algorithms are based on the Sequest algorithm, Comet has been shown to be superior. Because Sequest HT is a commercial software no detailed description of the algorithm is available. Therefore, we can only assume that the slight changes in the base algorithm made by Comet and two decades of development led to the massive gain of identifications[38]. As a consequence of the benchmark, we used Comet + Percolator to identify our peptides. There are many other identification algorithms like Mascot[94], OMMSA[45], MaxQuant[30], or MSGF+[63], which we did not consider in this benchmark. The results of multiple identification algorithms could also be combined, which might result in even better identifications. However, a complete benchmark of all identification algorithms would go beyond the scope of this thesis.

5.3.6 Data overview

In total, we performed 435 DDA mass spectrometry runs (class I = 216, class II = 219). These resulted in 64,534 unique class I peptide identifications and 90,958 class II peptide identifications. These peptides mapped uniquely to 13,516 class I proteins and 13,398 class II proteins. Figure 5.13 and 5.14 provide an overview of the number of identified peptides per tissue and patient. The samples varied massively in the number of identified peptides. The highest number of identified peptides for class I was 4,953 and the lowest 13, for class II the highest was 5,564 and the lowest 93. This variation was not tissue- or patient-specific but rather based on technical issues, such as heterogeneity of the tissue material or the amount of peptides after immunoprecipitation. Furthermore, the number of identified peptides did not depend on the quantity of sample used as immunoprecipitation input. Next, we predicted the binding affinity of the identified peptides to the individual's HLA type. Then, the number of binders divided by the number of identification, provided an estimation of the quality of the sample (Figure 5.15). Using this metric we again identified the deviant properties of the stomach samples. However, due to the lack of reliable HLA class II binding predictors, this quality measurement was not usable for HLA class II runs.

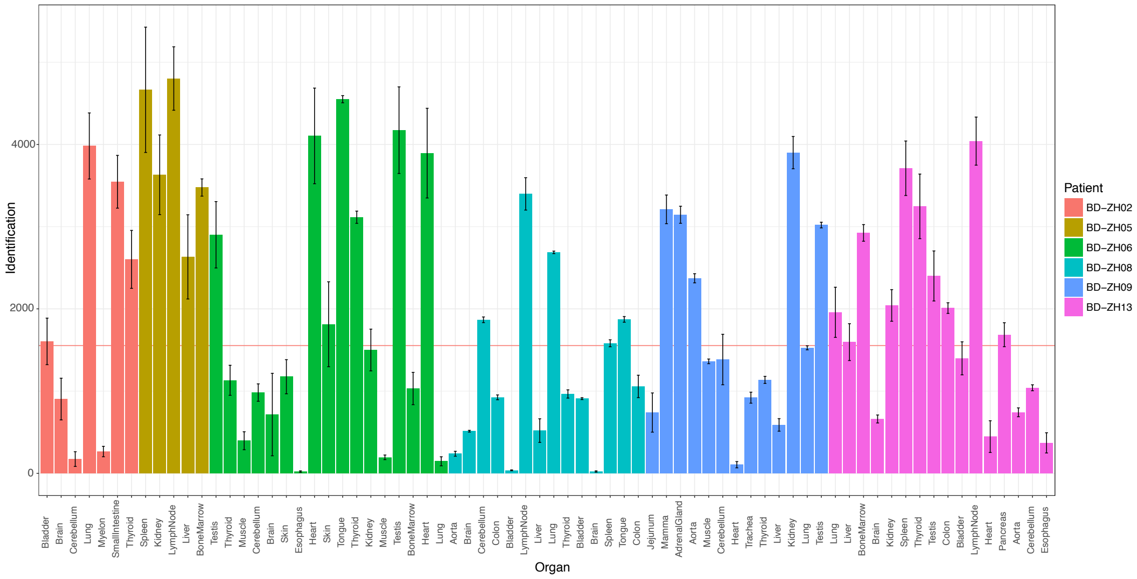


Figure 5.13: Number of identified class I peptides for each tissue and patient. The error bars show the standard deviation across the three technical replicates.

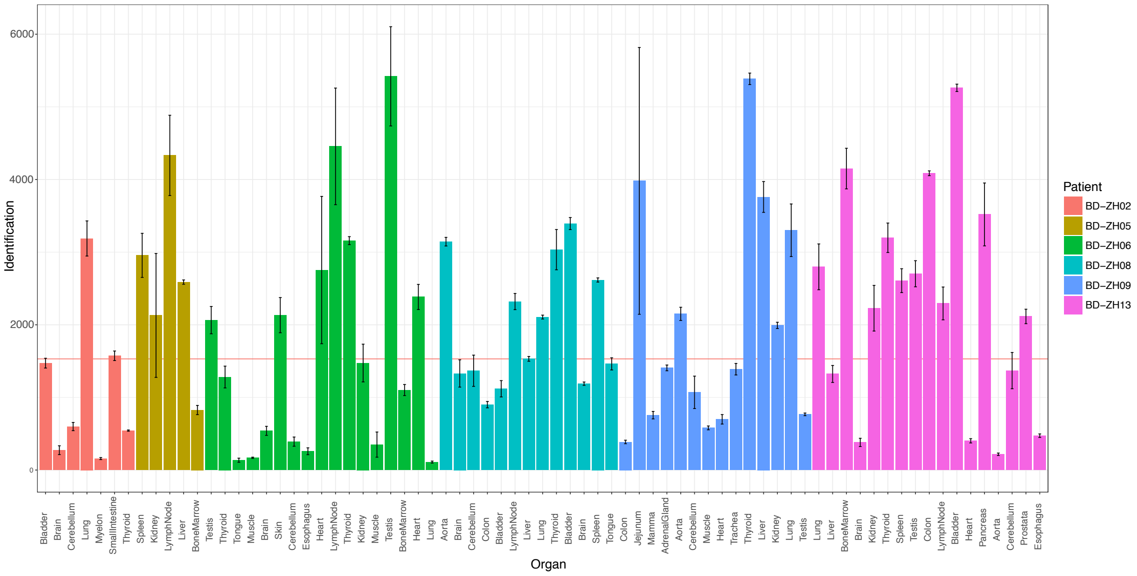


Figure 5.14: Number of identified class II peptides for each tissue and patient. The error bars show the standard deviation across the three technical replicates.

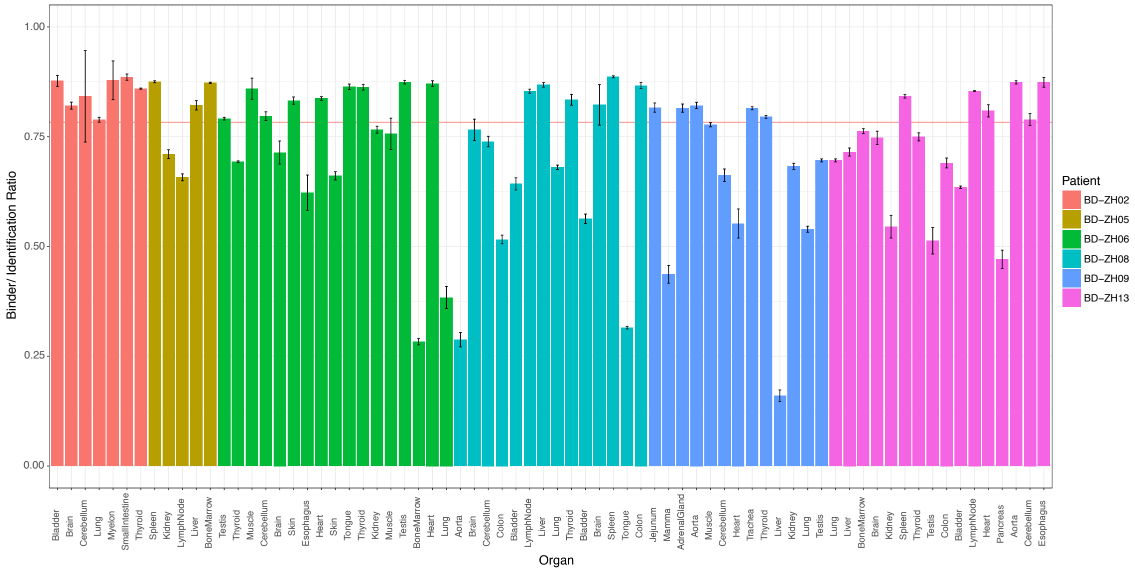


Figure 5.15: Ratio of HLA binding class I peptides to identified peptides for each tissue and patient. The error bars show the standard deviation across the three technical replicates. Binding prediction was performed with netMHCpan-3.0.

5.3.7 Properties of the immunopeptidome

After the first overview of data contained in the HLA Ligand Atlas, this section focuses on the description of the properties of the immunoprecipitation. First, we describe the length distribution of the peptides in conjunction with the HLA type to which they are predicted to bind. Second, we investigate the overlap on presented proteins on HLA class I and class II. Last, we look into class I peptides which can be found nested in class II peptides and describe how often length variations of the same core peptide can be found in HLA class I and class II separately.

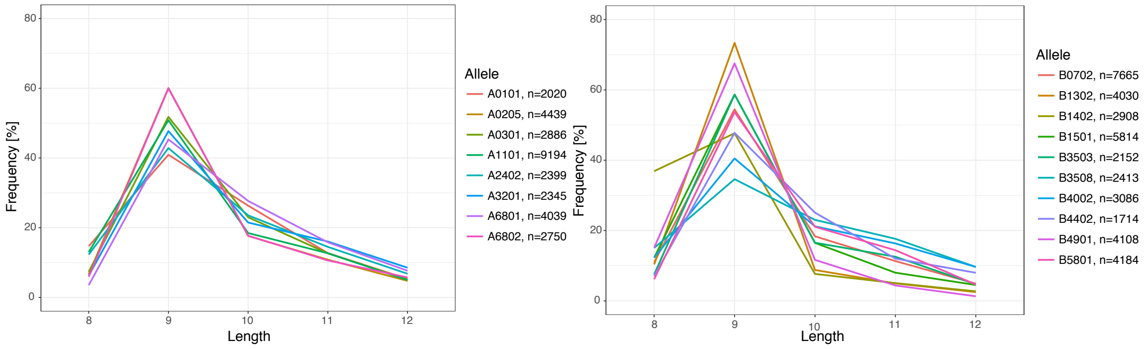
Peptide length distributions of different HLA alleles

Different HLA alleles bind different peptides. This hypothesis does not only apply to the sequence of the peptides, but also to the length of the binding peptide. Therefore, we analyzed the length distribution of the peptides belonging to each HLA class I type. In the first step, we assigned the peptides of each individual to an HLA allele using Gibbs clustering. After the clustering, we assigned found motifs by visually comparing the cluster motif to the binding motif of the individual's HLA type. We assigned only distinct motifs to each HLA allele. Due to imperfectly described binding motifs of HLA-C, we did not assign any motifs to HLA-C. Supplementary Table C.2 - C.7 show the resulting clusters for each individual and their HLA allele assignments. After the assignment to an HLA allele, we calculated the peptide length distribution (Figure 5.16).

As we had expected, most of the found peptides had length nine. However, we found small changes in the distribution for HLA-A alleles (Figure 5.16a). The relative abundance in length ten varied between the alleles. We observed a larger variation for HLA-B alleles (Figure 5.16b). HLA-B*14:02 had a high relative abundance for length eight peptides and only very few peptides with a length larger than nine. Furthermore, B*13:02 and B*49:01 had very few 10 to 12mers. A similar observation has recently been made by Abelin et al. (2017)[1]. They show an even more variable length distribution for HLA-B alleles. Furthermore, they achieve a relative abundance of 80% for length 9 for some alleles, which is not the case in our dataset. However, the comparison between their distribution and ours in general showed very similar results. Because Abelin et al. obtained their peptides from monoallelic HLA cell lines, this indirectly confirmed the use of the Gibbs clustering method for allelic separation of the peptides. To conclude, the length of the peptides presented by each HLA allele varied, but still most of the HLA class I peptides were of length nine. We did not use this method for HLA class II, because no clearly separated clusters could be obtained, presumably due to the shifting binding core of HLA class II.

HLA class I and II overlap

HLA class I presents endogenous proteins whereas HLA class II presents exogenous proteins. Although this paradigm holds, both can theoretically present similar proteins. However, due to their different antigen processing pathway (see Section 2.2) and the allele-specific presentation preferences, the proteins from which the presented peptides originate differ between class I



(a) HLA-A

(b) HLA-B

Figure 5.16: Frequency of the length of the peptides for each allele. We assigned the peptides to their allele with Gibbs clustering. We used only distinct clusters to assign peptides to HLA allele and merged alleles that occurred in more than one individual. *n* is the number of peptides found for the allele.

and II. Here, we calculated how much overlap exists on protein level between both HLA classes. First, we mapped all found HLA class I and II peptides back to their source protein and excluded all peptides that mapped to multiple proteins (Figure 5.17). Next, we calculated the overlaps between the proteins found on each patient, tissue, and across all samples. Based on the computed overlap, we calculated the Jaccard-similarity score for the two sets. The average Jaccard-similarity score was 0.109 (median 0.104). On average, we found 18.2% (median 15.5%) of all class I proteins also in class II. The comparison between tissue- and individual-specific overlap and the overall overlap showed, that most of the known human proteins were presented on both HLA classes if all samples are combined. However, if the proteins of only one sample are compared (e.g., only liver ZH05) there is a very low overlap. These results suggest that, although both HLA classes can present most of the human proteins, the two HLA classes in each tissue do not redundantly present source proteins.

Next, we calculated the overlap between the sets of peptides found in each sample, which led to a low average Jaccard-similarity score of 0.024 (median 0.016) and an overlap of class I peptides of 6.3% (median 3.0%). Because HLA class I and class II present peptides of different length, these results were not surprising. Therefore, we subsequently searched for class I peptides nested in class II peptides. On average, we found 9.6% of the class I peptides in class II peptides (median 5.7%).

In addition, we were interested in the nestedness of the peptides within each HLA class. To identify nested peptides, we searched for length variations of the peptides within each class and counted the peptides with length variations. A nested peptide was defined as a peptide with a length variation (e.g., SYFPEITH and SYPEITH**I**). Each peptide with a length variation was counted as one nested peptide and was not combined with the other peptides of the same core peptide. On average, 8.0% (median 6.4%) of HLA class I and 59.3% (median 61.8%) of HLA class II peptides were nested. This was consistent with the 3D structure of HLA. As described in Subsection 2.2.3, the protein structure of HLA class I is like a bathtub and allows only short dangling ends. Furthermore, the anchor positions of the binding motifs are at the beginning and end of the peptide sequence, which prevents a shortening of the peptide. In contrast, the 3D structure of HLA class II allows dangling ends and only needs a matching binding core to bind the peptides. Therefore, the large number of length variations in HLA class II peptides was consistent with the structure of HLA class II. However, if there is a biological function of nested peptides is not yet discovered.

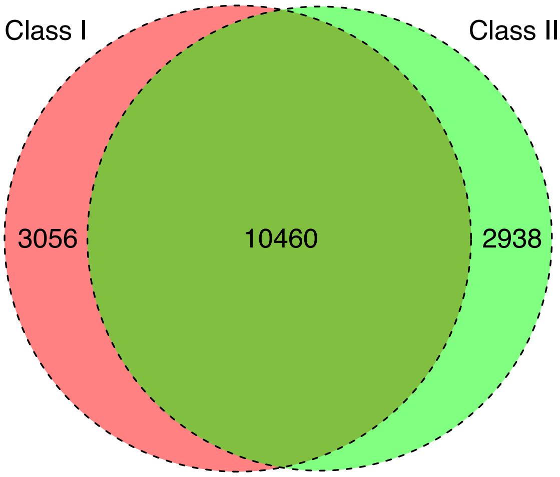


Figure 5.17: Source protein overlap for HLA class I and class II. The overlap was calculated based on the proteins of all patients and tissue types.

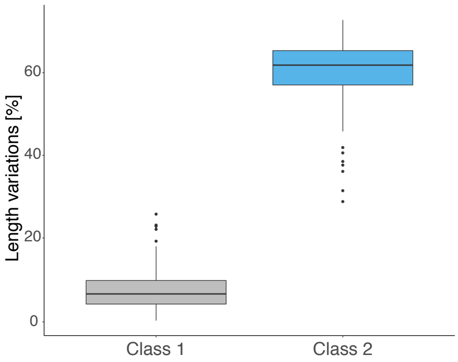


Figure 5.18: Peptide length variations inside each HLA class, which is defined as the number of nested peptides divided by the total number of peptides for each run. A nested peptide was defined as a peptide with a length variation (e.g., SYFPEITH and SYPEITH**I**). Each peptide with a length variation was counted as one nested peptide and was not combined with the other peptides of the same core peptide.

5.3.8 Individual and organ differences in immunopeptidome

In former studies of the immunopeptidome, the focus was mostly on specific tumor types and therefore on specific tissues[12,15,68]. In contrast, the HLA Ligand Atlas contains various tissue types from multiple individuals. To assess the similarity of the immunopeptidome of different tissues, we calculated the pairwise Jaccard-similarity score of the sets of peptides and proteins found on each tissue. Based on these distances, we performed a hierarchical clustering of individuals and tissues (HLA class I: Figure 5.19; HLA class II: Supplementary Figure C.8). The clustering does not show any similarities between tissues of different individuals. However, in our dataset the number of shared HLA class I alleles between the individuals was very small (Allele: occurrence, A*11:01: 3, A*68:01: 2, B*07:02: 2, B*15:01: 2, C*03:03: 2, C*04:01: 2, C*07:01: 2, C*07:02: 2, other alleles: 1) and therefore we found a clear HLA binding specificity separation on peptide level. Next, we tried to remove the influence of the HLA binding. We mapped the peptides to their source protein and calculated the Jaccard-similarity score and the hierarchical clustering (HLA class I: Figure 5.20; HLA class II: Figure C.9). Again, the resulting heat maps showed a clear clustering by individual and not by tissue type. This result implies that the HLA type of an individual also defines the majority of the presented proteins. In addition, we tried to remove the individual's background. We subtracted the mean/median presentation frequency of each peptide/protein per individual and recreated the heat map with the Euclidean distance (Figure C.10-C.17). In the resulting clustering the individuals are still clustered and only some tissues types group together. However, removing the mean/median peptide presentation of an individual is a simple method and more complex methods like deconvolution using Bayesian models might yield a better separation on the tissue level. Furthermore, a larger HLA type overlap might result in more distinct clustering of tissues after subtraction of the individual background. To this end, we will have to expand the database and measure the immunopeptidome of more individuals with similar HLA types.

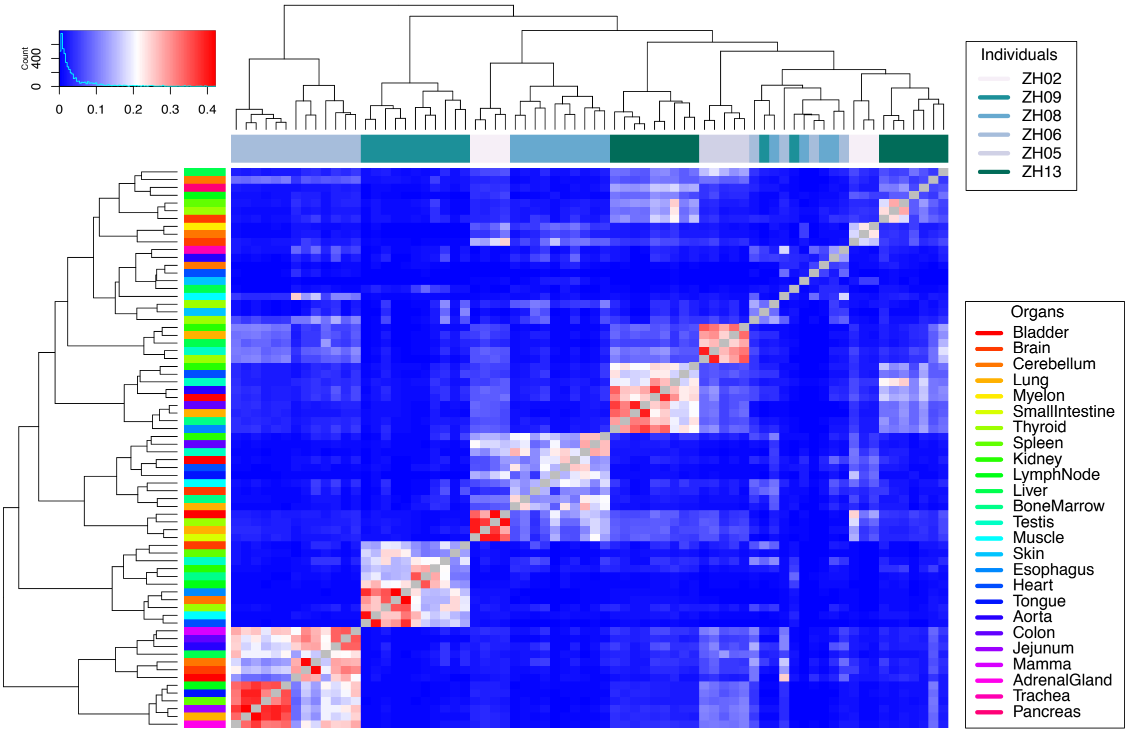


Figure 5.19: Heat map of pariwise distances of sets of identified HLA class I peptides for patients and tissues. We calculated the pairwise distance as the Jaccard-similarity score on presence and absence of identifications. The color coded dendogram on top shows individuals' ID, the dendogram on the left shows the tissue types.

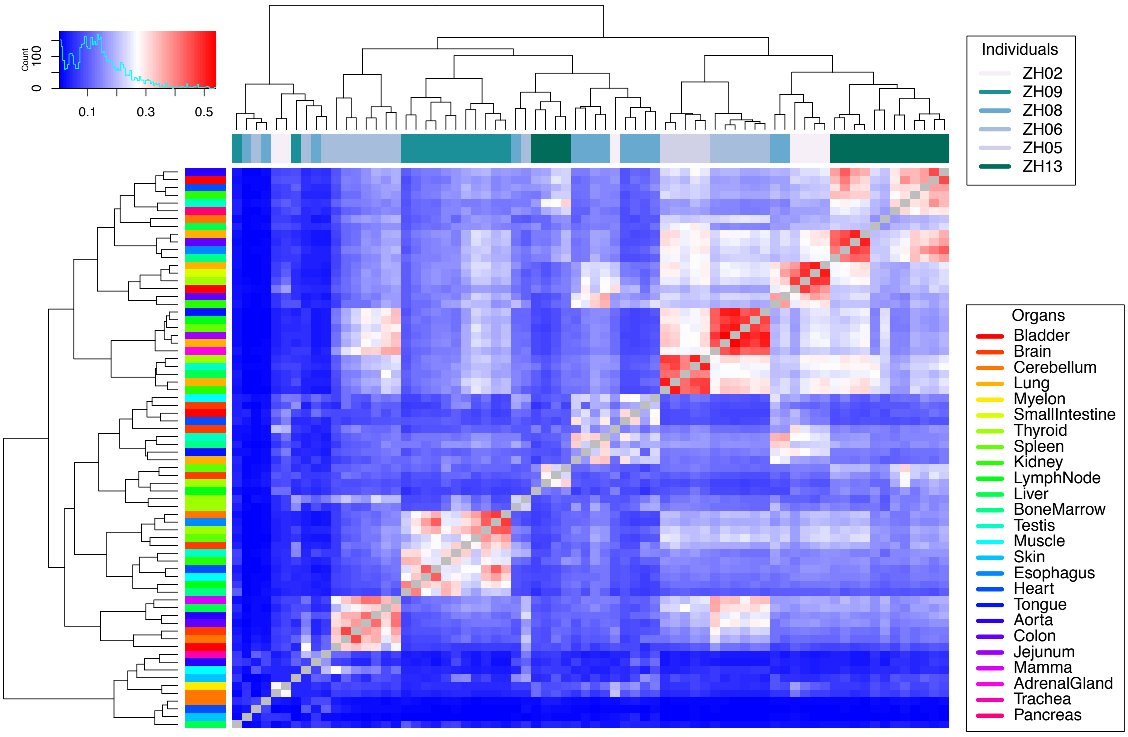


Figure 5.20: Heat map of pariwise distances of sets of identified HLA class I proteins for patients and tissues. We calculated the pairwise distance as the Jaccard-similarity score on presence and absence of proteins. The color coded dendogram on top shows individuals' ID, the dendogram on the left shows the tissue types

5.3.9 Tissue-specific Proteins

Above we demonstrated with a clustering that the immunopeptidome is more similar within individuals than within tissue samples obtained from the same tissue in different individuals. Nevertheless, we searched for tissue-specific proteins, which we defined as proteins that were exclusively found on one tissue and never on any other tissue. We analyzed only tissues, for which we had obtained at least three different samples to avoid random hits. Table 5.1 shows that we had at least three samples for 17 different tissue types. However, we only considered proteins that were found on at least three different samples, which reduced the number of tissues to five for HLA class I (brain, kidney, liver, lung, thyroid) and six for HLA class II (brain, colon, heart, kidney, liver, lung, thyroid). Table 5.4 lists the tissue-specific proteins for HLA class I and Table 5.5 lists those for HLA class II. The protein expression data was obtained from `www.proteomicsdb.org`[143].

We found only one protein - Opalin - to be both tissue-specific in its presentation on HLA and its expression on protein level. In addition, for five tissue-specific proteins presented on HLA, we found no protein expression at all in the specific tissue (TAF7L, KCNJ9, PREB, TMPS2, and WDR75). All other tissue-specific proteins presented on HLA were at least expressed in their tissue on protein level and many were expressed in high levels. In addition, we more frequently found tissue-specific HLA peptides with high compared to low expression. However, it has to noted that the tissue-specific protein presented on HLA may be present on other tissues with high-protein expression that we did not include in our dataset yet. Therefore, the expansion of the HLA Ligand Atlas should include further tissue types to ensure a more diverse view on the immunopeptidome. In addition, with more data we could combine similar tissue types such as brain and cerebellum to search for peptides or proteins specific for biological systems or functions rather than individual tissues.

Table 5.4: Tissue-specific HLA class I proteins. The count column shows the number of tissue samples on which we found peptides from proteins. Protein-expression data obtained from `www.proteomicsdb.org`.

Tissue	Protein	Gene	Count	High protein expression	Tissue-specific protein	Name
Brain	Q16650	TBR1	3	Yes	No	T-box brain protein 1
Brain	Q96PE5	OPALI	3	Yes	Yes	Opalin
Kidney	O00476	NPT4	3	No	No	Sodium-dependent phosphate transport protein 4
Kidney	Q9UHE5	NAT8	4	Yes	No	N-acetyltransferase 8
Kidney	Q8TCC7	S22A8	3	Yes	No	Solute carrier family 22 member 8
Liver	Q9NQ94	A1CF	3	Yes	No	APOBEC1 complementation factor
Lung	O95436	NPT2B	3	Yes	No	Sodium-dependent phosphate transport protein 2B
Lung	P11686	PSPC	4	Yes	No	Pulmonary surfactant-associated protein C
Lung	O95171	SCEL	4	Yes	No	Sciellin
Thyroid	Q5H9L4	TAF7L	4	Not found	Not found	Transcription initiation factor TFIID subunit 7-like
Thyroid	Q92806	KCNJ9	3	Not found	Not found	G protein-activated inward rectifier potassium channel 3
Thyroid	Q9HCU5	PREB	3	Not found	Not found	Prolactin regulatory element-binding protein

Table 5.5: Tissue-specific HLA class II proteins. The count column shows the number of tissue samples on which we found peptides from proteins. Protein-expression data obtained from `www.proteomicsdb.org`.

Tissue	Protein	Gene	Count	High protein expression	Tissue-specific protein	Name
Brain	O76070	SYUG	3	Yes	No	Gamma-synuclein
Brain	P42658	DPP6	3	Yes	No	Dipeptidyl aminopeptidase-like protein 6
Brain	P06307	CCKN	3	Yes	No	Cholecystokinin
Colon	P01282	VIP	3	Yes	No	VIP peptides
Heart	P19429	TNNI3	3	Yes	No	Troponin I, cardiac muscle
Heart	O14639	ABLM1	3	Yes	No	Actin-binding LIM protein 1
Kidney	P02489	CRYAA	3	Yes	No	Alpha-crystallin A chain
Liver	P06133	UD2B4	4	Yes	No	UDP-glucuronosyltransferase 2B4
Liver	P22310	UD14	3	Yes	No	UDP-glucuronosyltransferase 1-4
Liver	P22760	AAAD	3	Yes	No	Arylacetamide deacetylase
Lung	P08476	INHBA	3	Yes	No	Inhibin beta A chain
Lung	P11686	PSPC	5	Yes	No	Pulmonary surfactant-associated protein C
Lung	O15393	TMPS2	3	Not found	Not found	Transmembrane protease serine 2
Lung	Q8IWA0	WDR75	3	Not found	Not found	WD repeat-containing protein 75
Lung	Q6UY14	ATL4	3	No	No	ADAMTS-like protein 4
Lung	P07988	PSPB	3	Yes	No	Pulmonary surfactant-associated protein B
Thyroid	P07202	PERT	6	Yes	No	Thyroid peroxidase

5.4 Discussion and Outlook

With 85 samples, the HLA Ligand Atlas is currently the largest publicly available benign immunopeptidome dataset. Here, we described a detailed analysis of the immunopeptidome of HLA class I and II. We performed a statistical analysis of the immunopeptidome and described in detail the several quality control steps during data generation and analysis. First, we showed with a time-series experiment that when the individuals were stored for up to 72 h at 4 °C, the quality and quantity of the immunopeptidome barely change over time. Next, we checked for outlier runs with mass spectrometry run properties. We used Gibbs clustering and identified atypical HLA binding motifs for two stomach samples that we found to be outliers. These binding motifs were explained by the cleavage specificity of pepsin. However, we could not determine when (before or during immunoprecipitation) and where (*in vivo* or *in vitro*) the pepsin cleavage occurred. Further experiments with additional pepsin inhibition during the immunoprecipitation may help to evaluate if the digestion occurs during the immunoprecipitation. If the pepsin digestion is not an artifact of the immunoprecipitation, we have to perform more experiments to ensure that the digestion already occurs in the living organism and not during the storage of the body at 4 °C. This could be done by using fresh stomach samples from biopsies instead of samples from autopsies. To avoid non-HLA specific peptide contamination, we excluded the stomach samples from our further analyses. In the next step of the quality control we looked for typical contaminants in the immunoprecipitation. We benchmarked different quality measurements and chose the protein coverage, which performed best. Next, we discarded peptides from proteins with more than 40% coverage in any run. With this method we cannot exclude all contaminants. In particular, peptide contaminants from proteins with low coverage will pass this filter. One way to mark these peptides as contaminants and remove them from the analysis would be to use a blacklist of peptides that could be created based on expert knowledge and years of experience. However, this blacklist could be biased by the presumption of the expert. Therefore, we only use the unbiased approach described above.

After ensuring the quality of the immunopeptidome data, we benchmarked our identification and processing pipeline. We compared Sequest HT and Comet including Percolator and showed that Comet + Percolator outperformed Sequest HT, with and without Percolator. Here, we focused on two identification tools and did not include others such as Mascot[94], OMMSA[45], MaxQuant[30], or MSGF+[63]. Since, a complete benchmark of all identification algorithms would go beyond the scope of this thesis.

The FDR accumulation in these large datasets is an often-raised concern. When we combine many runs that were individually processed, the FDR can accumulate, which results in a total FDR higher than 5%. This problem does not only concern the HLA Ligand Atlas but also proteomics data bases like the ProteomicsDB[143] and the Human Protein Atlas[126]. In proteomics databases, a protein-based FDR can be calculated to reduce the problem, which is not possible for our immunopeptidome dataset. We were not able to correct for the FDR accumulation because no method dealing with the problem has been published yet.

In the next section, we defined properties of the immunopeptidome. First, we analyzed the length distribution of the peptides belonging to specific HLA allele. We assigned the peptides using Gibbs Clustering and a manual assignment of the motifs to its HLA type. This manual assignment is very time-consuming and will no longer be possible when the database grows and the number of patient increases. To solve this problem, we assigned each cluster to an HLA allele with netMHC and used the cluster assignment for the peptides. However, in our analysis we did not assign all clusters to avoid trash clusters. Therefore, we did a quality measurement to remove these clusters before an automatic assignment was used. Next, we assessed the overlap between HLA class I and class II, length variations within each HLA class, and between the classes. We showed that when we combine the dataset most of the proteins can be found in both HLA classes but when only one sample is analyzed, the protein overlap is rather small. Furthermore, the length variation analyses showed that only few HLA class I peptides can be found in HLA class II peptides. However, nested peptides are of general interest for vaccination approaches[64,113]. Finally, we identified a large group of nested peptides in HLA class II, but not in HLA class I.

In the last two sections, we compared the samples across individuals and tissue types. First, we calculated a Jaccard distance matrix on peptide and protein level and showed that the HLA type of the individuals is critical for the selection of presented peptides and proteins. In addition, we subtracted the individual-specific background presentation. However, the resulting clustering did only slightly change to a tissue-based grouping. Our background subtraction method was a very simple approach and more advanced methods to deconvolute the individual- and tissue-specific presentation could be used (e.g., Bayesian-based methods). Furthermore, our analyzed individuals shared only very few HLA types. Therefore, additional samples should be obtained from individuals with similar HLA types.

In the last section, we tried to identify tissue-specific proteins. However, we found only one protein that was both tissue-specific on immunopeptidome and proteomics level. The rather small amount of tissue-specific proteins was likely caused by the few samples that were available for each tissue type, which could be solved by obtaining more samples from specific tissues. In addition, new tissue types would be needed to ensure that these tissue-specific proteins are not presented on other not analyzed tissues. To sum up, we presented different quality control steps, an identification method benchmark, and a comprehensive description of the benign immunopeptidome.

Chapter 6

A meta-analysis of HLA peptidome composition in different hematological entities

This part of the thesis describes a meta-analysis of the HLA peptidome in different hematological entities. It shall be an example of how immunopeptidome data of malignant and benign tissues can be analyzed and especially demonstrate the importance of a large dataset of benign tissues like the HLA Ligand Atlas. The content is based on and has been published by Backert et al.[12]

A meta-analysis of HLA peptidome composition in different hematological entities: Entity-specific dividing lines and "pan-leukemia" antigens
Backert L, Kowalewski DJ*, Walz S, Schuster H, Berlin C, Neidert MC, Schemionek M, Brümmendorf T, Vucinic V, Niederwieser D, Kanz L, Salih HR, Kohlbacher O, Weisel K, Rammensee HG, Stevanovic S, Stickel JS*

6.1 Introduction

In contrast to the recent breakthrough advances in the treatment of solid malignancies by antigen-unspecific immune-checkpoint blockade[22,29,48,81,86,100] the success of this highly promising treatment modality has so far been limited in hematological cancers[8,77] with the prominent exception of Hodgkin lymphoma[9,10]. As clinical effectiveness of checkpoint inhibition has been shown to be directly correlated to mutational load in solid tumors[102,117] and mutation-derived neoepitopes have been identified as targets of the resultant anti-tumor T-cell responses[49,50,121], it may be surmised that the suboptimal effectiveness in hematologic malignancies (HM) may at least in part be attributed to the predominantly low mutational burden of these cancer entities[4,131]. On the other hand, HM can be effectively treated by stem cell transplantation[21,101,141], donor lymphocyte infusion[107,108,111] or the more recently developed adoptive approaches utilizing chimeric antigen receptor (CAR) T cells, which showed breakthrough effectiveness, even in previously therapy-resistant forms of malignancy[44,80,95]. However, apart from the latter, these approaches are hampered by their infrequent effectiveness and, more importantly, severe off-target toxicity such as graft-versus-host disease. As CAR T-cell therapy likely will remain restricted to only a handful of cell surface (differentiation) antigens (e.g. CD19[80], HER2[145], CEA[51]), there is a pressing need to identify new targets and suitable treatment strategies for hematological malignancies not amenable to CAR T-cell therapy. For this aim, the identification of HLA-restricted T-cell epitopes on HM and their implementation in adoptive, engineering- or vaccine-based T-cell immunotherapy is a highly attractive option, rendering a vast array of intracellular - and potentially more specific - HM antigens amenable to immunological targeting. To this end our group and others have extensively studied the HLA-presented immunopeptidome of hematological cancers including acute myeloid leukemia (AML)[19], chronic myeloid leukemia (CML)[119], chronic lymphocytic leukemia (CLL)[67] and multiple myeloma (MM)[135], which led to the identification of multiple pathophysiologically relevant epitopes of anti-HM T-cell responses and inspired the notion that immune control in these low-mutational entities may effectively be mediated by T cells targeting non-mutated epitopes[70]. As the development of novel immunotherapeutic compounds is a highly cost- and time-intensive enterprise[133,138], such non-mutant, common antigens represent highly attractive targets for off-the-shelf immunotherapy, which may be suited for the effective treatment of a substantial proportion of the patient population.

In this study we present a meta-analysis of our previous studies on the immunopeptidomes of the four major hematologic cancers in adults, AML[19], CML[119], CLL[67] and MM[135], addressing the similarity of these malignancies on the immunologically pivotal level of HLA-restricted presentation with the dedicated aim of investigating the existence and prevalence of potential "pan-leukemia antigens".

6.2 Material and Methods

6.2.1 Patient blood and bone marrow samples

Peripheral mononuclear cells (PBMC) from AML, CLL and CML patients and bone marrow mononuclear cells (BMNC) from MM patients (provided by the Departments of Hematology and Oncology in Tübingen, Leipzig and Aachen, Germany) at the time of initial diagnosis or relapse prior to therapy were isolated by density gradient centrifugation (Biocoll, Biochrom GmbH, Berlin, Germany) and erythrocyte lysis (EL buffer, Qiagen, Venlo, Netherlands) (Table 6.1). For all AML and CLL samples the frequency of malignant cells within the PBMC isolate was >

80%. For MM samples the percentage of malignant plasma cells within the BMNC fraction was $> 60\%$. For CML we analyzed whole blood samples of 12 CML patients in the chronic phase (no blasts), two in the accelerated phase (18-20% myeloid blasts) and two in a blast crisis (50-60% myeloid blasts). Informed consent was obtained in accordance with the Declaration of Helsinki protocol. The study was performed according to the guidelines of the local ethics committee (373/2011BO2, 142/2013BO2). HLA typing was carried out by the Department of Hematology and Oncology, Tübingen, Germany. Samples were stored at -80°C until further use.

6.2.2 Healthy control tissue samples

PBMC and bone marrow mononuclear cells (BMNC) from healthy volunteers were isolated by density gradient centrifugation (Biocoll, Biochrom GmbH, Berlin, Germany) and erythrocyte lysis (EL buffer, Qiagen, Venlo, Netherlands). Normal tissue samples from patients and autopsy material were provided by the University Hospital Tübingen, Germany and the University Hospital Zürich, Switzerland (Table 6.1). Specimens were frozen in liquid nitrogen immediately after resection. Informed consent was obtained in accordance with the Declaration of Helsinki protocol

6.2.3 Myeloma cell lines (MCL)

For HLA ligandome analysis the myeloma cell lines (MCLs, U266, RPMI8226 and JJN3) were cultured in the recommended cell media (RPMI1640 (Gibco, Carlsbad, CA, USA), IMDM (Lonza, Basel, Switzerland)) supplemented with fetal calf serum, 100 IU/L penicillin, 100 mg/L streptomycin, and 2 mmol/L glutamine at 37°C and 5% CO_2.

6.2.4 Isolation of HLA ligands from primary samples and MCLs

HLA class I molecules were isolated using standard immunoaffinity purification as described before in Berlin et al.[19], Kowalewski et al.[69], and section 2.3, using the pan-HLA class I specific mAb W6/32 (produced in house) to extract HLA ligands.

6.2.5 Analysis of HLA ligands by LC-MS/MS

HLA ligand extracts were analyzed in five technical replicates as described previously[66] . In brief, peptide samples were separated by nanoflow HPLC (RSLCnano, Thermo Fisher, Waltham, MA, USA) using a 50 $\mu m \times$ 25 cm PepMap RSLC column (Thermo Fisher) and a gradient ranging from 2.4 to 32.0% acetonitrile over the course of 90 min. Eluting peptides were analyzed in an online-coupled LTQ Orbitrap XL mass spectrometer (Thermo Fisher) using a top 5 CID (collision-induced dissociation) fragmentation method.

Table 6.1: Tissue samples and peptide yields comprised in the non-malignant primary tissue database. Reproduced with permission from Backert et al.[12]

Tissue	Number of analyzed samples	Unique peptide IDs
Brain	5	2208
Kidney	30	16045
Lung	3	5427
Muscle	2	583
Small Intestine	2	3098
Spleen	2	4585
Bladder	1	1416
Heart	1	1128
Myelon	1	383
Pancreas	2	1576
Skin	2	543
Stomach	1	1198
Thyroid	1	1451
Adrenal Gland	1	690
Esophagus	1	392
Liver	13	10081
Testicle	1	1736
Trachea	1	334
Bone Marrow	9	3591
PBMC	30	17322
Granulocytes	3	4224
Colon	32	12539
Ovary	3	1036

6.2.6 Database search and HLA annotation

Data processing was performed using the software Proteome Discoverer (v1.3, ThermoFisher) and the Mascot search engine (Mascot 2.2.04; Matrix Science, London, UK)[94]. The search database was the human proteome as comprised in the Swiss-Prot database (20,279 reviewed protein sequences, September 27th, 2013) without enzymatic restriction. Precursor mass tolerance was set to 5 ppm, and fragment mass tolerance was set to 0.5 Da. Oxidized methionine was allowed as a dynamic modification. The peptide-level false discovery rate (FDR) was estimated using a decoy database consisting of the shuffled target database and the Percolator algorithm (v2.04)[57]. The results were filtered with $q \leq 0.05$ (5% FDR). Peptide lengths were limited to 8-12 amino acids. Protein inference was disabled, allowing for multiple protein annotations of peptides. HLA annotation was performed using NetMHCpan (v3.0)[89], annotating peptides with IC_{50} scores $\leq$ 500 nM and/or percentile ranks $\leq$ 2% as ligands of the corresponding HLA allotype. Samples for which only two-digit HLA typings were available, the missing sub-alleles were inferred based on the assumption of the most frequent four-digit allotype. For quality control, yield thresholds of ≥200 unique HLA class I ligands for primary samples and ≥1,000 unique HLA class I ligands for MCL were applied.

6.2.7 Software and statistical analysis

Data processing and analysis was performed in Python 2.7.10 and FRED[42] (binding prediction). Statistical analyses were conducted using in R 3.3.1[96]. All R scripts can be accessed at `https://github.com/linusb/pan_leukemia`. The heat maps were created using the R package gplots[139]. The Jaccard index graphs were visualized using the igraph[32] package, and the Venn diagrams were plotted using the R package VennDiagram[25]. The clustering and distance graph analysis was performed using complete linkage clustering and the Jaccard distance, which measures the qualitative dissimilarity between two HLA peptidomes. The Jaccard distance in these analyses was calculated as the difference of the size of the union minus the intersection of HLA peptidomes, divided by the size of their union.

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{6.1}$$

d_J is the Jaccard distance and J the Jaccard index. A and B represent the sets of peptides for the two samples. For all possible sample combinations, this pairwise comparison was computed. The Jaccard distance was selected as it is equivalent to overlap visualization by Venn diagrams commonly utilized in HLA peptidomics studies. The thresholds were defined empirically, with Jaccard similarities of 0.1 yielding optimal sensitivity and specificity. For the clustering, the Jaccard distance graphs, and the overlap analysis of "cancer-exclusive" HLA ligand, the normal tissue immunopeptidome was subtracted. In addition, HLA ligands occurring only once across all samples were discarded and only samples containing more than five unique HLA class I ligands were included. The layout of the distance graph analysis was computed using the Fruchterman-Reingold layout algorithm.

6.3 Results

6.3.1 Hierarchical clustering of HLA-restricted antigens on source protein level does not discern specific hematological malignancies

In order to obtain a comprehensive overview of the antigenic landscape of the four major HM and their immunological relatedness, we first performed an unsupervised hierarchical clustering of source proteins (8,053 unique source proteins) represented by HLA-restricted peptides (40,361 unique peptide IDs) in the immunopeptidomes of primary AML (n=19), CML (n=16), CLL (n=35) and MM/myeloma cell line MCL (n=9/4, Figure 6.1). Without stratification of patients for expression of specific HLA allotypes, this source protein level analysis did not delineate clusters along entity lines but rather revealed that the antigenic landscape is divided into a smaller subset of non-entity specific common antigens (Figure 6.1, upper box) juxtaposed with a larger, highly heterogeneous set of sample-specific antigens (Figure 6.1, lower box). Whereas the larger group of sample and subset-specific source proteins clearly reflects a high degree of tumor/patient individuality, the presence of a smaller common subset of antigens hints at the potential presence of highly frequent and entity-spanning pan-leukemia antigens. In order to evaluate the presence of such targets in the HM dataset, we shifted our analysis to the HLA peptidome level, specifically filtered for HLA ligands which were exclusively detected on tumor tissue and subsequently performed HLA allotype-specific immunopeptidome profiling and cluster analyses for the seven most common HLA allotypes (A*01:01, A*02:01, A*03:01, A*24:02, B*07:02, B*08:01, B*18:01; >95% population coverage in the Caucasian population)[110]. To this end, we first subtracted from the dataset of HM-derived HLA ligands any peptide (irrespective of HLA restriction) also contained in our comprehensive in-house database of HLA ligands detected on non-malignant primary tissue specimens (n=147, number of unique peptides: 44,541, Supplementary Table 6.1). For the remaining set of peptides, which were only detected on malignant samples (from now on referred to as "cancer-exclusive") we computationally assigned the restricting HLA allotypes and compiled allotype-specific HLA ligand datasets for further analysis (Supplementary Table D.5).

Importantly, we cannot rule out presentation of these "cancer-exclusive" HLA ligands on normal (sub-)tissues or cell populations at levels below the limit of the detection or on sample types missing in our normal tissue database.

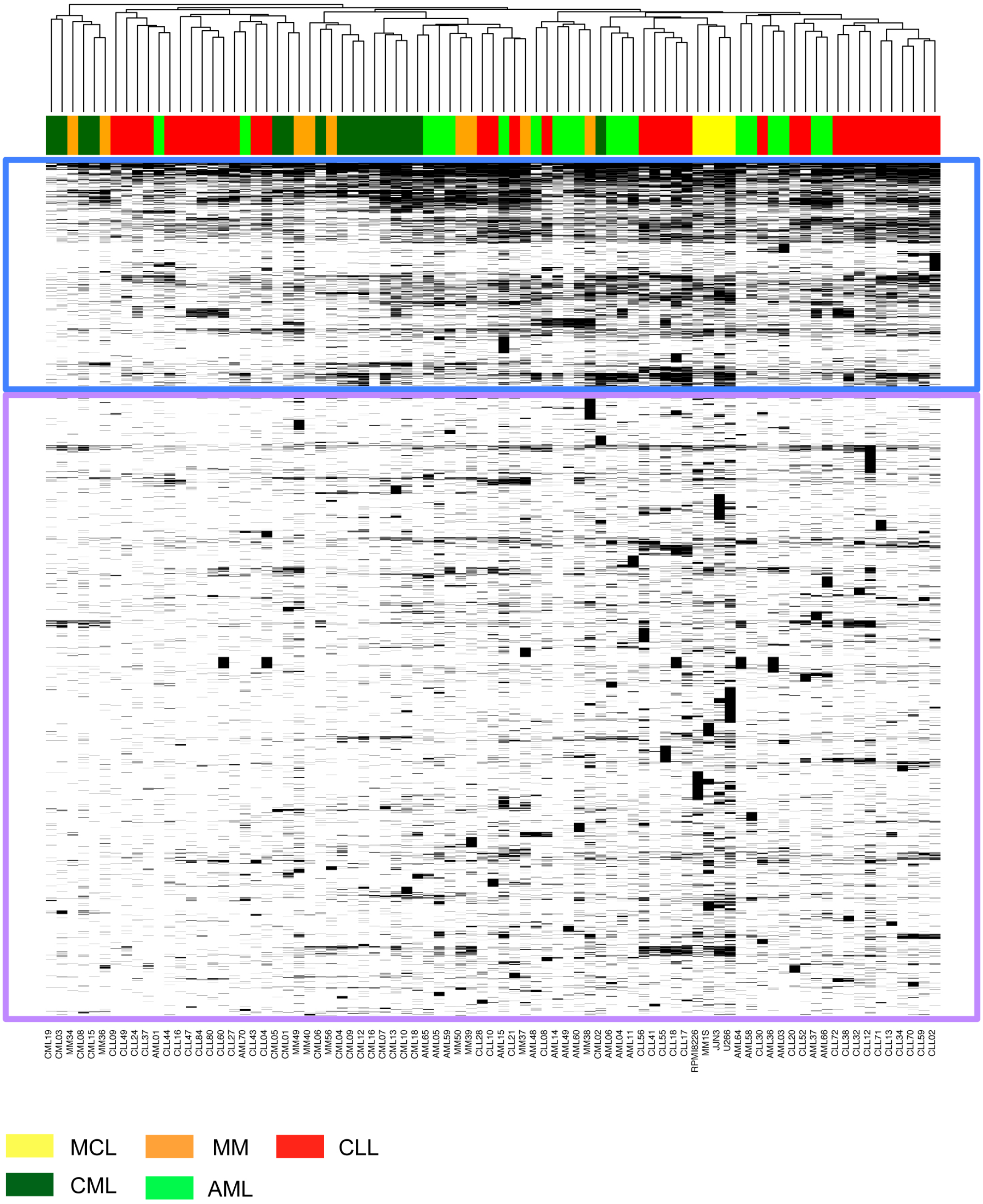


Figure 6.1: Unsupervised clustering analysis of HLA ligand source proteins represented in the immunopeptidomes of HM. Peptides identified by LC-MS/MS in HLA class I ligand extracts of AML (light green, n=19), CML (dark green, n=16), CLL (red, n=35) and MM/MCL (orange/yellow, n=9/4) were mapped to their source proteins. For conserved sequences mapping to multiple proteins all protein annotations were retained. Complete linkage clustering was performed based on the Jaccard similarity coefficient of HLA ligand source proteins. A subset of source proteins shared across samples and entities with high frequencies of presentation is highlighted in the blue box; infrequent, sample/entitity-specific source proteins are highlighted in the purple box. Reproduced with permission from Backert et al.[12]

6.3.2 HM entities and lineages can be distinguished purely based on HLA allotype-specific immunopeptidome composition

Unsupervised clustering analysis of the HLA-A*02:01-restricted HM immunopeptidomes (AML (n=9), CML (n=6), CLL (n=16), MM/MCL (n=4/1)) resulted in clear clustering of samples belonging to the same hematological cancer entities, as well as coherent clustering of the lineages these malignancies arise from (Figure 6.2A). This suggests that the HLA ligandome directly reflects tumor/lineage-specific biology, which is further underscored by the findings of gene ontology analyses (GO Term BP) using DAVID[53], which identified B-cell receptor signaling (GO ID: 0050853) as a significantly enriched biological process (P<0.05 after Benjamini-Hochberg correction for multiple testing) represented selectively in the immunopeptidome of the lymphatic lineage (CLL and MM/MCL).

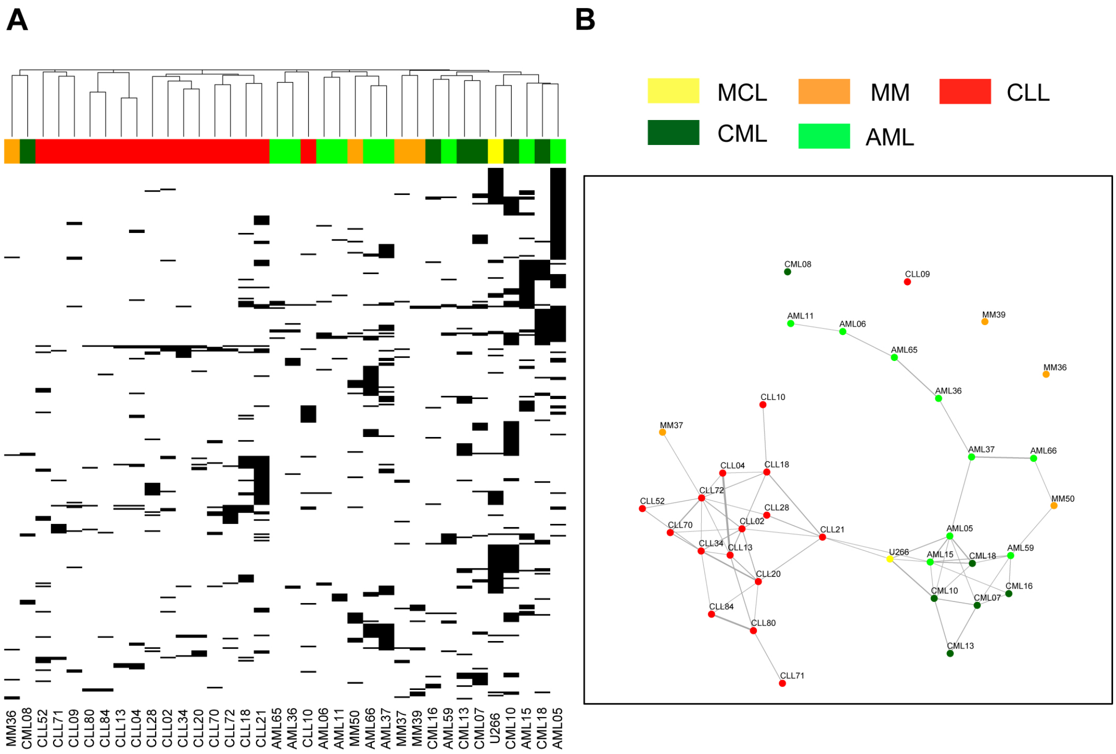


Figure 6.2: Unsupervised clustering analysis and Jaccard distance graphs of "cancer-exclusive" HLA-A*02:01 ligands on hematological cancers "Cancer-exclusive" HLA-A*02:01 ligands identified on AML (light green, n=9), CML (dark green, n=6), CLL (red, n=16) and MM/MCL (orange/yellow, n=4/1) were analyzed by: A. Complete linkage clustering based on the Jaccard similarity coefficient of A*02:01 immunopeptidomes. B. Jaccard distance graphs. Samples showing ≥10% Jaccard similarity of their "cancer-exclusive" HLA-A*02:01 immunopeptidomes were linked by edges, with the thickness of the edge positively correlating with the degree of similarity. Reproduced with permission from Backert et al.[12]

To further investigate and visualize the inter-relatedness of samples and to assess lineage-specific dividing lines, we performed Jaccard distance graph analysis, which identified sub-networks of closely related (≥10% immunopeptidome overlap, linked by edges) CLL and CML samples. AML on the other hand showed a chainlike structure of related samples, which covers a vast range of possible A*02:01 immunopeptidome compositions. Connections across entity boundaries were only identified in two isolated cases (Figure 6.2B). For the other HLA allotypes similar observations were made, with clear entity-specific dividing lines detected in all cases and CLL universally clustering in centralized subnetworks (Supplementary Figures 6.3 & 6.4).

Together, these findings suggest that peptide-specific T-cell immunotherapy in hematologic malignancies may have to be designed in an entity-specific fashion. However, it has to be noted that the underlying analysis was selectively implemented to assess similarities in immunopeptidome composition and, by proxy, tumor biology-and does not provide the sensitivity to detect individual shared pan-leukemia antigens. To specifically achieve this goal and evaluate the potential presence of broadly applicable targets for off-the-shelf immunotherapy of multiple HM with a single peptide cocktail, we further sought for individual, shared HLA ligands across the different HM entities.

6.3.3 Overlap analysis identifies a small panel of naturally presented "pan-leukemia" antigens

For the allotype-specific overlap analysis, we assigned the HLA ligands to an HLA type using netMHC and computed the 'cancer-exclusive' presentation of these HLA ligands. The allotype-specific overlap analysis of HM-derived HLA ligands identified 25 unique HLA ligands (A*01:01: 0; A*02:01: 11; A*03:01: 9; A*24:02: 0; B*07:02: 2; B*08:01: 0; B*18:01: 3) showing "cancer-exclusive" presentation on all four HM simultaneously (Figure 6.5A, supplementary Table 6.2), supplementary Figure S3 in Backert et al.[12]). Thus, universal antigen presentation across entity boundaries is a very rare phenomenon, which is further aggravated by the fact that shared antigens typically show low presentation frequencies within the different HM entities (Figure 6.5B). A single pan-leukemia peptide with presentation frequency of more than 20% across all entities was identified for HLA-A*02:01 (POLA2470-480, GLTSTDLLFHL). Furthermore, lineage-specific analysis highlights three myeloid lineage-specific antigens and six lymphatic lineage-specific antigens with presentation frequencies above 20% (with a minimum value of $n \geq 4$ allotype positive samples applied for the calculation of presentation frequencies, Figure 6.5B, supplementary Table 6.2). However, these targets were so far not evaluated for immunogenicity or tumor-specific cytotoxicity. Together, these results clearly argue in favor of entity-specific antigen discovery for T-cell immunotherapy in HM, albeit our analysis identified very few novel, broadly presented candidate targets, which may be amenable for further drug development.

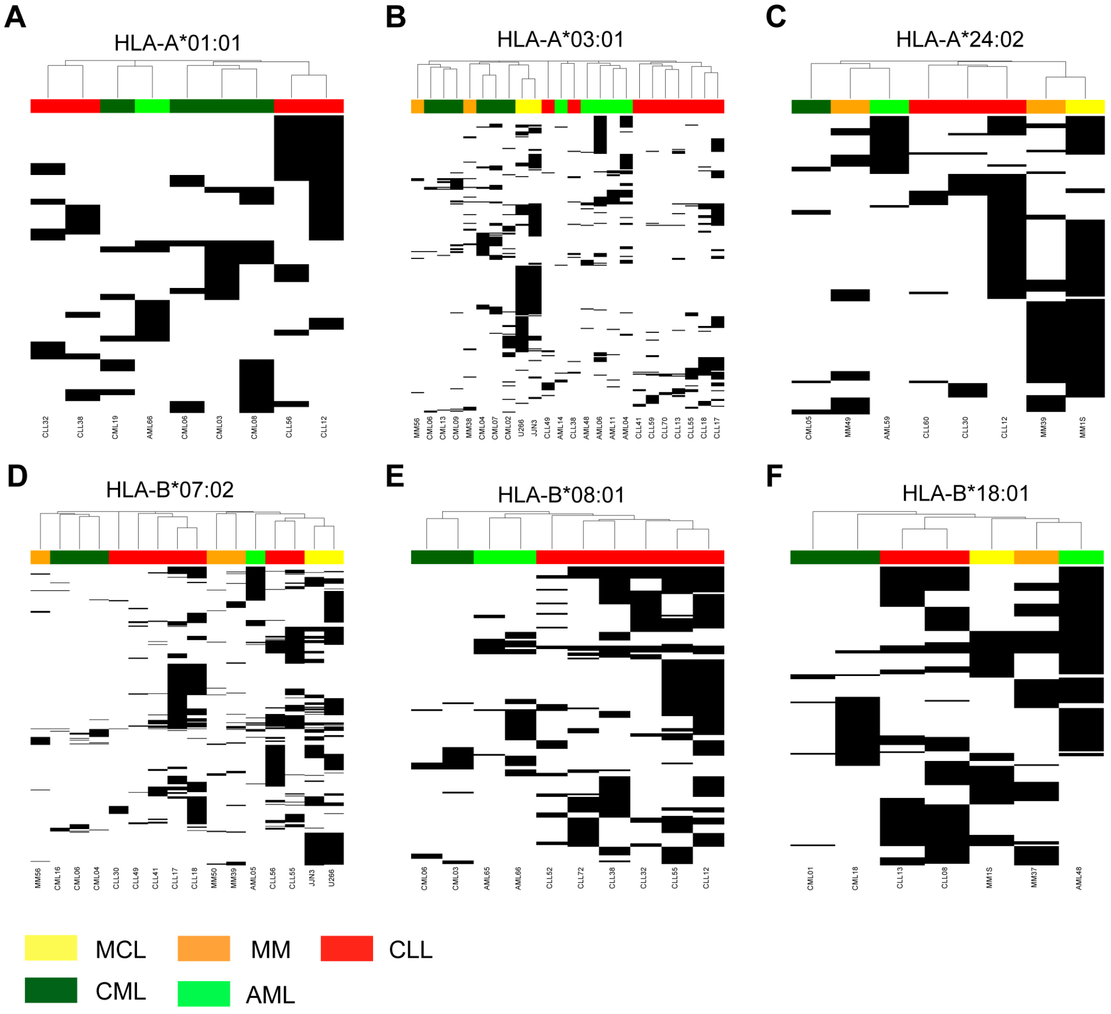


Figure 6.3: HLA allotype-specific clustering analysis of "cancer-exclusive" HLA ligands on hematological cancers. A: "Cancer-exclusive" HLA-A*01:01 ligands identified on AML (light green, n=1), CML (dark green, n=4), CLL (red, n=4) and MM/MCL (orange/yellow, n=0/0) were analyzed by complete linkage clustering based on the Jaccard similarity coefficient. B: "Cancer-exclusive" HLA-A*03:01 ligands identified on AML (light green, n=5), CML (dark green, n=6), CLL (red, n=9) and MM/MCL (orange/yellow, n=2/2) were analyzed by complete linkage clustering based on the Jaccard similarity coefficient. C: "Cancer-exclusive" HLA-A*24:02 ligands identified on AML (light green, n=1), CML (dark green, n=1), CLL (red, n=3) and MM/MCL (orange/yellow, n=2/1) were analyzed by complete linkage clustering based on the Jaccard similarity coefficient. D: "Cancer-exclusive" HLA-B*07:02 ligands identified on AML (light green, n=1), CML (dark green, n=3), CLL (red, n=7) and MM/MCL (orange/yellow, n=3/2) were analyzed by complete linkage clusterin based on the Jaccard similarity coefficient. E: "Cancer-exclusive" HLA-B*08:01 ligands identified on AML (light green, n=2), CML (dark green, n=2), CLL (red, n=6) and MM/MCL (orange/yellow, n=0/0) were analyzed by complete linkage clustering based on the Jaccard similarity coefficient. F: "Cancer-exclusive" HLA-B*18:01 ligands identified on AML (light green, n=1), CML (dark green, n=2), CLL (red, n=2) and MM/MCL (orange/yellow, n=1/1) were analyzed by complete linkage clustering based on the Jaccard similarity coefficient. Reproduced with permission from Backert et al.[12]

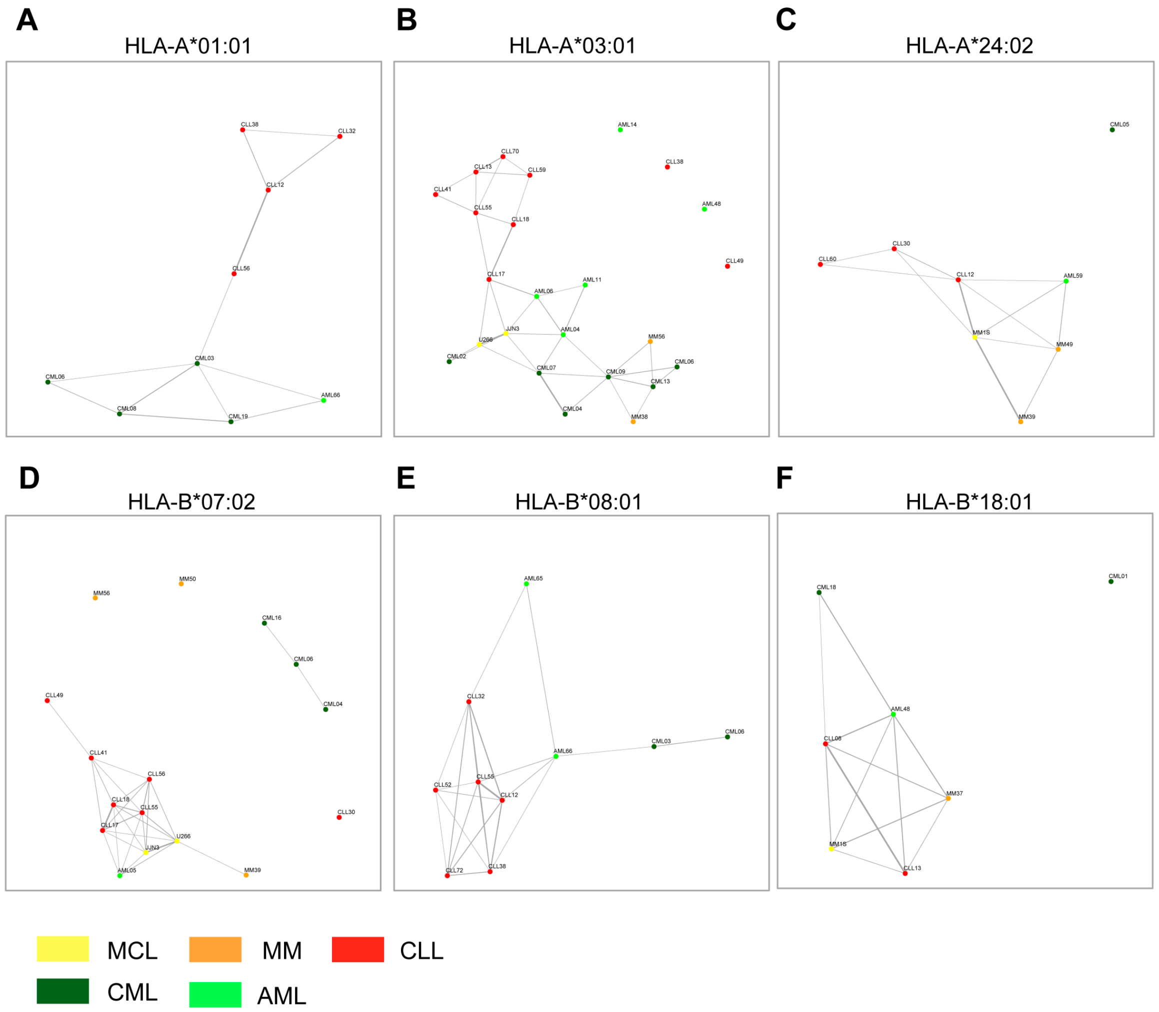
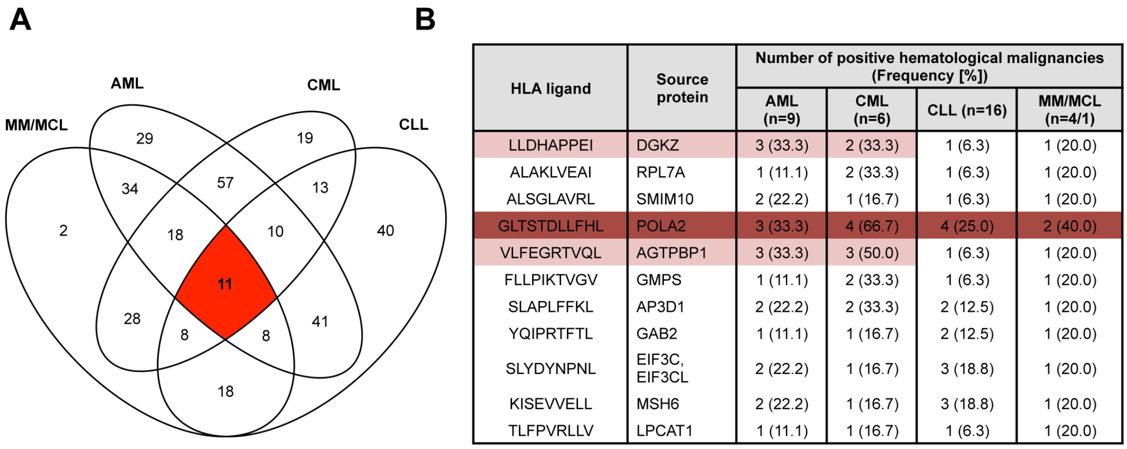


Figure 6.4: HLA allotype-specific Jaccard distance graphs of "cancer-exclusive" HLA ligands on hematological cancers. Samples showing ≥10% Jaccard similiarity of their "cancer-exclusive" HLA peptidomes were linked by edges, with the thickness of the edge positively correlating with the degree of similarity. A: Jaccard distance graph based on "cancer-exclusive" HLA-A*01:01 ligands identified on AML (light green, n=1), CML (dark green, n=4), CLL (red, n=4) and MM/MCL (orange/yellow, n=0/0). B: Jaccard distance graph based on "cancer-exclusive" HLA-A*03:01 ligands identified on AML (light green, n=5), CML (dark green, n=6), CLL (red, n=9) and MM/MCL (orange/yellow, n=2/2). C: Jaccard distance graph based on "cancer-exclusive" HLA-A*24:02 ligands identified on AML (light green, n=1), CML (dark green, n=1), CLL (red, n=3) and MM/MCL (orange/yellow, n=2/1). D: Jaccard distance graph based on "cancer-exclusive" HLA-B*07:02 ligands identified on AML (light green, n=1), CML (dark green, n=3), CLL (red, n=7) and MM/MCL (orange/yellow, n=3/2). E: Jaccard distance graph based on "cancer-exclusive" HLA-B*08:01 ligands identified on AML (light green, n=2), CML (dark green, n=2), CLL (red, n=6) and MM/MCL (orange/yellow, gn=0/0). F: Jaccard distance graph based on "cancer-exclusive" HLA-B*18:01 ligands identified on AML (light green, n=1), CML (dark green, n=2), CLL (red, n=2) and MM/MCL (orange/yellow, n=1/1). Reproduced with permission from Backert et al.[12]



HLA ligand	Source protein	Number of positive hematological malignancies (Frequency [%])			
		AML (n=9)	CML (n=6)	CLL (n=16)	MM/MCL (n=4/1)
LLDHAPPEI	DGKZ	3 (33.3)	2 (33.3)	1 (6.3)	1 (20.0)
ALAKLVEAI	RPL7A	1 (11.1)	2 (33.3)	1 (6.3)	1 (20.0)
ALSGLAVRL	SMIM10	2 (22.2)	1 (16.7)	1 (6.3)	1 (20.0)
GLTSTDLLFHL	POLA2	3 (33.3)	4 (66.7)	4 (25.0)	2 (40.0)
VLFEGRTVQL	AGTPBP1	3 (33.3)	3 (50.0)	1 (6.3)	1 (20.0)
FLLPIKTVGV	GMPS	1 (11.1)	2 (33.3)	1 (6.3)	1 (20.0)
SLAPLFFKL	AP3D1	2 (22.2)	2 (33.3)	2 (12.5)	1 (20.0)
YQIPRTFTL	GAB2	1 (11.1)	1 (16.7)	2 (12.5)	1 (20.0)
SLYDYNPNL	EIF3C, EIF3CL	2 (22.2)	1 (16.7)	3 (18.8)	1 (20.0)
KISEVVELL	MSH6	2 (22.2)	1 (16.7)	3 (18.8)	1 (20.0)
TLFPVRLLV	LPCAT1	1 (11.1)	1 (16.7)	1 (6.3)	1 (20.0)

Figure 6.5: Presentation of "cancer"-exclusive HLA-A*02:01 ligands across different hematological malignancies. A: Overlap analysis of "cancer-exclusive" HLA-A*02:01 ligands identified on AML (n=9), CML (n=6), CLL (n=16) and MM/MCL (n=4/1). B: HLA-A*02:01 restricted "pan-leukemia" antigens identified across all four hematological malignancies. Peptides represented with frequencies ≥20% across all entities are marked in dark red, peptides represented with frequencies ≥20% across entities of the same lineage are marked in light red. A minimum value of n≥4 allotype positive samples was required for the calculation of presentation frequencies. Reproduced with permission from Backert et al.[12]

Table 6.2: "Pan-leukemia" antigens for HLA-A*03:01, HLA-B*07:02, and HLA-B*18:01 identified across all four hematological malignancies. Peptides represented with frequencies ≥20% across all entities are marked in dark red, peptides represented with frequencies ≥20% across entities of the same lineage are marked in light red. A minimum value of $n \geq 4$ allotype positive samples was required for the calculation of presentation frequencies. Reproduced with permission from Backert et al.[12]

HLA ligand	Source protein	Number of positive hematological malignancies (Frequency [%])			
HLA-A*03:01		AML (n=5)	CML (n=6)	CLL (n=9)	MM/MCL (n=2/2)
GLDDPRLEK	LRPAP1	3 (60.0)	3 (50.0)	1 (11.1)	1 (25.0)
GLDPSQRPK	CHTF18	2 (40.0)	1 (16.7)	3 (33.3)	2 (50.0)
KLYEKKLLKL	TMPO	1 (20.0)	1 (16.7)	3 (33.3)	2 (50.0)
KLYPTLVIR	ELP3	4 (80.0)	1 (16.7)	4 (44.4)	1 (25.0)
KMKEALLSIGK	FAM136A	2 (40.0)	1 (16.7)	1 (11.1)	2 (50.0)
RIAKLEAAY	UTP20	1 (20.0)	2 (33.3)	1 (11.1)	2 (50.0)
RLMDRPIFY	GALNS	1 (20.0)	2 (33.3)	1 (11.1)	1 (25.0)
RLNHYVLYK	PSMD3	2 (40.0)	1 (16.7)	2 (22.2)	2 (50.0)
RVVDGKDLTTK	FANCD2	1 (20.0)	1 (16.7)	4 (44.4)	2 (50.0)
HLA-B*07:02		AML (n=1)	CML (n=3)	CLL (n=7)	MM/MCL (n=3/2)
APKRPPSAFF	HMGB1P1, HMGB1, HMGB2	1 (100.0)	1 (33.3)	2 (28.6)	2 (40.0)
SPIEKSGVL	CASC5	1 (100.0)	1 (33.3)	1 (14.3)	2 (40.0)
HLA-B*18:01		AML (n=1)	CML (n=2)	CLL (n=2)	MM/MCL (n=1/1)
DEAPPEHSF	DGKZ	1 (100.0)	1 (50.0)	2 (100.0)	2 (100.0)
DEHHSVNF	RPL7A	1 (100.0)	1 (50.0)	2 (100.0)	2 (100.0)
DETSALKF	SMIM10	1 (100.0)	1 (50.0)	1 (50.0)	1 (50.0)

6.4 Discussion

In the wake of the clinical success of immune checkpoint modulation, it became more and more evident that novel, supplementary therapeutic interventions may be required for a range of malignancies and patient collectives showing low response rates to checkpoint inhibitor monotherapy[75,102]. For this reason therapeutic strategies aimed at inducing antigen-specific anti-tumor T-cell responses have experienced a surge of renewed interest[93]. Common prerequisite to all these approaches is the exact knowledge of clinically effective targets specifically presented on HLA molecules on malignant cells. While the current paradigm views mutation-derived neoepitopes as the most highly effective targets of anti-tumor T-cell responses[103,117,128], this mutation-centric view severely limits the range of malignancies deemed eligible for T-cell immunotherapy[4,131]. Furthermore, mutation-specific strategies would, at least in most cases, be patient-individualized and thus require massively time-and cost intensive target discovery and validation, which currently poses a severe limitation to the number of patients eligible for such approaches[133,138]. Together, these circumstances prompted us and others to comprehensively investigate the non-mutant antigenic landscape presented by HLA molecules on different low-mutational cancer entities[17,36,88,115,119,134]. Importantly, our previous studies in hematological malignancies demonstrated that 1) vast arrays of non-mutated but nevertheless cancer-specific HLA ligands are presented on these cancers, which may be explained by altered antigen processing in malignant cells[70] 2) these peptides are immunogenic and targeted by physiologically occurring T-cell responses in patients[19,135] and 3) that anti-leukemia T-cell responses do correlate with improved patient survival in CLL patients underlining their central role in cancer immune control[67].

Based on these studies we herein conducted a meta-analysis aimed at assessing the particularities of four major HM on the immunopeptidome level and gauged the possibility of identifying a set of universal "pan-leukemia" antigens. In order to evaluate whether the immunopeptidome directly reflects the different biology of the four HM, we assessed the relatedness of all samples on the HLA ligand source protein level. This did not result in grouping of samples according to their respective entities but revealed the existence of a common set of "housekeeping" antigens represented across all entities. This was expected based on our previous studies and is in line with findings of another study on the immunopeptidome of cell lines derived from different tissue origins[17].

Even though our analysis was aimed at removing the impact of different HLA types from the equation by clustering on the level of HLA ligand source proteins, a pattern of HLA allotype-dependent selection for specific source proteins is clearly evident when comparing this cluster analysis with the results of HLA allotype-stratified clustering of HM ligands. Where source protein clustering did not result in coherent grouping of samples along entity lines, the allotype-specific analysis clearly delineated samples according to their entity and lineage of origin, indicating a major influence of sample HLA types on protein representation in the immunopeptidome. Importantly, robust clustering of entity and lineage subgroups was observed for all seven HLA allotypes analyzed in this study. This underscores the robustness of our analytical pipeline and demonstrates that tumor- and lineage specific biology is reflected in the HLA peptidome, which points to the possibility of confidently identifying and assigning pathology purely based on immunopeptidome data (an approach which was previously presented for proteomics data[14]). On the other hand, this finding hints at the limited occurrence of broadly shared antigens, which led us to employ simple overlap analysis as a sensitive means to identify this sparse population of entity-spanning HLA ligands. This verified the rarity of "pan-leukemia" antigens, as such peptides were only detectable for four out of seven HLA allotypes and furthermore typically showed only low frequencies of presentation within the different entities. None of these "pan-leukemia" antigens derives from established tumor-associated genes, which may be explained by a distorted correlation of gene expression and HLA restricted antigen presentation and underscores the importance of direct antigen discovery by mass-spectrometry[17,142]. However -importantly- it also has to be noted that several factors pose central challenges for mass spectrometry based antigen discovery: limited sensitivity and dynamic range as well as the stochasticity of sampling in data-dependent mass spectrometry may lead to false-positive tumor-exclusive detection.

Our central finding is the presence of entity- and lineage-specific dividing lines, which may vitally impede the development of entity-spanning antigen-specific compounds. This strongly argues in favor of entity-specific approaches for the development of antigen-specific T-cell immunotherapy in hematological malignancies.

Chapter 7

Conclusion and Outlook

Immunotherapies are among the most promising recently developed treatment options for cancer[99]. However, the development of immunotherapies is based on the identification of therapeutic targets. One group of potential targets are antigens that are presented by HLA, called epitopes. Mass spectrometry-based immunoprecipitation allows measuring the immunopeptidome (i.e. all HLA-bound peptides) of a tissue or cancer type[19,36,66,70,135]. One major concern in immunotherapies is cross-reactivity, which occurs if the immune system attacks other benign or healthy parts of the body besides the target tissue. To control for potential cross-reactivity of new targets, the immunopeptidome of healthy tissues is needed.

The HLA Ligand Atlas

The anticipated user base of the HLA Ligand Atlas consists of immunologists. Therefore, the raw MS data and the result of subsequent analyses have to be presented, so that they can quickly and comfortably access the information they are invested in. To this end, we developed a browser-based web interface, which allows wet-lab scientists to perform simple analyses that build on preprocessed MS data and the search for specific peptides and proteins. The HLA Ligand Atlas also provides tissue- or HLA-specifc lists of the peptides that we identified with our automated QC and processing pipeline. Currently, our dataset contains 435 mass spectrometry runs, which encompass up to 5,000 peptides each. As the backend for the HLA Ligand Atlas, we use an MySQL database, which provides fast access to the data. Because the access to the HLA Ligand Atlas is restricted until publication and only available for test users, we cannot foresee how much traffic will occur when many users access the database simultaneously. We assume that minor changes in both the database and the interface have to be made to allow a smooth user experience. In addition, we hope to enhance the interactive analysis and statistics provided trough the HLA Ligand Atlas based on user requests and feedback. With this feedback, we will implement analyses to answer frequent questions, which will add value to the website. Currently, the HLA Ligand Atlas only contains benign samples. However, adding malignant tissues would add significant new content and would allow answering cancer-related questions. However, adding these cancer samples will require significant changes to the metadata structure of the database.

Analysis of the benign tissue immunopeptidome

In addition to the developing of the HLA Ligand Atlas, we perfomred a detailed analysis of the benign immunopeptidome data that is contained in the database. Before we analyzed the immunopeptidome data, we performed multiple quality control steps. First, we performed a time-series experiment, to ensure the stability of the immunopeptidome over time at 4 °C. Furthermore, we discovered that both samples from stomach tissue were pepsin digested. However, we could not determine when (before or while immunoprecipitation) and where (*in vivo* or *in vitro*) the digestion occurred. Further experiments are needed to verify that this pepsin-specific cleavage is not an immunoprecipitation artifact. Finally, we identify typical protein contaminants based on their coverage. Although we removed these contaminants, peptide contaminants and known peptide impurities may still be confounders of the analysis. To identify and remove these contaminants, a blacklist could be compiled based on expert knowledge, similar to the Crapome in proteomics studies[83]. After we had excluded low quality data from the HLA Ligand Atlas, we showed that the open-source software Comet[38,39] performed better than the proprietary software Sequest HT on immunopeptidome data. However, we did not compare Comet and Sequest HT with other identification software such as Mascot[94], OMMSA[45], MaxQuant[30], or MSGF+[63], which can therefore not be judged. Next, we assessed common properties of HLA class I and class II ligands. We observed that the length variation varies between HLA types. A deconvolution with Gibbs clustering yielded results that are similar to those shown by Abelin et al. (2017)[1]. In addition, we calculated the peptide and protein overlap between HLA class I and class II and found minor overlaps on sample level and a large protein overlap when we combined all samples from one class. Furthermore, we showed that HLA class II peptides often contain length variations of the same core peptide (nested peptides). In contrast, HLA class I peptides had only few length variants. In the last part of the analysis of the HLA Ligand Atlas data, we compared inter- and intra-individual differences. We found that the HLA type has a major impact on both the presented peptides and the proteins. In both cases, an unsupervised hierarchical cluster analysis grouped samples by individuals rather by tissues. Only when we subtracted each individual's background, minor tissue clusters were identified. Although the individuals had only a small HLA type overlap between, we identified tissue-specific proteins. In total we found 12 tissue-specific proteins for HLA class I and 17 for HLA class II. A comparison of these proteins and their protein expression obtained from the ProteomicsDB[143] showed that only one of these proteins was also tissue-specific on the protein expression level. This implies that more data is needed to find clearly tissue-specific HLA proteins. This data should be obtained from individuals with more overlapping HLA alleles. Furthermore, the number of different tissue types should be expanded. In general, the aggregation of more samples is the next crucial step before the final publication of the HLA Ligand Atlas. Furthermore, the integration of public data, like links to the IMGT[76] (peptides and HLA) or Uniprot[124] (proteins) database or integration of information contained in SYFPEITHI[97] would provide new insights to the user.

A meta-analysis of HLA peptidome composition in different hematological entities

In the last chapter of this thesis, we describe a meta-analysis of the HLA immunopeptidome of different hematological entities. The chapter explains how tumor-associated antigens can be found and how an analysis of a larger cancer dataset can be performed. This analysis used a large benign dataset similar to the HLA Ligand Atlas as a negative dataset to avoid cross-reactivity and to remove benign background presentation.

The study was motivated by new breakthroughs in therapies using immune checkpoint modulation[75,102]. However, these new therapies are not suitable for all types of malignancies and furthermore not all patients benefit from them. Therefore, other therapies have to be developed. These new approaches have in common that they need a tumor-specific target, which can be identified using HLA immunopeptidome analysis. We analyzed 83 HLA immunopeptidome samples of four different hematologic malignancies (19 AML, 16 CLM, 35 CLL, and 13 MM/MCL), which resulted in 40,361 unique HLA class I peptides. This meta-analysis aimed to find "pan-leukemia" antigens and to describe shared properties between the malignancies. These properties were analyzed using the source proteins of the identified peptides. However, neither a hierarchical clustering nor a graph- and Jaccard-similarity-score-based approach resulted in a clustering of hematological entities but revealed a group of common "housekeeping" antigens presented across all malignancies. Next, we assigned our immunopeptidome to the patients HLA types, and rerun the analysis only on peptides assigned to and patients with a specific HLA type. This HLA type-specific analysis resulted in a clustering of the hematological malignancies for all seven analyzed HLA types. Furthermore, it showed evidence that the HLA type has a strong influence on which proteins are presented. Next, we searched for "pan-leukemia" antigens, using again the HLA type assigned peptides and a simple overlap analysis. We found in only four of the seven analyzed HLA types broadly shared antigens. None of the found shared antigens were known tumor-associated genes defined in the literature. However, this could be explained by the weak correlation between gene expression and HLA presentation and demonstrates the importance of immunoprecipitation-based antigen discovery[17,142]. This analysis showed entity-lineage-specific dividing lines, which might be of importance when we try to identify "pan-leukemia"-antigens. However, this finding also supports the continuation of the approach of finding entity-specific antigens for T-cell immunotherapy in hematological malignancies.

General conclusions and outlook

The development of new immunotherapy treatment options relies on the discovery of new target antigens. These antigens can be identified by analyzing the HLA immunopeptidome. In the context of this thesis, we presented the HLA Ligand Atlas: a large publicly available database and web page for benign immunopeptidome data. It was developed with the aim of providing easy access and fast data queries through an intuitive web interface. During the development of the HLA Ligand Atlas, the SysteMHC Atlas[114] has been published, which also provides access to many different immunopeptidome datasets. However, the interface of SysteMHC does only allow querying for peptides, proteins and HLAs, and does not support further information besides a spectral library. Furthermore, SysteMHC aims to collect all immunopeptidome datasets and to process them in consistent way. It provides access to them, but does not allow to analyze the data in the interface. We think that both databases can coexist, since they focus on different features.

In addition, this thesis contains an extensive analysis of the benign dataset. It describes quality control criteria and properties of HLA class I and class II peptides and proteins. Furthermore, an inter- and intra-individual analysis described herein shows a clear clustering of the immunopeptidome data by individuals. Finally, we identified 27 tissue-specific antigens. Although the presented dataset is to the best of our knowledge the largest benign dataset available, it certainly has to be enlarged. The acquisition of new samples should be focused on individuals with similar HLA types. In addition, new tissue types should be collected to capture the full complexity of the HLA immunopeptidome. This new data will support the search for new targets for cancer treatment, but also the development of new immunoinformatics software. For example, the recently published new version of netMHCpan-4.0[56] uses HLA ligand data to train and to enhance the binding prediction. The dataset presented here can be included in such algorithms and will allow better binding predictions. However, since we focus on the acquisition of samples with similar HLA types, the data will be more likely to improve predictions of the binding of specific HLA alleles than the prediction across all HLA alleles. Furthermore, new methods for the identification of HLA peptides from MS data can be developed with the provided raw files. This would be especially interesting, since the immunopeptidome analysis differs from the standard peptidomics analysis (e.g., enzymatic digestion).

We characterized the immunopeptidome from four different hematologic malignancies. This analysis resulted in a clustering along entity-specific dividing lines when only peptides belonging to one HLA type were considered at a time. In addition, we defined "pan-leukemia" antigens. However, the complete meta-analysis showed that entity-specific antigens were to be favored over "pan-leukemia" ones. Therefor, entity-specific research projects should be conducted in the future to discover new hematologic malignancy targets.

To conclude, we presented and analyzed a large dataset of benign HLA immunopeptidome samples, which hopefully helps future researchers to discover new immunotherapy targets and to develop new, safe, and effective immunotherapies for cancer patients.

Bibliography

[1] J. G. Abelin, D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, J. Stevens, W. Lane, G. L. Zhang, T. M. Eisenhaure, K. R. Clauser, N. Hacohen, M. S. Rooney, S. A. Carr, and C. J. Wu. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*, 46(2):315–326, feb 2017. 3, 75, 100

[2] S. Abiteboul, R. Hull, and V. Vianu. Foundations of Databases. *Journal of Chemical Information and Modeling*, 53(9):1689–1699, 2013. 25, 26

[3] A. Admon and M. Bassani-Sternberg. The Human Immunopeptidome Project, a Suggestion for yet another Postgenome Next Big Thing. *Molecular & Cellular Proteomics*, 10(10):O111.011833–O111.011833, 2011. 1

[4] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. Andrew Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, aug 2013. 86, 97

[5] M. Andreatta, B. Alvarez, and M. Nielsen. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic acids research*, apr 2017. 24

[6] M. Andreatta, O. Lund, and M. Nielsen. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics (Oxford, England)*, 29(1):8–14, jan 2013. 24

[7] M. Andreatta and M. Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, feb 2016. 13

[8] S. M. Ansell, S. A. Hurvitz, P. A. Koenig, B. R. LaPlant, B. F. Kabat, D. Fernando, T. M. Habermann, D. J. Inwards, M. Verma, R. Yamada, C. Erlichman, I. Lowy, and J. M. Timmerman. Phase I Study of Ipilimumab, an Anti-CTLA-4 Monoclonal Antibody, in Patients with Relapsed and Refractory B-Cell Non-Hodgkin Lymphoma. *Clinical Cancer Research*, 15(20):6446–6453, oct 2009. 85

[9] S. M. Ansell, A. M. Lesokhin, I. Borrello, A. Halwani, E. C. Scott, M. Gutierrez, S. J. Schuster, M. M. Millenson, D. Cattry, G. J. Freeman, S. J. Rodig, B. Chapuy, A. H. Ligon, L. Zhu, J. F. Grosso, S. Y. Kim, J. M. Timmerman, M. A. Shipp, and P. Armand. PD-1 Blockade with Nivolumab in Relapsed or Refractory Hodgkin's Lymphoma. *New England Journal of Medicine*, 372(4):311–319, jan 2015. 4, 85

[10] P. Armand, M. A. Shipp, V. Ribrag, J.-M. Michot, P. L. Zinzani, J. Kuruvilla, E. S. Snyder, A. D. Ricart, A. Balakumaran, S. Rose, and C. H. Moskowitz. Programmed Death-1 Blockade With Pembrolizumab in Patients With Classical Hodgkin Lymphoma After Brentuximab Vedotin Failure. *Journal of Clinical Oncology*, jun 2016. 4, 85

[11] M. Baccarani. Evolving concepts in the management of chronic myeloid leukemia: recommendations from an expert panel on behalf of the European LeukemiaNet. *Blood*, 108(6):1809–1820, sep 2006. 11

[12] L. Backert, D. Johannes Kowalewski, S. Walz, H. Schuster, C. Berlin, M. Christoph Neidert, M. Schemionek, T. H. Brummendorf, V. Vucinic, D. Niederwieser, L. Kanz, H. Rainer Salih, O. Kohlbacher, K. Weisel, H.-G. Rammensee, S. Stevanovic, J. Sarah Walz, L. Backert, D. Johannes Kowalewski, S. Walz, H. Schuster, C. Berlin, M. Christoph Neidert, M. Schemionek, T. H. Brummendorf, V. Vucinic, D. Niederwieser, L. Kanz, H. Rainer Salih, O. Kohlbacher, K. Weisel, H.-G. Rammensee, S. Stevanovic, and J. Sarah Walz. A meta-analysis of HLA peptidome composition in different hematological entities: Entity-specific dividing lines and "pan-leukemia" antigens. *Oncotarget*, 5(0), jan 2017. 78, 85, 88, 91, 92, 93, 94, 95, 96, 150

[13] L. Backert and O. Kohlbacher. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Medicine*, 7(1):119, dec 2015. 12, 22, 23

[14] J. Balog, L. Sasi-Szabo, J. Kinross, M. R. Lewis, L. J. Muirhead, K. Veselkov, R. Mirnezami, B. Dezso, L. Damjanovich, A. Darzi, J. K. Nicholson, and Z. Takats. Intraoperative Tissue Identification Using Rapid Evaporative Ionization Mass Spectrometry. *Science Translational Medicine*, 5(194):194ra93–194ra93, jul 2013. 98

[15] M. Bassani-Sternberg, E. Bräunlein, R. Klar, T. Engleitner, P. Sinitcyn, S. Audehm, M. Straub, J. Weber, J. Slotta-Huspenina, K. Specht, M. E. Martignoni, A. Werner, R. Hein, D. H. Busch, C. Peschel, R. Rad, J. Cox, M. Mann, and A. M. Krackhardt. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature Communications*, 7(November):13404, nov 2016. 2, 3, 31, 53, 78

[16] M. Bassani-Sternberg, C. Chong, P. Guillaume, M. Solleder, H. S. Pak, P. O. Gannon, L. E. Kandalaft, G. Coukos, and D. Gfeller. Deciphering HLA-I motifs across HLA peptidomes improves neoantigen predictions and identifies allostery regulating HLA specificity. *PLoS computational biology*, 13(8):e1005725, 2017. 53

[17] M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, and M. Mann. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular & cellular proteomics : MCP*, 14(3):658–73, mar 2015. 68, 97, 98, 101

[18] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. 20

[19] C. Berlin, D. J. Kowalewski, H. Schuster, N. Mirza, S. Walz, M. Handel, B. Schmid-Horch, H. R. Salih, L. Kanz, H.-G. Rammensee, S. Stevanović, and J. S. Stickel. Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. *Leukemia*, 29(3):647–59, mar 2015. 1, 4, 86, 87, 97, 99

[20] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, jan 2000. 9

[21] M. Bleakley and S. R. Riddell. Exploiting T cells specific for human minor histocompatibility antigens for therapy of leukemia. *Immunology and Cell Biology*, 89(3):396–407, mar 2011. 4, 86

[22] J. Brahmer, K. L. Reckamp, P. Baas, L. Crinò, W. E. Eberhardt, E. Poddubskaya, S. Antonia, A. Pluzanski, E. E. Vokes, E. Holgado, D. Waterhouse, N. Ready, J. Gainor, O. Arén Frontera, L. Havel, M. Steins, M. C. Garassino, J. G. Aerts, M. Domine, L. Paz-Ares, M. Reck, C. Baudelet, C. T. Harbison, B. Lestini, and D. R. Spigel. Nivolumab versus Docetaxel in Advanced Squamous-Cell NonâĂŞSmall-Cell Lung Cancer. *New England Journal of Medicine*, 373(2):123–135, jul 2015. 4, 85

[23] F. Caligaris-Cappio and T. J. Hamblin. B-cell chronic lymphocytic leukemia: a bird of a different feather. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 17(1):399–408, jan 1999. 11

[24] P. Charoentong, F. Finotello, M. Angelova, C. Mayer, M. Efremova, D. Rieder, H. Hackl, and Z. Trajanoski. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Reports*, 18(1):248–262, 2017. 31, 52

[25] H. Chen and P. C. Boutros. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, 12(1):35, 2011. 89

[26] CLL Trialists' Collaborative Group. Chemotherapeutic options in chronic lymphocytic leukemia: a meta-analysis of the randomized trials. CLL Trialists' Collaborative Group. *Journal of the National Cancer Institute*, 91(10):861–8, may 1999. 11

[27] D. K. Cole, A. M. Bulek, G. Dolton, A. J. Schauenberg, B. Szomolay, W. Rittase, A. Trimby, P. Jothikumar, A. Fuller, A. Skowera, J. Rossjohn, C. Zhu, J. J. Miles, M. Peakman, L. Wooldridge, P. J. Rizkallah, and A. K. Sewell. Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity. *Journal of Clinical Investigation*, 126(6):2191–2204, may 2016. 9

[28] J. E. Cortes, M. Talpaz, S. O'Brien, S. Faderl, G. Garcia-Manero, A. Ferrajoli, S. Verstovsek, M. B. Rios, J. Shan, and H. M. Kantarjian. Staging of chronic myeloid leukemia in the imatinib era: an evaluation of the World Health Organization proposal. *Cancer*, 106(6):1306–15, mar 2006. 11

[29] J. Couzin-Frankel. Cancer Immunotherapy. *Science*, 342(6165):1432–1433, dec 2013. 4, 85

[30] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, dec 2008. 3, 16, 72, 83, 100

[31] R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, jun 2004. 3, 16, 32

[32] G. Csárdi and T. Nepusz. The igraph software package for complex network research, 2006. 89

[33] M. Di Marco, H. Schuster, L. Backert, M. Ghosh, H.-G. Rammensee, and S. Stevanović. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. *The Journal of Immunology*, page ji1700938, 2017. 53

[34] H. Döhner, S. Stilgenbauer, A. Benner, E. Leupolt, A. Kröber, L. Bullinger, K. Döhner, M. Bentz, and P. Lichter. Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 343(26):1910–1916, dec 2000. 11

[35] P. Dönnes and A. Elofsson. Prediction of MHC class I binding peptides, using SVMHC. *BMC bioinformatics*, 3:25, sep 2002. 22

[36] V. Dutoit, C. Herold-Mende, N. Hilf, O. Schoor, P. Beckhove, J. Bucher, K. Dorsch, S. Flohr, J. Fritsche, P. Lewandrowski, J. Lohr, H.-G. Rammensee, S. Stevanovic, C. Trautwein, V. Vass, S. Walter, P. R. Walker, T. Weinschenk, H. Singh-Jasuja, and P.-Y. Dietrich. Exploiting the glioblastoma peptidome to discover novel tumour-associated antigens for immunotherapy. *Brain*, 135(4):1042–1054, apr 2012. 1, 97, 99

[37] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–10, jan 2002. 2

[38] J. K. Eng, M. R. Hoopmann, T. A. Jahan, J. D. Egertson, W. S. Noble, and M. J. MacCoss. A Deeper Look into Comet - Implementation and Features. *Journal of The American Society for Mass Spectrometry*, 26(11):1865–1874, nov 2015. 3, 16, 18, 19, 69, 71, 72, 100

[39] J. K. Eng, T. A. Jahan, and M. R. Hoopmann. Comet: An open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24, jan 2013. 3, 16, 32, 69, 100

[40] J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, nov 1994. 3, 16, 17

[41] F. H. Epstein, S. Faderl, M. Talpaz, Z. Estrov, S. O'Brien, R. Kurzrock, and H. M. Kantarjian. The Biology of Chronic Myeloid Leukemia. *New England Journal of Medicine*, 341(3):164–172, jul 1999. 11

[42] M. Feldhahn, P. Dönnes, P. Thiel, and O. Kohlbacher. FRED - A framework for T-cell epitope detection. *Bioinformatics*, 25(20):2758–2759, 2009. 89

[43] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5):E359–E386, mar 2015. 1

[44] A. L. Garfall, M. V. Maus, W.-T. Hwang, S. F. Lacey, Y. D. Mahnke, J. J. Melenhorst, Z. Zheng, D. T. Vogl, A. D. Cohen, B. M. Weiss, K. Dengel, N. D. Kerr, A. Bagg, B. L. Levine, C. H. June, and E. A. Stadtmauer. Chimeric Antigen Receptor T Cells against CD19 for Multiple Myeloma. *New England Journal of Medicine*, 373(11):1040–1047, 2015. 4, 86

[45] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research*, 3(5):958–964, oct 2004. 72, 83, 100

[46] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. 24

[47] C. Gerstner, A. Dubnovitsky, C. Sandin, G. Kozhukh, H. Uchtenhagen, E. A. James, J. Rönnelid, A. J. Ytterberg, J. Pieper, E. Reed, C. Tandre, M. Rieck, R. A. Zubarev, L. Rönnblom, T. Sandalova, J. H. Buckner, A. Achour, and V. Malmström. Functional and Structural Characterization of a Novel HLA-DRB1*04:01-Restricted α-Enolase T Cell Epitope in Rheumatoid Arthritis. *Frontiers in Immunology*, 7:494, nov 2016. 9

[48] S. N. Gettinger, L. Horn, L. Gandhi, D. R. Spigel, S. J. Antonia, N. A. Rizvi, J. D. Powderly, R. S. Heist, R. D. Carvajal, D. M. Jackman, L. V. Sequist, D. C. Smith, P. Leming, D. P. Carbone, M. C. Pinder-Schenck, S. L. Topalian, F. S. Hodi, J. A. Sosman, M. Sznol, D. F. McDermott, D. M. Pardoll, V. Sankar, C. M. Ahlers, M. Salvati, J. M. Wigginton, M. D. Hellmann, G. D. Kollia, A. K. Gupta, and J. R. Brahmer. Overall Survival and Long-Term Safety of Nivolumab (Anti-Programmed Death 1 Antibody, BMS-936558, ONO-4538) in Patients With Previously Treated Advanced Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology*, 33(18):2004–2012, jun 2015. 4, 85

[49] A. Gros, M. R. Parkhurst, E. Tran, A. Pasetto, P. F. Robbins, S. Ilyas, T. D. Prickett, J. J. Gartner, J. S. Crystal, I. M. Roberts, K. Trebska-McGowan, J. R. Wunderlich, J. C. Yang, and S. A. Rosenberg. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nature medicine*, 22(4):433–8, apr 2016. 85

[50] M. M. Gubin, X. Zhang, H. Schuster, E. Caron, J. P. Ward, T. Noguchi, Y. Ivanova, J. Hundal, C. D. Arthur, W.-J. Krebber, G. E. Mulder, M. Toebes, M. D. Vesely, S. S. K. Lam, A. J. Korman, J. P. Allison, G. J. Freeman, A. H. Sharpe, E. L. Pearce, T. N. Schumacher, R. Aebersold, H.-G. Rammensee, C. J. M. Melief, E. R. Mardis, W. E. Gillanders, M. N. Artyomov, and R. D. Schreiber. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*, 515(7528):577–81, nov 2014. 85

[51] R. D. Guest, N. Kirillova, S. Mowbray, H. Gornall, D. G. Rothwell, E. J. Cheadle, E. Austin, K. Smith, S. M. Watt, K. Kühlcke, N. Westwood, F. Thistlethwaite, R. E. Hawkins, and D. E. Gilham. Definition and application of good manufacturing process-compliant production of

CEA-specific chimeric antigen receptor expressing T-cells for phase I/II clinical trial. *Cancer Immunology, Immunotherapy*, 63(2):133–145, feb 2014. 86

[52] S. P. Haen and H.-G. Rammensee. The repertoire of human tumor-associated epitopes - identification and selection of antigens and their application in clinical trials. *Current Opinion in Immunology*, 25(2):277–283, apr 2013. 1

[53] D. W. Huang, B. T. Sherman, and R. a. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, dec 2008. 92

[54] W. Hugo, J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, E. Seja, S. Lomeli, X. Kong, M. C. Kelley, J. A. Sosman, D. B. Johnson, A. Ribas, and R. S. Lo. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, 165(1):35–44, 2016. 32

[55] G. Juliusson, V. Lazarevic, A.-S. Horstedt, O. Hagberg, and M. Hoglund. Acute myeloid leukemia in the real world: why population-based registries are needed. *Blood*, 119(17):3890–3899, apr 2012. 11

[56] V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, and M. Nielsen. NetMHCpan-4.0: Improved PeptideâĂŞMHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology*, page ji1700893, 2017. 1, 102

[57] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923–5, nov 2007. 20, 21, 88

[58] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, jan 2008. 3, 20

[59] L. Kall, J. D. Storey, and W. S. Noble. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*, 24(16):i42–i48, aug 2008. 20

[60] E. Karosiene, M. Rasmussen, T. Blicher, O. Lund, S. Buus, and M. Nielsen. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10):711–24, oct 2013. 22, 24

[61] B. Keil. *Specificity of Proteolysis*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992. 64

[62] M. S. Khodadoust, N. Olsson, L. E. Wagar, O. A. W. Haabeth, B. Chen, K. Swaminathan, K. Rawson, C. L. Liu, D. Steiner, P. Lund, S. Rao, L. Zhang, C. Marceau, H. Stehr, A. M. Newman, D. K. Czerwinski, V. E. H. Carlton, M. Moorhead, M. Faham, H. E. Kohrt, J. Carette, M. R. Green, M. M. Davis, R. Levy, J. E. Elias, and A. A. Alizadeh. Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature*, 543(7647):723–727, mar 2017. 3, 53

[63] S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5:5277, oct 2014. 72, 83, 100

[64] H. T. Kissick, M. G. Sanda, L. K. Dunn, and M. S. Arredouani. Immunization with a Peptide Containing MHC Class I and II Epitopes Derived from the Tumor Antigen SIM2 Induces an Effective CD4 and CD8 T-Cell Response. *PLoS ONE*, 9(4):e93231, apr 2014. 3, 84

[65] A. A. Klammer and M. J. MacCoss. Effects of Modified Digestion Schemes on the Identification of Proteins from Complex Mixtures. *Journal of Proteome Research*, 5(3):695–700, mar 2006. 20

[66] D. J. Kowalewski, H. Schuster, L. Backert, C. Berlin, S. Kahn, L. Kanz, H. R. Salih, H.-G. Rammensee, S. Stevanovic, and J. S. Stickel. HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). *Proceedings of the National Academy of Sciences*, 112(2):E166–E175, jan 2015. 1, 4, 87, 99

[67] D. J. Kowalewski, H. Schuster, L. Backert, C. Berlin, S. Kahn, L. Kanz, H. R. Salih, H.-G. Rammensee, S. Stevanovic, and J. S. Stickel. HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). *Proceedings of the National Academy of Sciences*, 112(2):E166–E175, jan 2015. 3, 53, 86, 97

[68] D. J. Kowalewski, H. Schuster, L. Backert, C. Berlin, S. Kahn, L. Kanz, H. R. Salih, H.-G. Rammensee, S. Stevanovic, and J. S. Stickel. HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). *Proceedings of the National Academy of Sciences*, 112(2):E166–E175, 2015. 78

[69] D. J. Kowalewski and S. Stevanović. Biochemical large-scale identification of MHC class I ligands. *Methods in molecular biology (Clifton, N.J.)*, 960:145–157, 2013. 9, 31, 87

[70] D. J. Kowalewski, S. Stevanovic, H.-G. Rammensee, and J. S. Stickel. Antileukemia T-cell responses in CLL - We don't need no aberration. *OncoImmunology*, 4(7):e1011527, jul 2015. 1, 86, 97, 99

[71] D. J. Kowalewski, S. Walz, L. Backert, H. Schuster, O. Kohlbacher, K. Weisel, S. M. Rittig, L. Kanz, H. R. Salih, H.-G. Rammensee, S. Stevanović, and J. S. Stickel. Carfilzomib alters the HLA-presented peptidome of myeloma cells and impairs presentation of peptides with aromatic C-termini. *Blood Cancer Journal*, 6(4):e411, apr 2016. 3, 53

[72] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel. RNA-Seq Atlas - A reference database for gene expression profiling in normal tissue by next generation sequencing. *Bioinformatics (Oxford, England)*, 28(8):1184–1185, 2012. 53

[73] W. M. Kuehl and P. L. Bergsagel. Multiple myeloma: evolving genetic events and host interactions. *Nature Reviews Cancer*, 2(3):175–187, mar 2002. 11

[74] R. A. Kyle and S. V. Rajkumar. Multiple Myeloma. *New England Journal of Medicine*, 351(18):1860–1873, oct 2004. 11

[75] D. T. Le, J. N. Uram, H. Wang, B. R. Bartlett, H. Kemberling, A. D. Eyring, A. D. Skora, B. S. Luber, N. S. Azad, D. Laheru, B. Biedrzycki, R. C. Donehower, A. Zaheer, G. A. Fisher, T. S. Crocenzi, J. J. Lee, S. M. Duffy, R. M. Goldberg, A. de la Chapelle, M. Koshiji, F. Bhaijee, T. Huebner, R. H. Hruban, L. D. Wood, N. Cuka, D. M. Pardoll, N. Papadopoulos, K. W. Kinzler, S. Zhou, T. C. Cornish, J. M. Taube, R. A. Anders, J. R. Eshleman, B. Vogelstein, and L. A. Diaz. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England journal of medicine*, 372(26):2509–20, jun 2015. 97, 101

[76] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Regnier, F. Ehrenmann, G. Lefranc, and P. Duroux. IMGT, the international ImMunoGeneTics information system. *Nucleic acids research*, 37(Database issue):D1006–12, jan 2009. 1, 2, 101

[77] A. M. Lesokhin, P. Armand, E. C. Scott, A. Halwani, M. Gutierrez., M. M. Millenson, A. D. Cohen, S. J. Schuster, D. Lebovic, M. V. Dhodapkar, D. Avigan, B. Chapuy, A. H. Ligon, S. J. Rodig, D. Catt, and S. M. Ansell. Preliminary Results of a Phase I Study of Nivolumab (BMS-936558) in Patients with Relapsed or Refractory Lymphoid Malignancies. *Blood ASH Abstract 124:291*, 2014. 85

[78] C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, and M. Nielsen. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic acids research*, 36(Web Server issue):W509–12, jul 2008. 22

[79] K. Ma, O. Vitek, and A. I. Nesvizhskii. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*, 13(Suppl 16):S1, 2012. 32

[80] S. L. Maude, N. Frey, P. A. Shaw, R. Aplenc, D. M. Barrett, N. J. Bunin, A. Chew, V. E. Gonzalez, Z. Zheng, S. F. Lacey, Y. D. Mahnke, J. J. Melenhorst, S. R. Rheingold, A. Shen, D. T. Teachey, B. L. Levine, C. H. June, D. L. Porter, and S. A. Grupp. Chimeric antigen receptor T cells for sustained remissions in leukemia. *The New England journal of medicine*, 371(16):1507–17, oct 2014. 4, 86

[81] D. F. McDermott, C. G. Drake, M. Sznol, T. K. Choueiri, J. D. Powderly, D. C. Smith, J. R. Brahmer, R. D. Carvajal, H. J. Hammers, I. Puzanov, F. S. Hodi, H. M. Kluger, S. L. Topalian, D. M. Pardoll, J. M. Wigginton, G. D. Kollia, A. Gupta, D. McDonald, V. Sankar, J. A. Sosman, and M. B. Atkins. Survival, Durable Response, and Long-Term Safety in Patients With Previously Treated Advanced Renal Cell Carcinoma Receiving Nivolumab. *Journal of Clinical Oncology*, 33(18):2013–2020, jun 2015. 4, 85

[82] A. Meier and M. Kaufmann. *SQL- & NoSQL-Datenbanken*. eXamen.press. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016. 25, 26, 28

[83] D. Mellacheruvu, Z. Wright, A. L. Couzens, J.-P. Lambert, N. A. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardiu, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. Al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z.-Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A.-C. Gingras, and A. I. Nesvizhskii. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature Methods*, 10(8):730–736, jul 2013. 67, 100

[84] B. T. Messmer, D. Messmer, S. L. Allen, J. E. Kolitz, P. Kudalkar, D. Cesar, E. J. Murphy, P. Koduru, M. Ferrarini, S. Zupo, G. Cutrona, R. N. Damle, T. Wasil, K. R. Rai, M. K. Hellerstein, and N. Chiorazzi. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. *Journal of Clinical Investigation*, 115(3):755–764, mar 2005. 11

[85] R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, apr 2002. 20

[86] R. J. Motzer, B. Escudier, D. F. McDermott, S. George, H. J. Hammers, S. Srinivas, S. S. Tykodi, J. A. Sosman, G. Procopio, E. R. Plimack, D. Castellano, T. K. Choueiri, H. Gurney, F. Donskov, P. Bono, J. Wagstaff, T. C. Gauler, T. Ueda, Y. Tomita, F. A. Schutz, C. Kollmannsberger, J. Larkin, A. Ravaud, J. S. Simon, L.-A. Xu, I. M. Waxman, and P. Sharma. Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *New England Journal of Medicine*, 373(19):1803–1813, nov 2015. 4, 85

[87] K. M. Murphy, P. Travers, and M. Walport. *Janeway's Immunobiology*. Garland Science, 8th edition, 2011. 5

[88] A. Neumann, H. Hörzer, N. Hillen, K. Klingel, B. Schmid-Horch, H. J. Bühring, H. G. Rammensee, H. Aebert, and S. Stevanović. Identification of HLA ligands and T-cell epitopes for immunotherapy of lung cancer. *Cancer Immunology, Immunotherapy*, 62(9):1485–1497, 2013. 97

[89] M. Nielsen and M. Andreatta. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine*, 8(1):33, dec 2016. 4, 65, 66, 88

[90] M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund, and S. Buus. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PloS one*, 2(8):e796, jan 2007. 22, 24

[91] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics (Oxford, England)*, 20(9):1388–97, jun 2004. 24

[92] K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *Journal of immunology (Baltimore, Md. : 1950)*, 152(1):163–75, jan 1994. 22

[93] J. L. Perez-Gracia, S. Labiano, M. E. Rodriguez-Ruiz, M. F. Sanmamed, and I. Melero. Orchestrating immune check-point blockade for cancer immunotherapy in combinations. *Current Opinion in Immunology*, 27:89–97, apr 2014. 97

[94] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, dec 1999. 3, 16, 17, 72, 83, 88, 100

[95] D. L. Porter, W.-t. Hwang, N. V. Frey, S. F. Lacey, P. a. Shaw, A. W. Loren, A. Bagg, K. T. Marcucci, A. Shen, V. Gonzalez, D. Ambrose, S. a. Grupp, A. Chew, Z. Zheng, M. C. Milone, B. L. Levine, J. J. Melenhorst, and C. H. June. Chimeric antigen receptor T cells persist and induce sustained remissions in relapsed refractory chronic lymphocytic leukemia. *Science translational medicine*, 7(303):303ra139, sep 2015. 4, 86

[96] R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, 0:{ISBN} 3–900051–07–0, 2016. 89

[97] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–9, nov 1999. 2, 22, 47, 101

[98] P. a. Reche, J.-P. Glutting, and E. L. Reinherz. Prediction of MHC class I binding peptides using profile motifs. *Human immunology*, 63(9):701–9, sep 2002. 22

[99] M. Reck, D. Rodríguez-Abreu, A. G. Robinson, R. Hui, T. Csöszi, A. Fülöp, M. Gottfried, N. Peled, A. Tafreshi, S. Cuffe, M. O'Brien, S. Rao, K. Hotta, M. A. Leiby, G. M. Lubiniecki, Y. Shentu, R. Rangwala, and J. R. Brahmer. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *New England Journal of Medicine*, 375(19):1823–1833, nov 2016. 1, 99

[100] J. L. Riley. Combination Checkpoint Blockade - Taking Melanoma Immunotherapy to the Next Level. *New England Journal of Medicine*, 369(2):187–189, jul 2013. 4, 85

[101] M. Ritgen, S. Stilgenbauer, N. von Neuhoff, A. Humpe, M. Brüggemann, C. Pott, T. Raff, A. Kröber, D. Bunjes, R. Schlenk, N. Schmitz, H. Döhner, M. Kneba, and P. Dreger. Graft-versus-leukemia activity may overcome therapeutic resistance of chronic lymphocytic leukemia with unmutated immunoglobulin variable heavy-chain gene status: implications of minimal residual disease measurement with quantitative PCR. *Blood*, 104(8):2600–2, oct 2004. 4, 86

[102] N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, and T. A. Chan. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (New York, N.Y.)*, 348(6230):124–8, apr 2015. 4, 85, 97, 101

[103] P. F. Robbins, Y.-C. Lu, M. El-Gamil, Y. F. Li, C. Gross, J. Gartner, J. C. Lin, J. K. Teer, P. Cliften, E. Tycksen, Y. Samuels, and S. A. Rosenberg. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nature medicine*, 19(6):747–52, jun 2013. 97

[104] J. Robinson, J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, and S. G. E. Marsh. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, 43, 2015. 13

[105] J. Robinson, J. a. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, and S. G. E. Marsh. The IPD and IMGT/HLA database: allele variant databases. *Nucleic acids research*, 43(Database issue):D423–31, jan 2015. 22

[106] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods*, 13(9):741–8, aug 2016. 69

[107] K. S. Roush and C. D. Hillyer. Donor lymphocyte infusion therapy. *Transfusion medicine reviews*, 16(2):161–76, apr 2002. 4, 86

[108] N. H. Russell, J. L. Byrne, R. D. Faulkner, M. Gilyead, E. P. Das-Gupta, and A. P. Haynes. Donor lymphocyte infusions can result in sustained remissions in patients with residual or relapsed lymphoid malignancy following allogeneic haemopoietic stem cell transplantation. *Bone Marrow Transplantation*, 36(5):437–441, sep 2005. 4, 86

[109] D. Samson. Diagnosis and management of multiple myeloma, dec 2001. 11

[110] R. Schipper, J. D'Amaro, J. Bakker, J. Bakker, J. van Rood, and M. Oudshoorn. HLA gene and haplotype frequencies in bone marrow donors worldwide registries. *Human Immunology*, 52(1):54–71, jan 1997. 90

[111] C. Schmid, M. Labopin, A. Nagler, M. Bornhauser, J. Finke, A. Fassas, L. Volin, G. Gurman, J. Maertens, P. Bordigoni, E. Holler, G. Ehninger, E. Polge, N.-C. Gorin, H.-J. Kolb, and V. Rocha. Donor Lymphocyte Infusion in the Treatment of First Hematological Relapse After Allogeneic Stem-Cell Transplantation in Adults With Acute Myeloid Leukemia: A Retrospective Risk Factors Analysis and Comparison With Other Strategies by the EBMT Acute Leukem. *Journal of Clinical Oncology*, 25(31):4938–4945, nov 2007. 4, 86

[112] S. Seidl, H. Kaufmann, and J. Drach. New insights into the pathophysiology of multiple myeloma. *The Lancet Oncology*, 4(9):557–564, sep 2003. 11

[113] D. SenGupta, P. J. Norris, T. J. Suscovich, M. Hassan-Zahraee, H. F. Moffett, A. Trocha, R. Draenert, P. J. R. Goulder, R. J. Binder, D. L. Levey, B. D. Walker, P. K. Srivastava, and C. Brander. Heat shock protein-mediated cross-presentation of exogenous HIV antigen on HLA class I and class II. *Journal of immunology (Baltimore, Md. : 1950)*, 173(3):1987–93, aug 2004. 3, 84

[114] W. Shao, P. G. Pedrioli, W. Wolski, C. Scurtescu, E. Schmid, J. A. Vizcaíno, M. Courcelles, H. Schuster, D. Kowalewski, F. Marino, C. S. Arlehamn, K. Vaughan, B. Peters, A. Sette, T. H. Ottenhoff, K. E. Meijgaarden, N. Nieuwenhuizen, S. H. Kaufmann, R. Schlapbach, J. C. Castle, A. I. Nesvizhskii, M. Nielsen, E. W. Deutsch, D. S. Campbell, R. L. Moritz, R. A. Zubarev, A. J. Ytterberg, A. W. Purcell, M. Marcilla, A. Paradela, Q. Wang, C. E. Costello, N. Ternette, P. A. vanÂăVeelen, C. A. vanÂăEls, A. J. Heck, G. A. deÂăSouza, L. M. Sollid, A. Admon, S. Stevanovic, H.-G. Rammensee, P. Thibault, C. Perreault, M. Bassani-Sternberg, R. Aebersold, and E. Caron. The SysteMHC Atlas project. *Nucleic Acids Research*, 2017. 1, 2, 31, 35, 51, 102

[115] B. Shraibman, D. M. Kadosh, E. Barnea, and A. Admon. Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. *Molecular & Cellular Proteomics*, 15(9):3058–3070, sep 2016. 97

[116] D. Shteynberg, E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold, and A. I. Nesvizhskii. iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Molecular & Cellular Proteomics*, 10(12):M111.007690, 2011. 32

[117] A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok, and T. A. Chan. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *New England Journal of Medicine*, 371(23):2189–2199, dec 2014. 4, 85, 97

[118] J. S. Stickel, C. Berlin, M. Schemionek, L. Kanz, H. R. Salih, T. H. Brümmendorf, H.-G. Rammensee, S. Stevanovic, and D. J. Kowalewski. HLA Ligandome Analysis of Chronic Myeloid Leukemia (CML), Revealed Novel Tumor Associated Antigens for Peptide Based Immunotherapy. *Blood, ASH Abstract*, 2013. 4

[119] J. S. Stickel, A. O. Weinzierl, N. Hillen, O. Drews, M. M. Schuler, J. Hennenlotter, D. Wernet, C. A. Müller, A. Stenzl, H.-G. Rammensee, and S. Stevanović. HLA ligand profiles of primary renal cell carcinoma maintained in metastases. *Cancer Immunology, Immunotherapy*, 58(9):1407–1417, sep 2009. 86, 97

[120] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, aug 2003. 20

[121] E. Stronen, M. Toebes, S. Kelderman, M. M. van Buuren, W. Yang, N. van Rooij, M. Donia, M.-L. Boschen, F. Lund-Johansen, J. Olweus, and T. N. Schumacher. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science*, 352(6291):1337–1341, jun 2016. 85

[122] A. Szolek, B. Schubert, C. Mohr, M. Sturm, M. Feldhahn, and O. Kohlbacher. OptiType: Precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, 2014. 1

[123] M. The, M. J. MacCoss, W. S. Noble, and L. Käll. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of The American Society for Mass Spectrometry*, 27(11):1719–1727, nov 2016. 20, 21, 69

[124] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, jan 2017. 55, 101

[125] M. C. F. Thomsen and M. Nielsen. 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 Research*, 40(Web Server issue):W281–W287, jul 2012. 47, 49

[126] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten. Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419, jan 2015. 2, 31, 52, 53, 83

[127] E. M. Van Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. Foppen, S. M. Goldinger, J. Utikal, J. C. Hassel, B. Weide, K. C. Kaehler, C. Loquai, P. Mohr, R. Gutzmer, R. Dummer, S. Gabriel, C. J. Wu, D. Schadendorf, and L. A. Garraway. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, 350(6257):207–211, 2015. 32

[128] N. van Rooij, M. M. van Buuren, D. Philips, A. Velds, M. Toebes, B. Heemskerk, L. J. van Dijk, S. Behjati, H. Hilkmann, D. el Atmioui, M. Nieuwland, M. R. Stratton, R. M. Kerkhoven, C. Kesmir, J. B. Haanen, P. Kvistborg, and T. N. Schumacher. Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. *Journal of Clinical Oncology*, 31(32):e439–e442, nov 2013. 97

[129] C. M. Verfaillie. Biology of chronic myelogenous leukemia. *Hematology/oncology clinics of North America*, 12(1):1–29, feb 1998. 11

[130] R. Vita, J. a. Overton, J. a. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. The immune epitope database (IEDB) 3.0. *Nucleic acids research*, 43(Database issue):D405–12, jan 2015. 1, 2

[131] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, mar 2013. 86, 97

[132] P. Voigt and D. Reinberg. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia The Cancer Genome Atlas Research Network. *The New England journal of medicine*, 368(22):2059–74, may 2013. 10

[133] A. Walker and R. Johnson. Commercialization of cellular immunotherapies for cancer. *Biochemical Society Transactions*, 44(2):329–332, apr 2016. 86, 97

[134] S. Walter, T. Weinschenk, A. Stenzl, R. Zdrojowy, A. Pluzanska, C. Szczylik, M. Staehler, W. Brugger, P.-Y. Dietrich, R. Mendrzyk, N. Hilf, O. Schoor, J. Fritsche, A. Mahr, D. Maurer, V. Vass, C. Trautwein, P. Lewandrowski, C. Flohr, H. Pohla, J. J. Stanczak, V. Bronte, S. Mandruzzato, T. Biedermann, G. Pawelec, E. Derhovanessian, H. Yamagishi, T. Miki, F. Hongo, N. Takaha, K. Hirakawa, H. Tanaka, S. Stevanovic, J. Frisch, A. Mayer-Mokler, A. Kirner, H.-G. Rammensee, C. Reinhardt, and H. Singh-Jasuja. Multipeptide immune response to cancer vaccine IMA901 after single-dose cyclophosphamide associates with longer patient survival. *Nature Medicine*, 18(8):1254–1261, jul 2012. 97

[135] S. Walz, J. S. Stickel, D. J. Kowalewski, H. Schuster, K. Weisel, L. Backert, S. Kahn, A. Nelde, T. Stroh, M. Handel, O. Kohlbacher, L. Kanz, H. R. Salih, H.-G. Rammensee, and S. Stevanović.

The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for T-cell-based immunotherapy. *Blood*, 126(10):1203–13, sep 2015. 1, 3, 4, 53, 86, 97, 99

[136] M. Walzer, L. E. Pernas, S. Nasso, W. Bittremieux, S. Nahnsen, P. Kelchtermans, P. Pichler, H. W. P. van den Toorn, A. Staes, J. Vandenbussche, M. Mazanek, T. Taus, R. A. Scheltema, C. D. Kelstrup, L. Gatto, B. van Breukelen, S. Aiche, D. Valkenborg, K. Laukens, K. S. Lilley, J. V. Olsen, A. J. R. Heck, K. Mechtler, R. Aebersold, K. Gevaert, J. A. Vizcaíno, H. Hermjakob, O. Kohlbacher, and L. Martens. qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments. *Molecular & Cellular Proteomics*, 13(8):1905–1913, 2014. 60

[137] J. Wan, W. Liu, Q. Xu, Y. Ren, D. R. Flower, and T. Li. SVRMHC prediction server for MHC-binding peptides. *BMC bioinformatics*, 7(1):463, oct 2006. 22

[138] X. Wang and I. Rivière. Manufacture of tumor- and virus-specific T lymphocytes for adoptive cell therapies. *Cancer Gene Therapy*, 22(2):85–94, mar 2015. 86, 97

[139] G. R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables. gplots: Various R Programming Tools for Plotting Data. *R package version 2.17.0.*, page 2015, 2015. 89

[140] G. R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables. *gplots: Various R Programming Tools for Plotting Data*, 2016. 56

[141] P. L. Weiden, N. Flournoy, E. D. Thomas, R. Prentice, A. Fefer, C. D. Buckner, and R. Storb. Antileukemic Effect of Graft-versus-Host Disease in Human Recipients of Allogeneic-Marrow Grafts. *New England Journal of Medicine*, 300(19):1068–1073, may 1979. 4, 86

[142] A. O. Weinzierl, C. Lemmel, O. Schoor, M. Muller, T. Kruger, D. Wernet, J. Hennenlotter, A. Stenzl, K. Klingel, H.-G. Rammensee, and S. Stevanovic. Distorted Relation between mRNA Copy Number and Corresponding Major Histocompatibility Complex Ligand Density on the Cell Surface. *Molecular & Cellular Proteomics*, 6(1):102–113, oct 2006. 98, 101

[143] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, may 2014. 2, 31, 35, 52, 53, 80, 83, 100

[144] L. Wilkinson. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics*, 67(2):678–679, 2011. 71, 72

[145] Y. Zhao, E. Moon, C. Carpenito, C. M. Paulos, X. Liu, A. L. Brennan, A. Chew, R. G. Carroll, J. Scholler, B. L. Levine, S. M. Albelda, and C. H. June. Multiple injections of electroporated autologous T cells expressing a chimeric antigen receptor mediate regression of human disseminated tumor. *Cancer research*, 70(22):9053–61, nov 2010. 86

Appendix A

Abbreviations

A	
ADCC	*antibody-dependent cell-mediated cytotoxicity*
AML	*Acute myeloid leukemia*
ANSI	*American National Standards Institute*
AUC	*Area under the curve*
B	
BCR	*B cell receptor*
C	
CML	*Chronic myeloid leukemia*
CLL	*Chronic lymphocytic leukemia*
CSS	*Cascading Style Sheets*
CTL	*Cytoxic killer T cells*
D	
DDA	*Data Dependent Acquisition*
DIA	*Data Independent Acquisition*
E	
ER	*Endoplasmic reticulum*
F	
FDR	*False-discovery rate*
FP	*False positive*
FPR	*False positive rate*

G	
GSEA	*Gene set enrichment analysis*

H	
HM	*Hematological Malignancies*
HLA	*Human leukocyte antigen*
HTML	*HyperText Markup Language*

I	
IMGT	*the international ImMunoGeneTics information system*
ISO	*International Organization for Standardization*

J	
JS	*JavaScript*
JSON	*JavaScript Object Notation*

K	
KLD	*Kullback-Leibler distance*

L	
LC-MS/MS	*Liquid chromatography coupled tandem mass spectrometry*

M	
ML	*Machine learning*
MS	*Mass spectrometry*
MOWSE	*Molecular weight search*
MM	*Multiple myeloma*

N	
NGS	*Next Generation Sequencing*

P	
PAMPs	*Pathogen-associated molecular patterns*
PRIDE	*PRoteomics IDEntifications database*
PSM	*Peptide Spectrum Match*
PSSM	*Position specific scoring matrix*

Q	
QC	*quality control*

R

ROC	*Receiver operating characteristic*

S

SQL	*Structured Query Language*

T

TAA	*Tumor associated antigen*
TAP	*Transporter associated with antigen processing*
TCR	*T-cell receptor*
TP	*True positive*
TRP	*True positive rate*
TSA	*Tumor specific antigen*

V

Y

Appendix B

Publications

Accepted manuscripts

2018

*Neidert MC, Kowalewski DJ, Silginer M, Kapolou K, **Backert L**, Freudenmann LK, Peper JK, Marcu A, Wang SS, Walz JS, Wolpert F, Rammensee HG, 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. 2018 Jun;135(6):923-938

*Löffler MW, Kowalewski DJ, **Backert L**, Bernhardt J, Adam P, Schuster H, Dengler F, Backes D, Kopp HG, Beckert S, Wagner S, Königsrainer I, Kohlbacher O, Kanz L, Königsrainer A, Rammensee HG, Stevanović S, Haen SP.* Mapping the HLA ligandome of Colorectal Cancer Reveals an Imprint of Malignant Cell Transformation. Cancer Res. 2018 May 22. pii: canres.1745.2017.

*Nelde A, Kowalewski DJ, **Backert L**, Schuster H, Werner JO, Klein R, Kohlbacher O, Kanz L, Salih HR, Rammensee HG, Stevanović S, Walz JS.* 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. 2018 Feb 14;7(4):e1316438.

*Walz JS, Kowalewski DJ, **Backert L**, Nelde A, Kohlbacher O, Weide B, Kanz L, Salih HR, Rammensee HG, Stevanović S.* Favorable immune signature in CLL patients, defined by antigen-specific T-cell responses, might prevent second skin cancers. Leuk Lymphoma. 2018 Jan 3:1-10

2017

*Fuchs S, Mehlan H, Bernhardt J, Hennig A, Michalik S, Surmann K, Pané-Farré J, Giese A, Weiss S, **Backert L**, Herbig A, Nieselt K, Hecker M, Völker U, Mäder U.* AureoWiki - The repository of the Staphylococcus aureus research and annotation community. Int J Med Microbiol. 2017 Nov 24. pii: S1438-4221(17)30462-9.

*Schuster H, Peper JK, Bösmüller HC, Röhle K, **Backert L**, Bilich T, Ney B, Löffler MW, Kowalewski DJ, Trautwein N, Rabsteyn A, Engler T, Braun S, Haen SP, Walz JS, Schmid-Horch B, Brucker SY, Wallwiener D, Kohlbacher O, Fend F, Rammensee HG, Stevanović S, Staebler A, Wagner P.* The immunopeptidomic landscape of ovarian carcinomas. Proc Natl Acad Sci U S A. 2017 Nov 14;114(46):E9942-E9951

*Di Marco M, Schuster H, **Backert L**, Ghosh M, Rammensee HG, Stevanović* Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. J Immunol. 2017 Oct 15;199(8):2639-2651.

*Alberer M, Gnad-Vogt U, Hong HS, Mehr KT, **Backert L**, Finak G, Gottardo R, Bica MA, Garofano A, Koch SD, Fotin-Mleczek M, Hoerr I, Clemens R, von Sonnenburg F.* Safety and immunogenicity of a mRNA rabies vaccine in healthy adults: an open label, non-randomised, prospective, first-in-human phase I clinical trial. Lancet. 2017 Sep 23;390(10101):1511-1520.

*Haen SP, Groh C, Schumm M, **Backert L**, Löffler MW, Federmann B, Faul C, Dörfel D, Vogel W, Handgretinger R, Kanz L, Beth WA.* Haploidentical hematopoietic cell transplantation using in vitro T cell depleted grafts as salvage therapy in patients with disease relapse after prior allogeneic transplantation. Ann Hematol 2017 96:817.

Backert L*, Kowalewski DJ*, Walz S, Schuster H, Berlin C, Neidert MC, Schemionek M, Brümmendorf TH, Vucinic V, Niederwieser D, Kanz L, Salih HR, Kohlbacher O, Weisel K, Rammensee HG, Stevanović S, Walz JS.* A meta-analysis of HLA peptidome composition in different hematological entities: Entity-specific dividing lines and "pan-leukemia" antigens. Oncotarget. 2017 Jul 4;8(27):43915-43924.

2016

Hong HS, Koch SD, Scheel B, Gnad-Vogt U, Schröder A, Kallen KJ, Wiegand V, ***Backert L****, Kohlbacher O, Hoerr I, Fotin-Mleczek M, Billingsley JM.* Distinct transcriptional changes in non-small cell lung cancer patients associated with multi-antigenic RNActive©CV9201 immunotherapy. with aromatic C-termini. Oncoimmunology. 2016 Nov 18;5(12):e1249560

Barth SM, Schreitmüller CM, Proehl F, Oehl K, Lumpp LM, Kowalewski DJ, Di Marco M, Sturm T, ***Backert L****, Schuster H, Stevanović S, Rammensee HG, Planz O.* Characterization of the Canine MHC Class I DLA-88*50101 Peptide Binding Motif as a Prerequisite for Canine T Cell Immunotherapy. PLoS One. 2016 Nov 28;11(11):e0167017

Kowalewski DJ, Walz S, ***Backert L****, Schuster H, Kohlbacher O, Weisel K, Rittig SM, Kanz L, Salih HR, Rammensee HG, Stevanović S, Stickel JS.* Carfilzomib alters the HLA-presented peptidome of myeloma cells and impairs presentation of peptides with aromatic C-termini. Blood Cancer J. 2016 Apr 8;6:e411

2015

Backert L*, Kohlbacher O.* Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med. 2015 Nov 20;7:119

Walz S, Stickel JS, Kowalewski DJ, Schuster H, Weisel K, ***Backert L****, Kahn S, Nelde A, Stroh T, Handel M, Kohlbacher O, Kanz L, Salih HR, Rammensee HG, Stevanović S.* The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for T-cell-based immunotherapy. Blood. 2015 Sep 3;126(10):1203-13

Kowalewski DJ, Schuster H, ***Backert L****, Berlin C, Kahn S, Kanz L, Salih HR, Rammensee HG, Stevanović S, Stickel JS.* HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). Proc Natl Acad Sci U S A. 2015 Jan 13;112(2):E166-75

2014

Römer M, ***Backert L****, Eichner J, Zell A.* ToxDBScan: Large-scale similarity screening of toxicological databases for drug candidates. Int J Mol Sci. 2014 Oct 21;15(10):19037-55

Appendix C

Supporting Figures

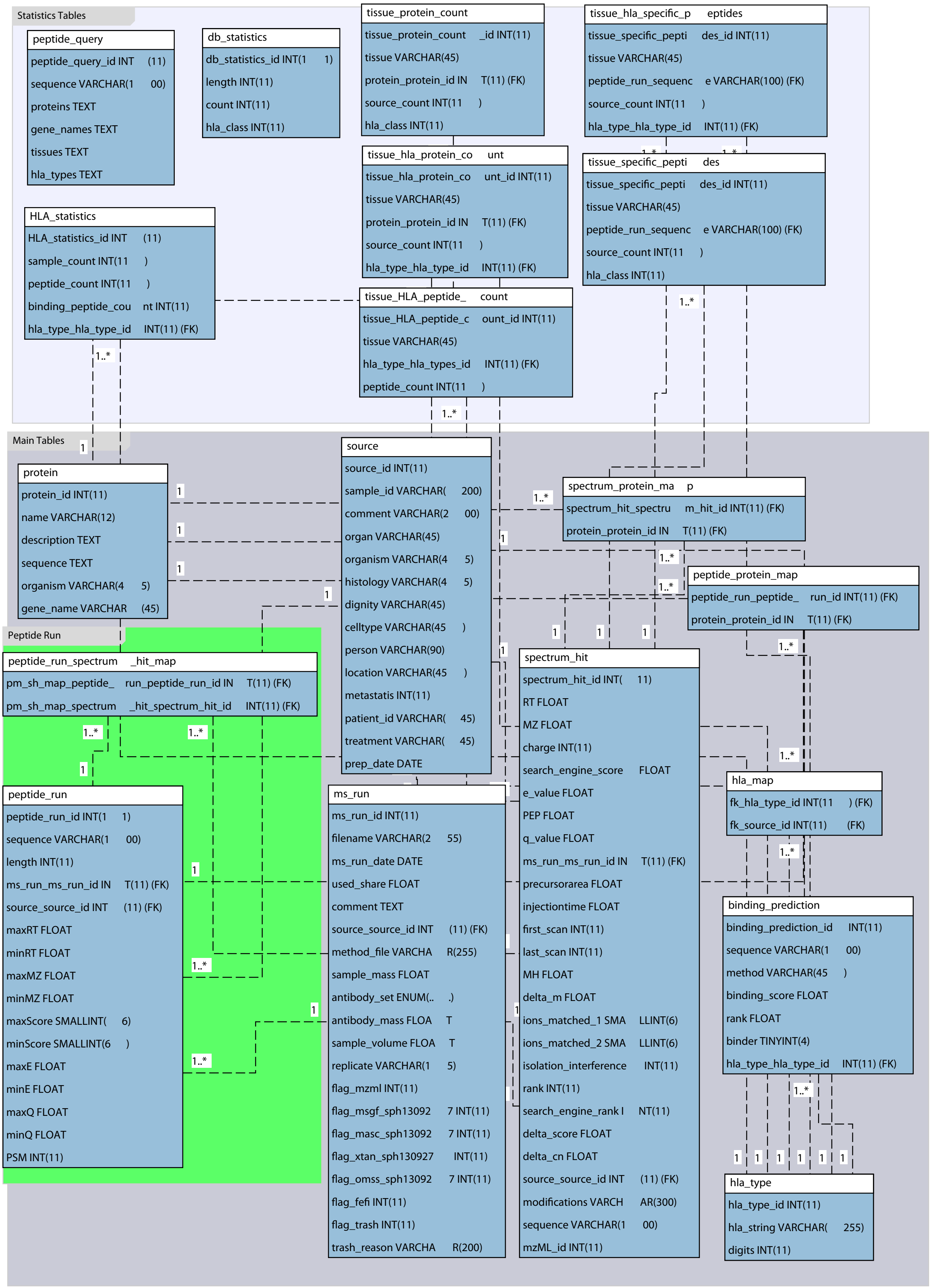
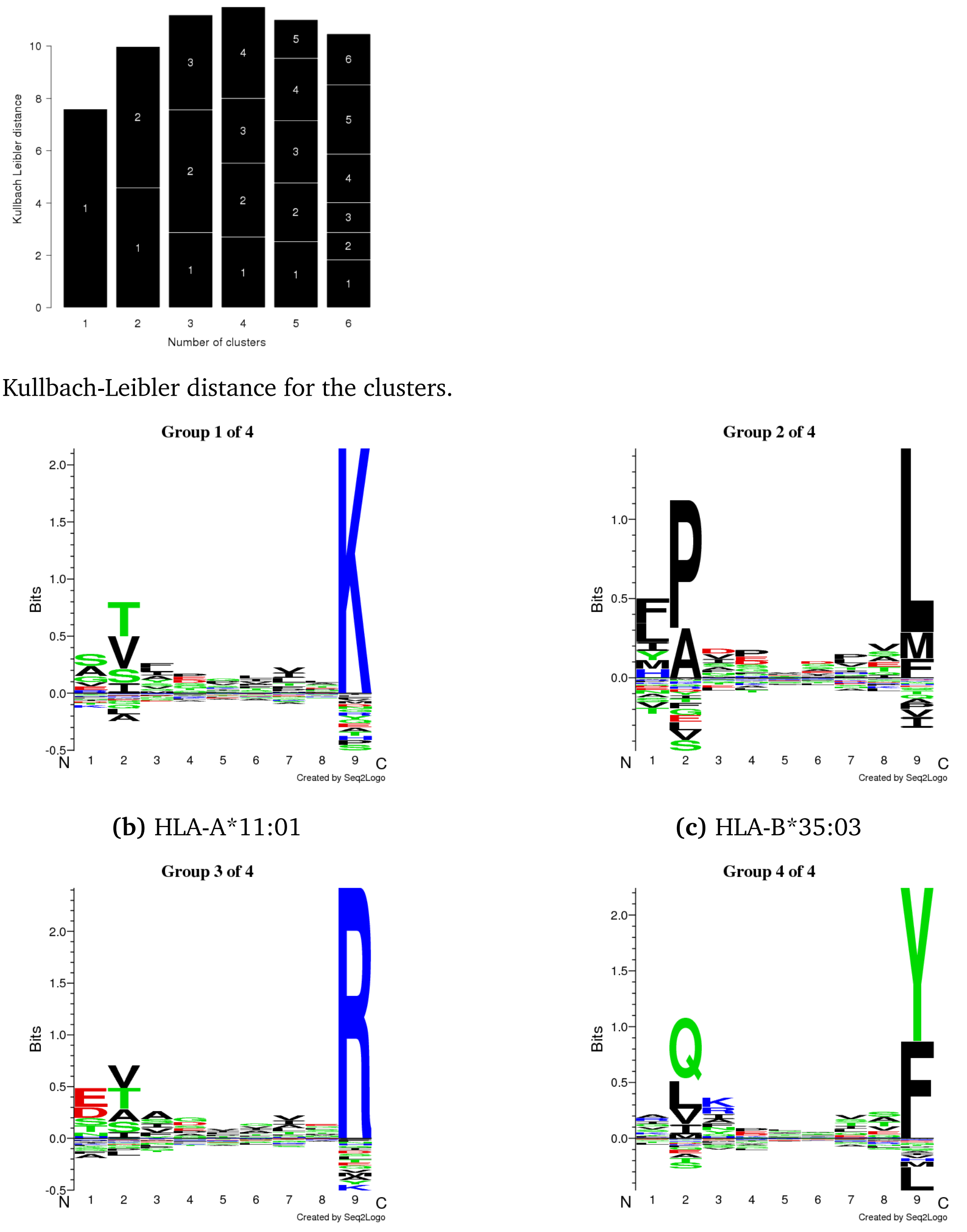


Figure C.1: ERM diagram of the whole MySQL database, including the main tables (gray), the peptide_run tables (green), and multiple precomputed statistic tables to allow a faster access to common queries (blue).



(a) Kullbach-Leibler distance for the clusters.

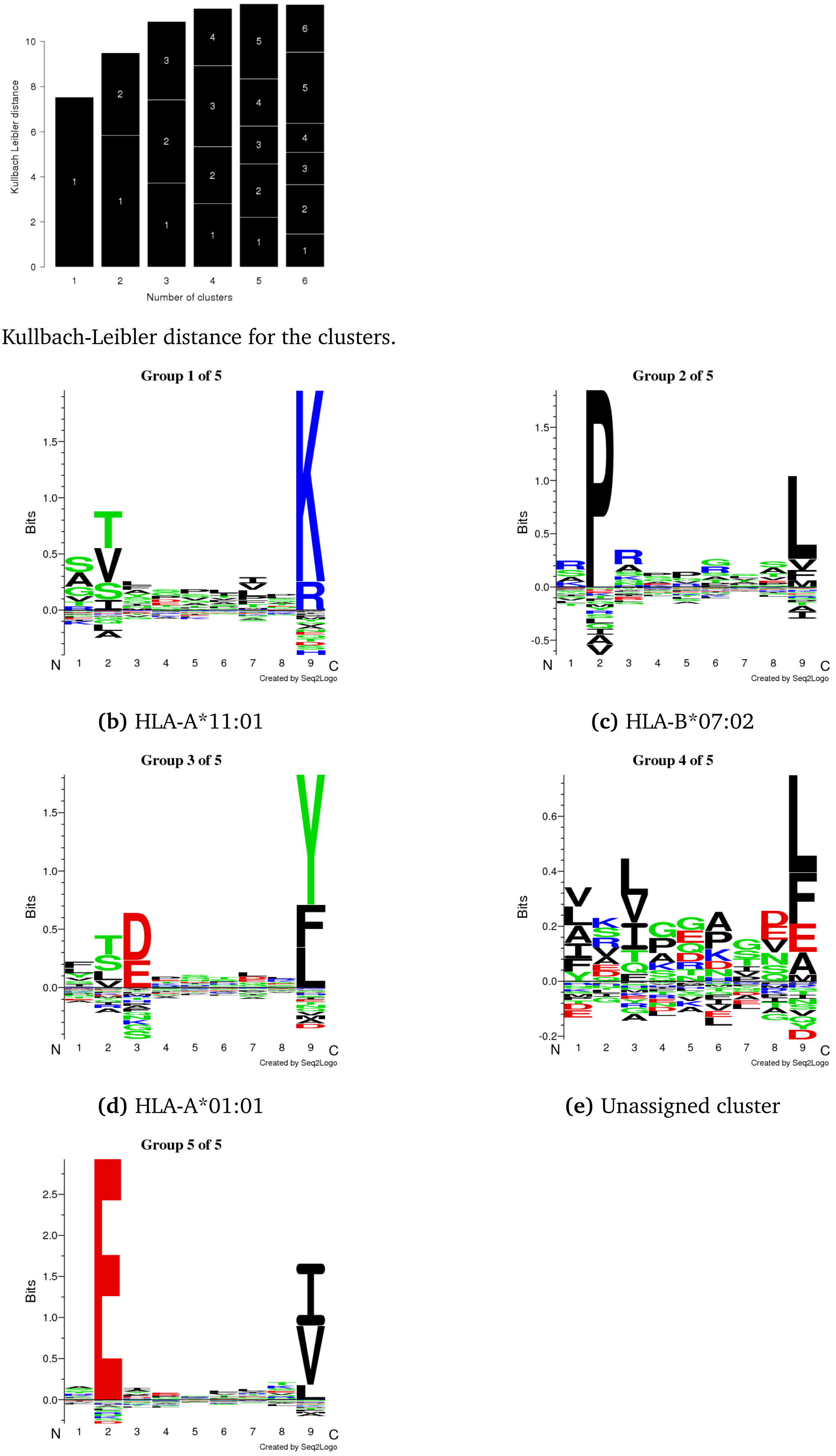
(b) HLA-A*11:01

(c) HLA-B*35:03

(d) HLA-A*68:01

(e) HLA-B*15:01

Figure C.2: Gibbs clustering of all HLA class I peptides found for ZH02



(a) Kullbach-Leibler distance for the clusters.

(b) HLA-A*11:01

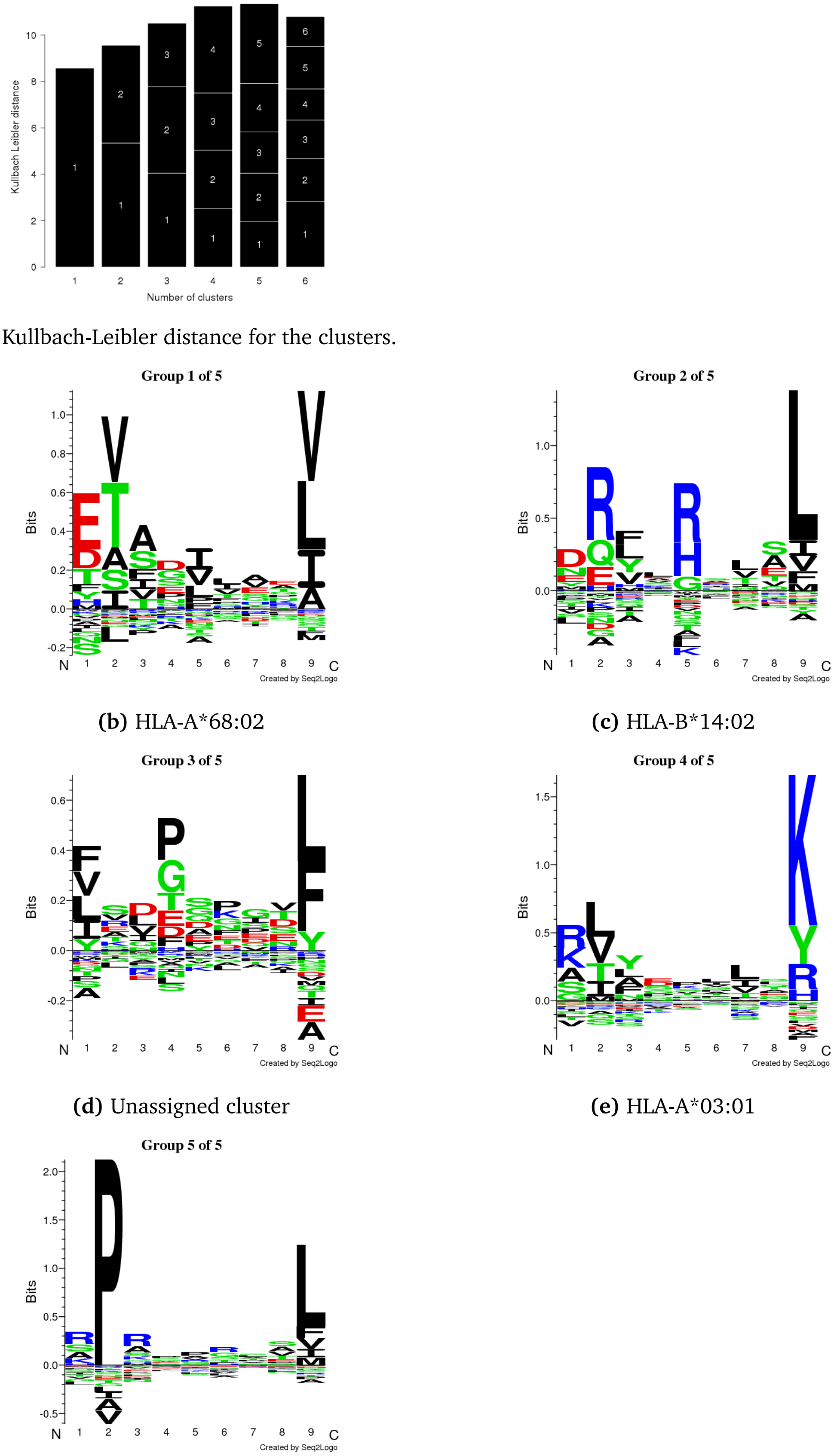
(c) HLA-B*07:02

(d) HLA-A*01:01

(e) Unassigned cluster

(f) HLA-B*49:01

Figure C.3: Gibbs clustering of all HLA class I peptides found for ZH05



(a) Kullbach-Leibler distance for the clusters.

(b) HLA-A*68:02

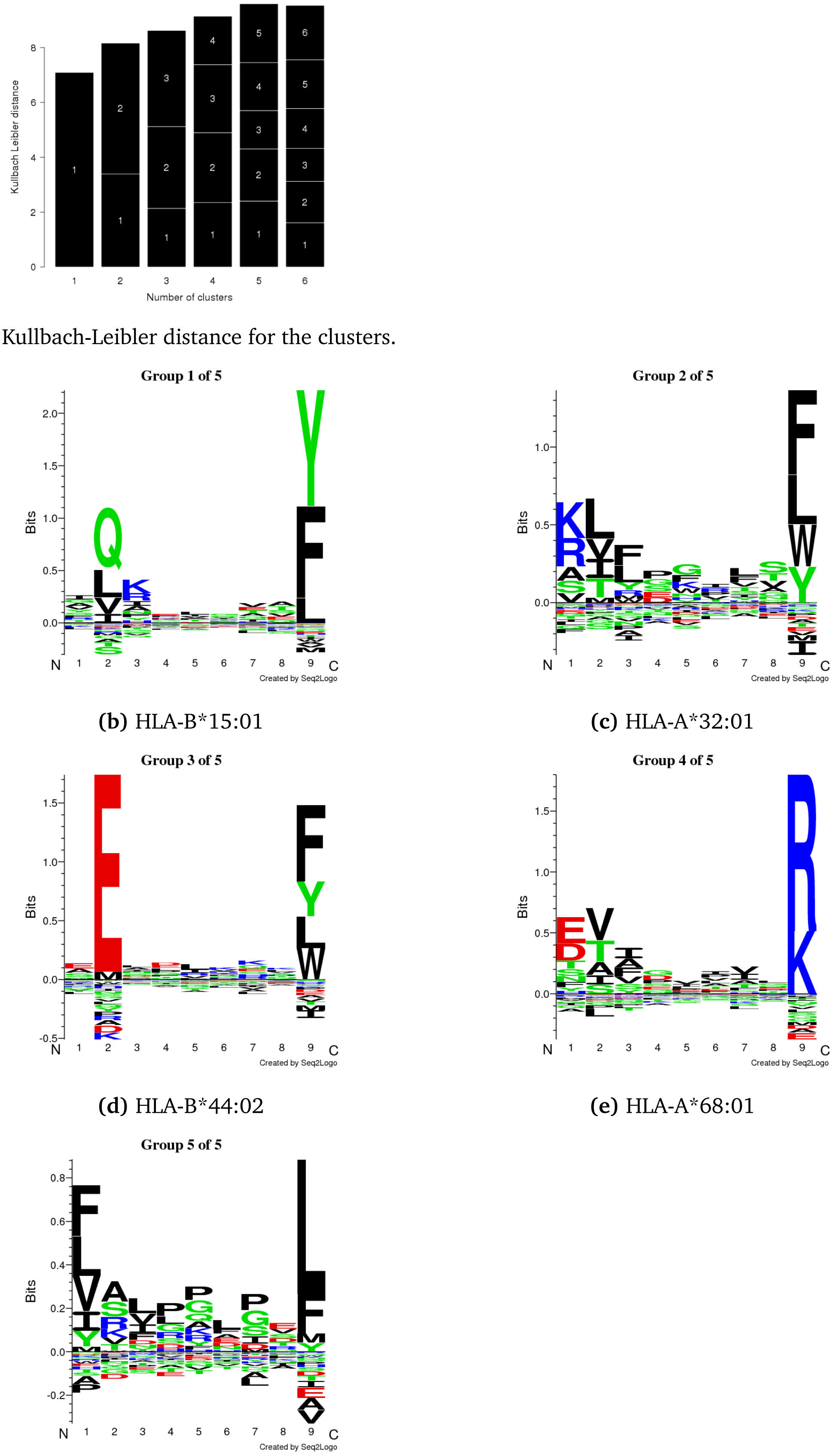
(c) HLA-B*14:02

(d) Unassigned cluster

(e) HLA-A*03:01

(f) HLA-B*07:02

Figure C.4: Gibbs clustering of all HLA class I peptides found for ZH06



(a) Kullbach-Leibler distance for the clusters.

(b) HLA-B*15:01

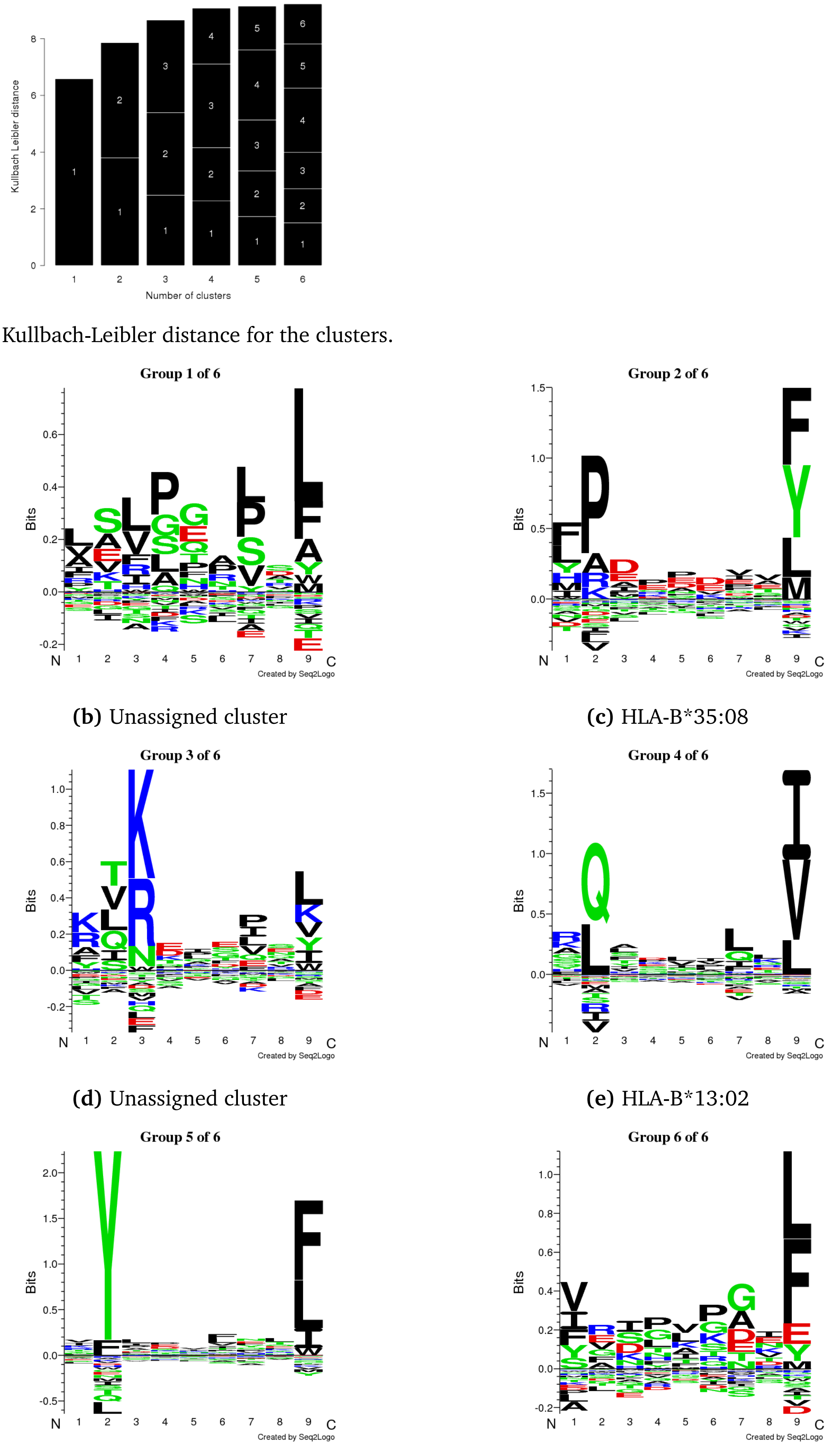
(c) HLA-A*32:01

(d) HLA-B*44:02

(e) HLA-A*68:01

(f) Unassigned cluster

Figure C.5: Gibbs clustering of all HLA class I peptides found for ZH08



(a) Kullbach-Leibler distance for the clusters.

(b) Unassigned cluster

(c) HLA-B*35:08

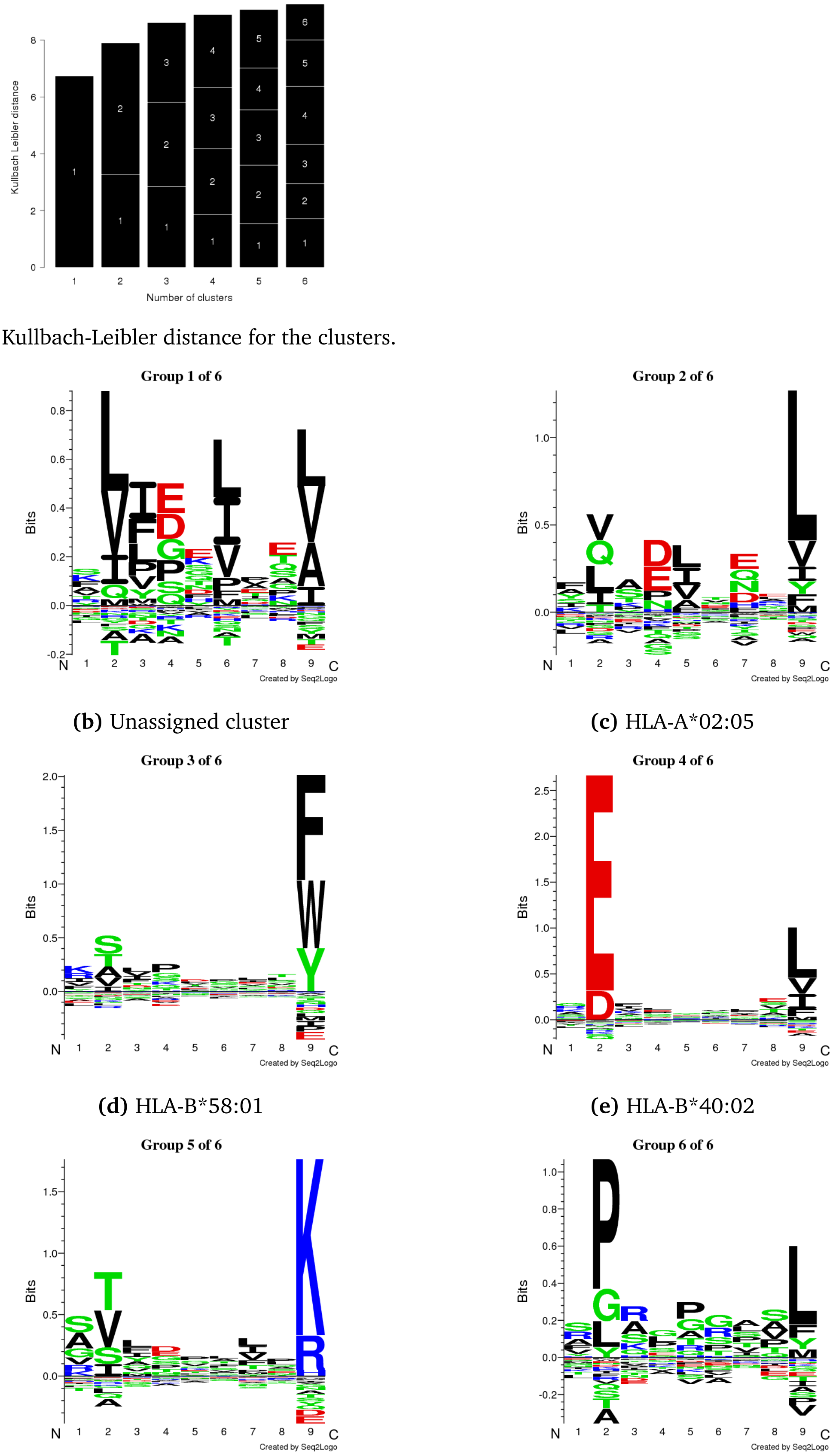
(d) Unassigned cluster

(e) HLA-B*13:02

(f) HLA-A*24:02

(g) Unassigned cluster

Figure C.6: Gibbs clustering of all HLA class I peptides found for ZH09. No distinct cluster was found for HLA-A*30:01.



(a) Kullbach-Leibler distance for the clusters.

(b) Unassigned cluster

(c) HLA-A*02:05

(d) HLA-B*58:01

(e) HLA-B*40:02

(f) HLA-A*11:01

(g) Unassigned cluster

Figure C.7: Gibbs clustering of all HLA class I peptides found for ZH13.

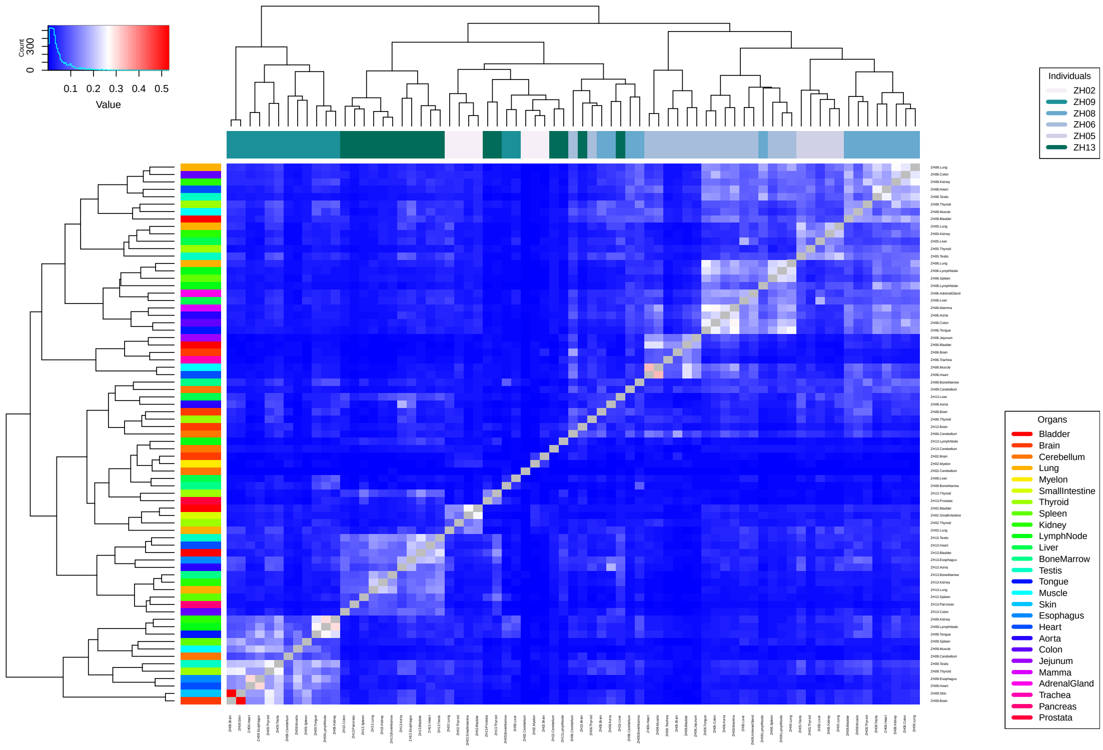


Figure C.8: Identification-based distance matrix heat map for HLA class II peptides. Pairwise distance was calculated using Jaccard-similarity score on presence and absence of identifications. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

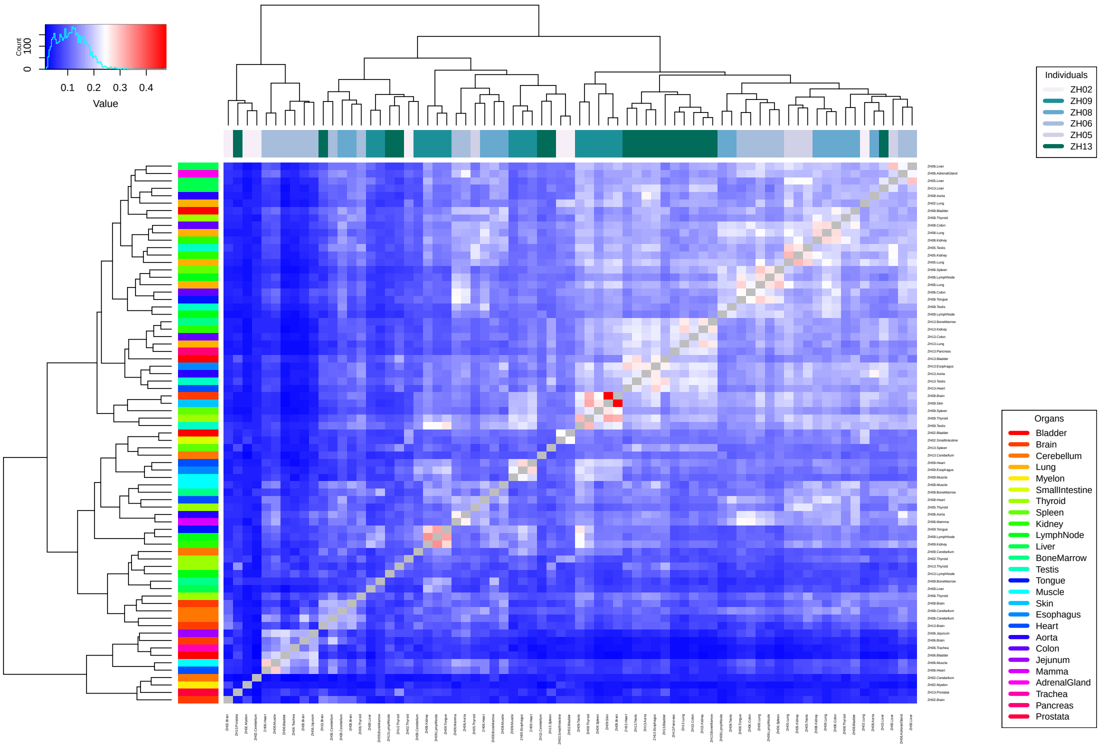


Figure C.9: Protein-based distance matrix heat map for HLA class II. Pairwise distance was calculated using Jaccard-similarity score on presence and absence of proteins. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

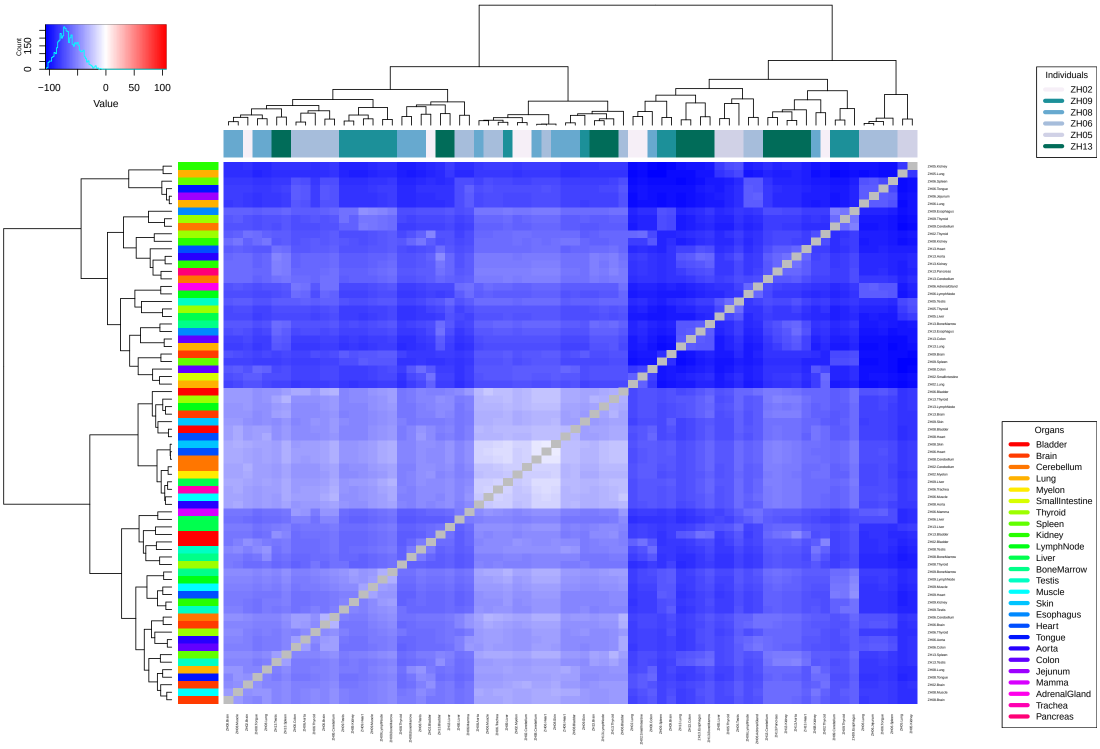


Figure C.10: Identification-based distance matrix heat map for HLA class I peptides with subtraction of the individual **mean** peptide occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

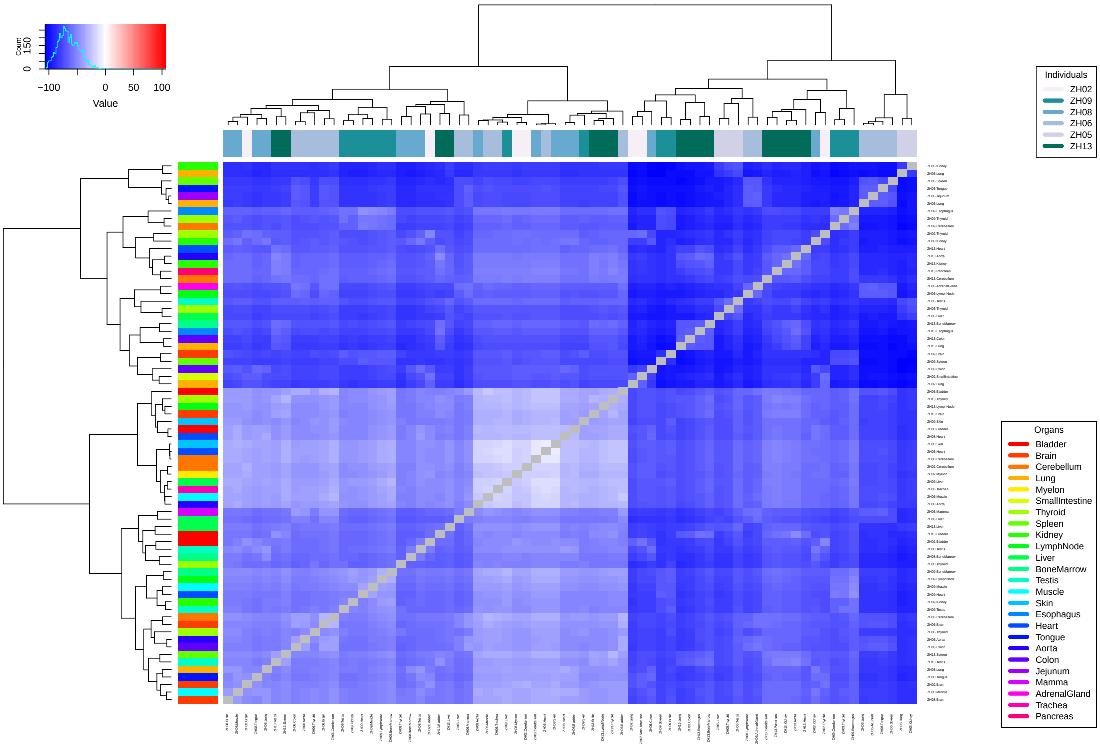


Figure C.11: Identification-based distance matrix heat map for HLA class I peptides with subtraction of the individual **median** peptide occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

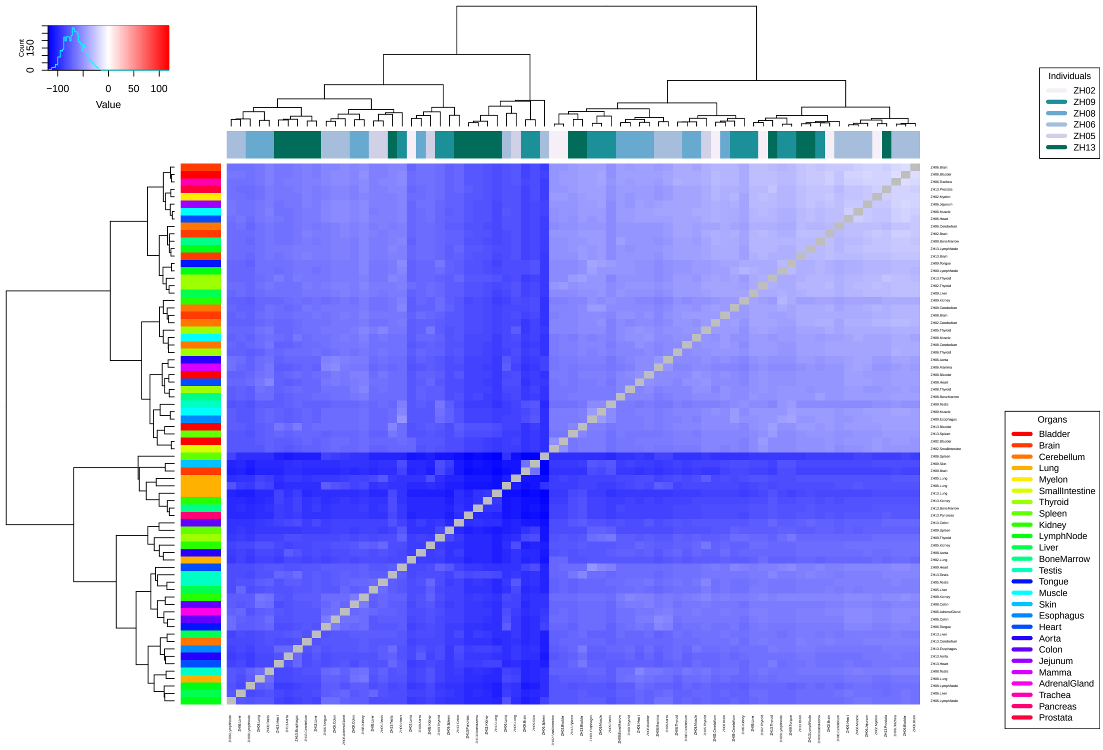


Figure C.12: Identification-based distance matrix heat map for HLA class II peptides with subtraction of the individual **mean** peptide occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

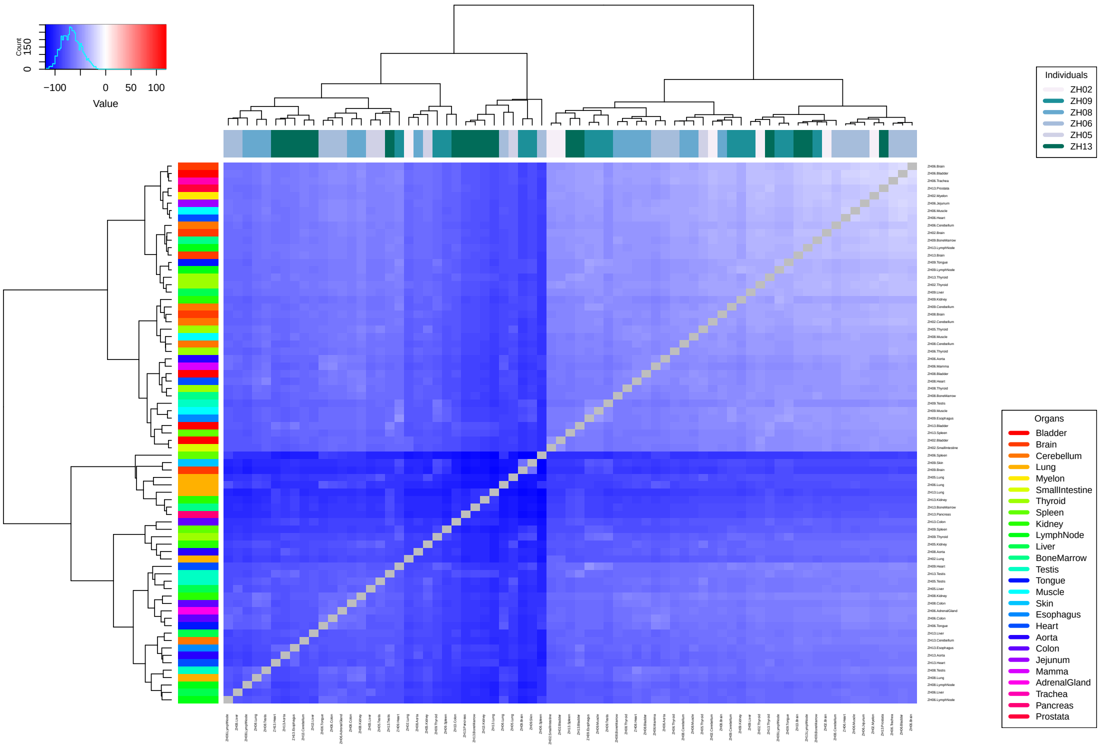


Figure C.13: Identification-based distance matrix heat map for HLA class II peptides with subtraction of the individual **median** peptide occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

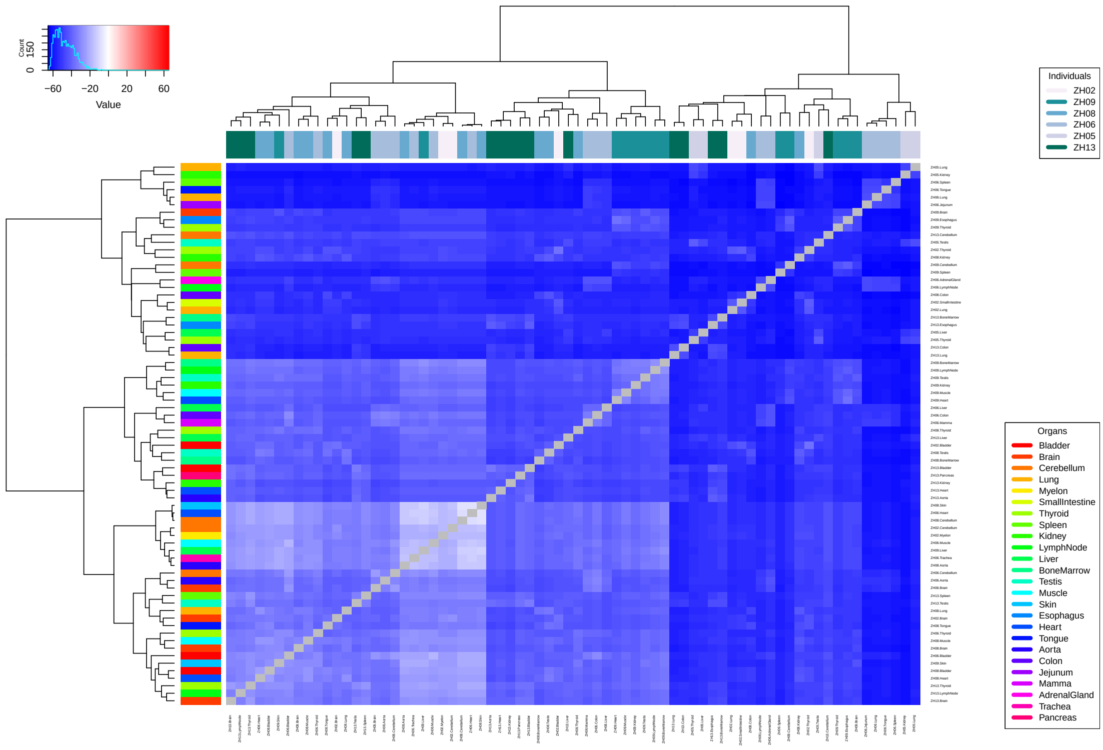


Figure C.14: Protein-based distance matrix heat map for HLA class I proteins with subtraction of the individual **mean** protein occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

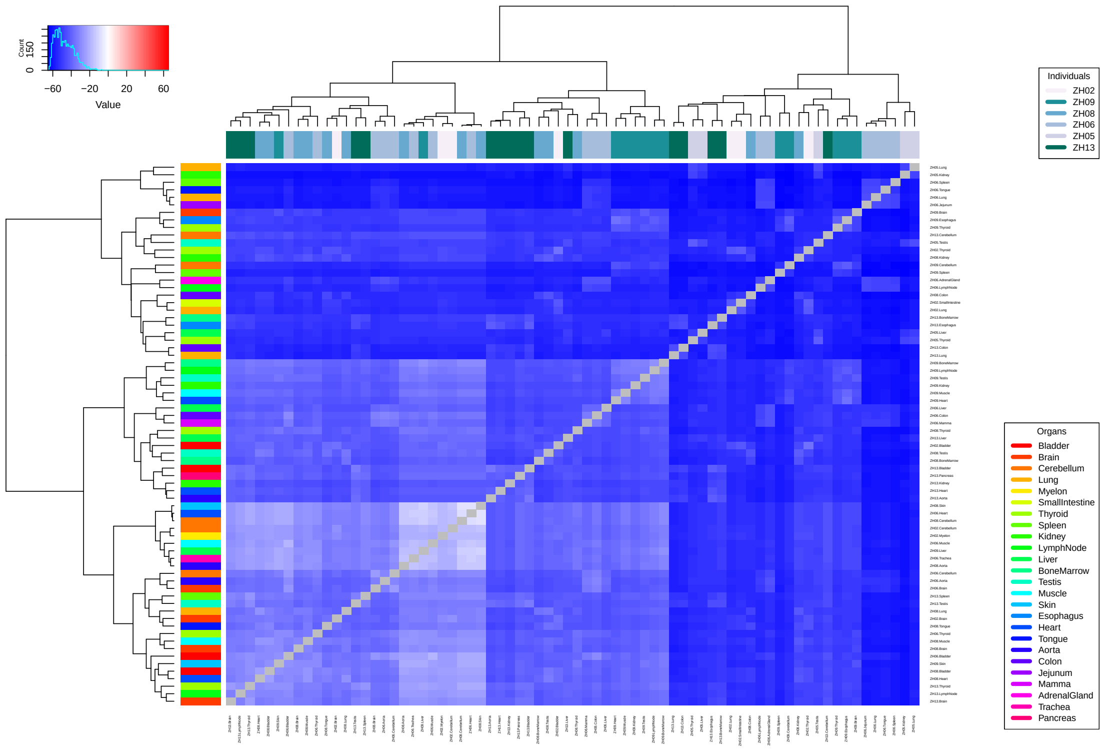


Figure C.15: Protein-based distance matrix heat map for HLA class I proteins with subtraction of the individual **median** protein occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

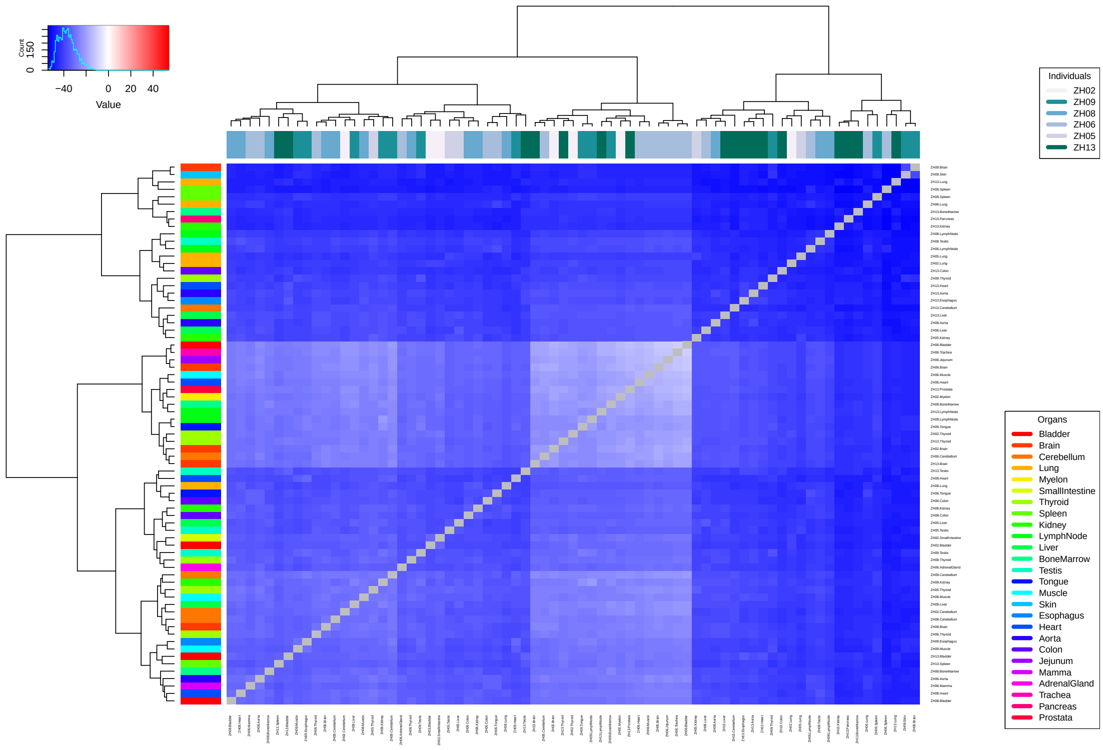


Figure C.16: Protein-based distance matrix heat map for HLA class II proteins with subtraction of the individual **mean** protein occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

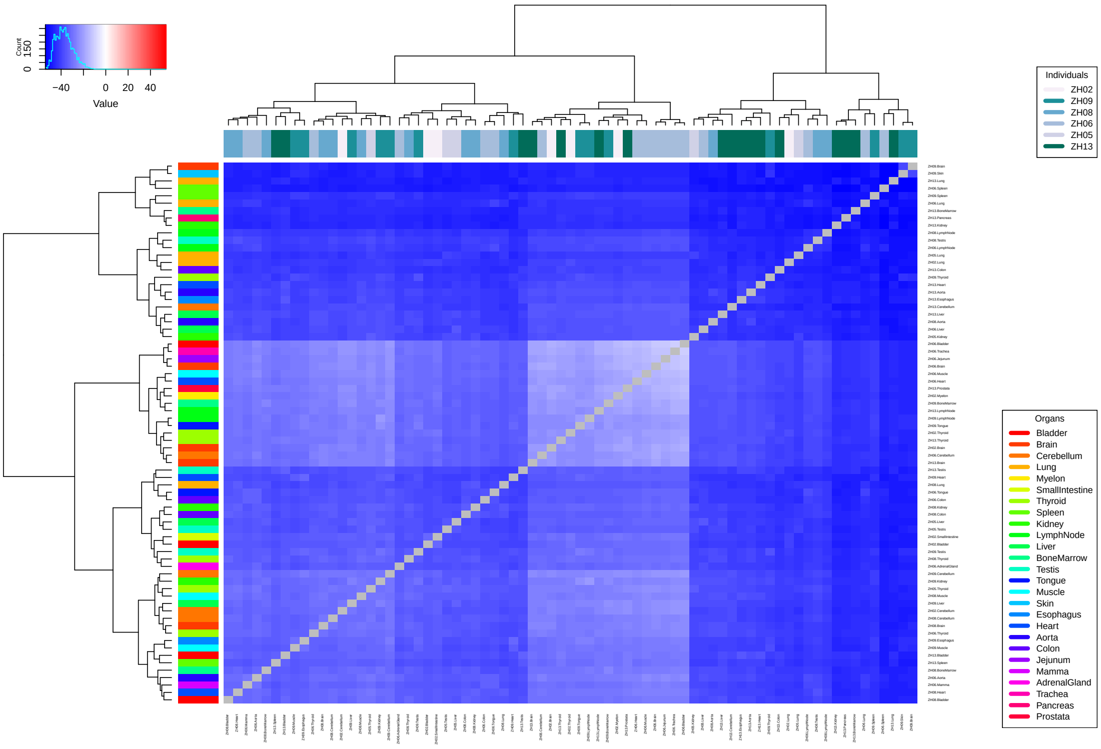


Figure C.17: Protein-based distance matrix heat map for HLA class II proteins with subtraction of the individual **median** protein occurrence. Pairwise distance was calculated using Euclidean-similarity score. The matrix is symmetrical and the upper colors show the individuals ID, whereas the left colors show the tissue type.

Appendix D

Supporting Tables

Algorithm D.1: MySQL query to select all peptides from all MS run and to calculate summarized properties. The peptides are grouped by each MS run. To calculate the different scores, the MySQL standard methods **MAX**, **MIN**, and **COUNT** are used.

```
SELECT
        distinct sp.sequence,
        length(sp.sequence),
        m.ms_run_id,
        s.source_id,
        max(sp.RT),
        min(sp.RT),
        max(sp.M),min(sp.M),
        max(sp.search_engine_score),
        min(sp.search_engine_score),
        max(sp.e_value),
        min(sp.e_value),
        max(sp.q_value),
        min(sp.q_value),
        count(sp.spectrum_hit_id)
from spectrum_hit sp
join
        ms_run m on m.ms_run_id = sp.ms_run_ms_run_id
join
        source s on s.source_id = m.source_source_id
join
        hla_map hm on hm.fk_source_id = s.source_id
join
        hla_type h on h.hla_type_id = hm.fk_hla_type_id
WHERE h.hla_string NOT LIKE 'D%'
group by m.ms_run_id, sp.sequence;
```

Table D.1: HLA class1 typing of the individuals contained in the HLA Ligand Atlas.

Individual	HLA-A	HLA-A	HLA-B	HLA-B	HLA-C	HLA-C
BD-ZH02	A*11:01	A*68:01	B*15:01	B*35:03	C*03:03	C*04:01
BD-ZH05	A*01:01	A*11:01	B*07:02	B*49:01	C*07:01	C*07:02
BD-ZH06	A*03:01	A*68:02	B*07:02	B*14:02	C*07:02	C*08:02
BD-ZH08	A*32:01	A*68:01	B*15:01	B*44:02	C*03:03	C*07:04
BD-ZH09	A*24:02	A*30:01	B*13:02	B*35:08	C*04:01	C*06:02
BD-ZH13	A*02:05	A*11:01	B*40:02	B*58:01	C*02:02	C*07:01

Table D.2: HLA class1 typing of the individuals contained in the HLA Ligand Atlas. Missing alleles are either caused by homozygosity or by missing sequencing.

Individual	HLA-DR	HLA-DR	HLA-DR	HLA-DQA	HLA-DQA	HLA-DQB	HLA-DQB
BD-ZH02	DRB1*04:01	DRB4*01:01		DQA1*03:01		DQB1*03:01	
BD-ZH05	DRB1*11:01	DRB1*14:01	DRB3*02:02	DQA1*01:01	DQA1*05:01	DQB1*03:01	DQB1*05:03
BD-ZH06	DRB1*13:03	DRB1*08:01	DRB3*01:01				
BD-ZH08	DRB1*13:03	DRB1*14:01		DQA1*05:05		DQB1*03:01	
BD-ZH09	DRB1*07:01			DQA1*02:01		DQB1*02:02	
BD-ZH13	DRB1*10:01	DRB1*15:01		DQA1*01:01	DQA1*01:02	DQB1*05:01	DQB1*06:02

Table D.3: List of shared peptides across all HLA class I samples.

Peptide	Occurrence	Occurrence [%]	Peptide	Occurrence	Occurrence [%]
AAMLDTVVFK	113	0.523148148	SVSNVVITK	144	0.666666667
AGDDAPRAVF	137	0.634259259	TPEEKSAVTAL	155	0.717592593
ASLSTFQQM	167	0.773148148	TVLTSKYR	118	0.546296296
ASVSTVLTSKY	108	0.5	VVYPWTQRF	174	0.805555556
AVALPLQTK	115	0.532407407	YASGRTTGIVL	109	0.50462963
DEVGGEALGRL	143	0.662037037	YASGRTTGIVM	152	0.703703704
ETFNTPAMY	124	0.574074074	GTMTGMLYK	116	0.537037037
FESFGDLSTPDA	110	0.509259259	IAVGYVDDTQF	150	0.694444444
FRLLGNVL	184	0.851851852	VAIQAVLSL	133	0.615740741
FTLGNVVGMY	113	0.523148148	VDIINAKQ	143	0.662037037
GTFGGLGSK	108	0.5	YEVSQLKD	113	0.523148148
GTYVSSVPR	116	0.537037037	EVGGEALGRL	130	0.601851852
KTYGEIFEK	120	0.555555556	PTTKTYFPHF	133	0.615740741
KVTEGSFVYK	119	0.550925926	VHLTPEEKSAVT	108	0.5
LASVSTVLTSKY	163	0.75462963	IYNEALKG	116	0.537037037
LLIENVASL	136	0.62962963	FGEHLLESDL	114	0.527777778
LPGQNEDLVLT	134	0.62037037	FLLFPDMEA	113	0.523148148
LRVAPEEHPTL	169	0.782407407	FRVVPQFVVF	119	0.550925926
LRVAPEEHPVL	183	0.847222222	KYPENFFLL	116	0.537037037
PENFRLLGNVL	147	0.680555556	LERMFLSF	135	0.625
RLFVGSIPK	142	0.657407407	LVGLFEDTNL	141	0.652777778
RVAPEEHPTL	146	0.675925926	MRYVASYL	152	0.703703704
RVAPEEHPVL	204	0.944444444	MRYVASYLL	153	0.708333333
SIFDGRVVAK	130	0.601851852	VAHVDDMPNAL	140	0.648148148
STIEYVIQR	108	0.5	YLVGLFEDTNL	114	0.527777778
SVQGIIIYR	109	0.50462963	YVAIQAVLSL	136	0.62962963

Table D.4: List of protein contaminant.

Protein	Gene	Protein	Gene	Protein	Gene
P30443	1A01	Q99878	H2A1J	P24844	MYL9
P04439	1A03	Q6FI13	H2A2A	A6NDD8	NBPFL
P13746	1A11	Q8IUE6	H2A2B	O15239	NDUA1
P30447	1A23	Q16777	H2A2C	O00483	NDUA4
P05534	1A24	Q7L7L0	H2A3	P0DJD8	PEPA3
P16188	1A30	Q9BTM1	H2AJ	P0DJD7	PEPA4
P30455	1A36	Q71UI9	H2AV	P0DJD9	PEPA5
P30464	1B15	P16104	H2AX	P62937	PPIA
P03989	1B27	P0C0S5	H2AZ	P62891	RL39
P18463	1B37	Q96A08	H2B1A	P22626	ROA2
Q04826	1B40	P33778	H2B1B	P62249	RS16
P30485	1B47	P62807	H2B1C	P62269	RS18
P62736	ACTA	P58876	H2B1D	P39019	RS19
P60709	ACTB	Q93079	H2B1H	P62308	RUXG
Q562R1	ACTBL	P06899	H2B1J	P06703	S10A6
P68032	ACTC	O60814	H2B1K	P05109	S10A8
P63261	ACTG	Q99880	H2B1L	P0DJI8	SAA1
P63267	ACTH	Q99879	H2B1M	P0DJI9	SAA2
P68133	ACTS	Q99877	H2B1N	Q01995	TAGL
P12235	ADT1	P23527	H2B1O	Q71U36	TBA1A
Q09666	AHNK	Q16778	H2B2E	P68363	TBA1B
P02647	APOA1	Q5QNW6	H2B2F	Q9BQE3	TBA1C
P02652	APOA2	Q8N257	H2B3B	Q13748	TBA3C
P02656	APOC3	P57053	H2BFS	P68366	TBA4A
P56381	ATP5E	P68431	H31	Q13885	TBB2A
P61769	B2MG	Q16695	H31T	Q9BVA1	TBB2B
P62158	CALM	Q71DI3	H32	Q13509	TBB3
Q5ZPR3	CD276	P84243	H33	P04350	TBB4A
P51911	CNN1	Q6NXT2	H3C	P68371	TBB4B
P23528	COF1	P62805	H4	P07437	TBB5
P09669	COX6C	P69905	HBA	Q9BUF5	TBB6
P15954	COX7C	P68871	HBB	P62328	TYB4
P10176	COX8A	P02042	HBD	O14604	TYB4Y
P31327	CPSM	P69891	HBG1	P0CG47	UBB
P17927	CR1	P69892	HBG2	Q96IX5	USMG5
P14406	CX7A2	Q9Y241	HIG1A	P08670	VIME
P81605	DCD	P04792	HSPB1	Q9HCL3	ZFP14
P17661	DESM	P52789	HXK2	Q9UII5	ZN107
Q16555	DPYL2	P02686	MBP	Q03924	ZN117
P04406	G3P	P10620	MGST1	Q99676	ZN184
P0C0S8	H2A1	P14174	MIF	O14709	ZN197
Q96QV6	H2A1A	P19105	ML12A	P0DKX0	ZN728
P04908	H2A1B	O14950	ML12B	Q8N4W9	ZN808
Q93077	H2A1C	P10916	MLRV	Q03923	ZNF85
P20671	H2A1D	Q9UKN1	MUC12		
Q96KK5	H2A1H	P60660	MYL6		

Table D.5: "Cancer-exclusive" & overall HLA ligand IDs on hematological malignancies. "Cancer-exclusive" peptides (overall HLA ligand IDs) identified from AML (n=16), CML (n=15), CLL (n=33) and MM/MCL (n=9/3) samples were annotated using NetMHCpan 3.0 and their respective sources' HLA type and only binding peptides (netMHC IC50âĽď500 nM and/or percentile rank ≤2) were retained for further analysis. Only samples expressing at least one of the 7 major HLA allotypes (A*01:01, A*02:01, A*03:01, A*24:02, B*07:02, B*08:01, B*18:01) are listed. For clustering, Jaccard distance graphs and overlap analysis of tumor-exclusive HLA ligand datasets, HLA ligands occurring only once across all samples were discarded and only samples containing ≥5 unique HLA class I ligands were included. Reproduced with permission from Backert et al.[12]

Sample	HLA Typing	Sample							
	A*01:01	A*02:01	A*03:01	A*24:02	B*07:02	B*08:01	B*18:01		
AML01	A*02, A*11, B*35, B*44	0.9	-	4 (57)	-	-	-	-	-
AML04	A*03:01, B*39:01, B*51:01	1.0	-	-	87 (438)	-	-	-	-
AML05	A*02, B*07, B*40	0.4	-	263 (641)	-	-	139 (447)	-	-
AML06	A*02, A*03, B*44:25, B*52:15,	1.6	-	19 (164)	118 (535)	-	-	-	-
AML11	A*02:01, A*03:01, B*38:01, B*44:02,	2.8	-	28 (235)	41 (373)	-	-	-	-
AML14	A*03:01, A*26:01, B*35:01, B*38:01,	1.7	-	-	11 (211)	-	-	-	-
AML15	A*02, A*66, B*40 , B*15,	8.4	-	112 (560)	-	-	-	-	-
AML36	A*02:01, A*23:01, B*44:02 , B*49:01	0.4	-	23 (208)	-	-	-	-	-
AML37	A*02:01, B*13:02, B*51:01	0.2	-	122 (610)	-	-	-	-	-
AML48	A*02:01, A*03:01, B*18:01	2.6	-	10 (153)	19 (218)	-	-	-	212 (511)
AML49	A*02, A*26:01, B*27:05	4.9	-	1 (28)	-	-	-	-	-
AML59	A*02, A*24, B*44, B*50	19.0	-	27 (29)	-	45 (294)	-	-	-
AML64	A*01:01, A*23:01, B*44:03	0.5	10 (175)	-	-	-	-	-	-
AML65	A*02:01, A*11:01, B*08:01, B*57:01	0.5	-	11 (209)	-	-	-	15 (154)	-
AML66	A*01, A*02, B*08, B*13	0.5	34 (255)	132 (600)	-	-	-	115 (322)	-
AML70	A*03:01, A*32:01, B*57:01, B*35:01	0.5	-	-	6 (143)	-	-	-	-
CML01	A*23, A*25, B*18, B*57	0.9	-	-	-	-	-	-	11 (81)
CML02	A*03, B*35, B*52	4.8	-	-	127 (691)	-	-	-	-
CML03	A*01, B*08	0.4	22 (228)	-	-	-	-	40 (132)	-
CML04	A*03, A*68, B*07, B*44	5.0	-	-	55 (294)	-	30 (216)	-	-
CML05	A*24, A*28, B*27, B*57	2.9	-	-	-	11 (148)	-	-	-
CML06	A*01, A*03, B*07, B*08	5.0	8 (103)	-	17 (145)	-	27 (205)	19 (143)	-
CML07	A*02, A*03, B*13, B*15	5	-	58 (382)	63 (387)	-	-	-	-
CML08	A*01, A*02, B*08, B*50	6	33 (212)	43 (165)	-	-	-	17 (72)	-
CML09	A*03, B*07, B*35	1,3	-	-	32 (287)	-	19 (170)	-	-
CML10	A*02, A*11, B*35, B*44	0.5	-	159 (570)	-	-	-	-	-
CML13	A*02:01, A*03:01, B*07:01, B*51:01	2.0	-	50 (340)	34 (364)	-	0 (0)	-	-
CML15	A*01:01, A*02:01, B*51:01	0.5	8 (86)	7 (118)	-	-	-	-	-
CML16	A*02:01, A*11:01, B*07	0.5	-	20 (197)	-	-	20 (141)	-	-
CML18	A*02, B*18, B*57	17	-	76 (396)	-	-	-	-	67 (217)
CML19	A*01:01, B*15:03	7.6	29 (132)	-	-	-	-	-	-
CLL02	A*02, A*11, B*39, B*40	20.0	-	22 (154)	-	-	-	-	-
CLL04	A*02:01, B*35:01, B*39:01	6.2	-	30 (207)	-	-	-	-	-
CLL08	A*25, A*26, B*18, B*38	1.8	-	-	-	-	-	-	184 (487)
CLL09	A*02, B*55, B*57	2.4	-	15 (109)	-	-	-	-	0 (0)
CLL10	A*02, A*23, B*15, B*41	5.8	-	32 (269)	-	-	-	-	-
CLL12	A*01, A*24, B*08, B*27	11.5	104 (484)	-	-	159 (555)	-	244 (534)	-
CLL13	A*02, A*03, B*18, B*35	5.2	-	12 (154)	30 (322)	-	-	-	107 (332)
CLL16	A*24, A*31, B*15, B*38	6.4	-	-	-	0 (33)	-	-	-
CLL17	A*03, A*30, B*07, B*13	0.8	-	-	231 (764)	-	269 (772)	-	-
CLL18	A*02, A*03, B*07, B*55	2.6	-	55 (294)	50 (390)	-	344 (983)	-	-
CLL20	A*01, A*02, B*27, B*37	0.5	8 (132)	19 (192)	-	-	-	-	-
CLL21	A*02, A*68, B*15, B*27	3.6	-	88 (474)	-	-	-	-	-
CLL27	A*02, A*03, B*35, B*57	0.9	-	2 (80)	3 (93)	-	-	-	-
CLL28	A*02, B*15, B*44	2.8	-	22 (266)	-	-	-	-	-
CLL30	A*24, B*07, B*49:01	2.4	-	-	-	26 (161)	36 (177)	-	-
CLL32	A*01, A*68, B*08, B*44	2.0	12 (197)	-	-	-	-	57 (193)	-
CLL34	A*02, A*03, B*40	1.4	-	21 (169)	4 (171)	-	-	-	-
CLL37	A*02, A*11, B*35, B*37	0.4	-	0 (24)	-	-	-	-	-
CLL38	A*01, A*03, B*08, B*51	0.8	14 (193)	-	12 (282)	-	-	114 (293)	-
CLL41	A*02, A*03, B*07, B*44	1.0	-	9 (120)	8 (196)	-	56 (401)	-	-
CLL43	A*24, B*35, B*50	1.5	-	-	-	2 (90)	-	-	-
CLL47	A*24, A*32, B*27, B*51	1.5	-	-	-	7 (91)	-	-	-
CLL49	A*03, B*07	1.2	-	-	34 (218)	-	53 (216)	-	-
CLL52	A*01, A*02, B*08, B*13	0.9	9 (183)	70 (369)	-	-	-	39 (211)	-
CLL55	A*03, A*26, B*07, B*08	2.0	-	-	36 (325)	-	221 (791)	233 (665)	-
CLL56	A*01, A*32, B*07, B*44	120.0	75 (277)	-	-	-	260 (759)	-	-
CLL59	A*02, A*03, B*35, B*40	3.9	-	9 (122)	26 (304)	-	-	-	-
CLL60	A*02, A*24, B*51, B*57	1.9	-	7 (104)	-	17 (189)	-	-	-
CLL70	A*02, A*03, B*40, B*44	0.2	-	29 (280)	8 (186)	-	-	-	-
CLL71	A*02, A*11, B*35	0.5	-	33 (179)	-	-	-	-	-
CLL72	A*02, B*08, B*51	0.5	-	30 (292)	-	-	-	60 (273)	-
CLL80	A*02, A*24, B*51	0.2	-	9 (79)	-	3 (76)	-	-	-
CLL84	A*02, B*40, B*51	1	-	7 (140)	-	-	-	-	-
MM34	A*01, A*24, B*08, B*18		2 (72)	-	-	0 (34)	-	1 (7)	3 (92)
MM36	A*01, A*02, B*08, B*37	1.2	12 (106)	9 (143)	-	-	-	7 (57)	-
MM37	A*02, A*33, B*15, B*18	1.8	-	27 (248)	-	-	-	-	120 (542)
MM38	A*03, A*26, B*40, B*55	1.2	-	-	39 (362)	-	-	-	-
MM39	A*02, A*24, B*07, B*27	4.5	-	32 (275)	-	129 (360)	74 (362)	-	-
MM40	A*02,A*03, B*07, B*35	0.1	-	1 (48)	1 (69)	-	9 (122)	-	-
MM49	A*24, A*25, B*39, B*40	1.3	-	-	-	49 (347)	-	-	-
MM50	A*02, B*07, B*44	0.5	-	54 (541)	-	-	26 (285)	-	-
MM56	A*03, A*33, B*07	0.4	-	-	12 (128)	-	59 (310)	-	-
MM1S	A*23:01, A*24:02, B*18:01, B*42:01	2	-	-	-	494 (953)	-	-	171 (372)
U266	A*02:01, A*03:01, B*07:02, B*40:01	2	-	246 (573)	260 (657)	-	581 (1084)	-	-
JJN3	A*03:01, A*33:01, B*07:02, B*14:02	2	-	-	295 (810)	-	308 (750)	-	-