PhD Dissertation

Leon Clemens Bichmann

Mass spectrometry guided immunoinformatics

LEON CLEMENS BICHMANN

MASS SPECTROMETRY GUIDED IMMUNOINFORMATICS

MASS SPECTROMETRY GUIDED IMMUNOINFORMATICS

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. Leon Bichmann
aus Frankfurt am Main

Tübingen
2021

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

Tag der mündlichen Qualifikation:	21.05.2021
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Oliver Kohlbacher
2. Berichterstatter:	Prof. Nico Pfeifer

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

"Mass Spectrometry guided Immunoinformatics"

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Unterschrift Leon Bichmann: Tübingen, 02.02.21, Leon

Abstract

A key task of the human immune system is the recognition and surveillance of peptides presented by the HLA complex on the surface of body cells. In this way, abnormalities can be discovered rapidly to elicit targeted immune responses. The identification of the HLA-presented immunopeptidome is thus of tremendous interest for research questions ranging from basic immunological processes to the design of immunotherapies such as vaccinations against infectious diseases and cancer. With the advancement of technical developments in biological high-throughput methods such as mass spectrometry it has become possible to identify thousands of sequences of HLA-presented peptides from a single sample of cells or human tissues. This has enabled researchers to directly investigate the peptide sequences presented in the human body and gain information on their properties. However, the acquisition and evaluation of large amounts of mass spectrometry measurements and HLA peptide sequence characteristics is a highly complex task that requires the development of sophisticated experimental and computational methods. This research work has focused on the evaluation and improvement of existing methodology to identify HLA-bound peptides and further to investigate various aspects of the immunopeptidome presented by human non-malignant and cancer tissues. An essential part of this effort was the development of novel automated, digital processing pipelines for HLA immunopeptidomics data. Specifically, two pipelines - "MHCquant" that achieved superior sensitivity in contrast to existing software solutions and "DIAproteomics" that allowed to explore the application of the novel method of data-independent acquisition to immunopeptidomics were developed. Application of the "MHCquant" pipeline to the currently largest existing immunopeptidomics data set of human non-malignant tissues, allowed to construct the novel data resource "The HLA Ligand Atlas". This benign reference data set is of great significance for the comparison with diseased-state tissues and was thoroughly evaluated for differences across the human population, tissue specificity and the presence of cryptic peptides from non-canonical genomic origins. Finally, the HLA immunopeptidome of multiple clinical hepatocellular carcinoma samples was analysed in combination with next generation genomic sequencing measurements in an in-depth multi-omics approach in order to discover tumor-associated mutated antigens as suitable targets for cancer immunotherapy. While the effort did not result in the determination of particular mutated antigens, it was possible to pinpoint tumor somatic mutations that are likely presented as epitopes. Ultimately, the missing findings are discussed as a consequence of technological limitations and the low mutational burden of hepatocellular carcinoma. The developed computational workflows as well as the investigated data sets were made publicly available to serve the scientific community of future generations as a standard to reanalyze and compare novel results with and advance the holistic understanding of immunological processes in the human body.

Zusammenfassung

Eine zentrale Komponente des humanen Immunsystems ist die Erkennung und Überwachung von Peptiden, die durch den HLA Komplex auf der Oberfläche von körpereigenen Zellen präsentiert werden. Auf diese Weise wird es ermöglicht Abnormalitäten frühzeitig zu erkennen und zielgerichtete Immunantworten auszulösen. Die Identifizierung des HLA-präsentierten Immunopeptidoms ist daher von enormen Interesse für Fragestellungen in der Erforschung von grundlegenden immunologischen Prozessen, sowie der Konzipierung von Immuntherapien wie z.B. Impfstoffen gegen Infektionskrankheiten und Krebs. Im Zuge der voranschreitenden technologischen Entwicklungen biologischer Hochdurchsatz-Methoden wie z.B. der Massenspektrometrie, ist es möglich geworden, tausende Sequenzen HLA-präsentierter Peptide aus einer einzelnen Zell- oder menschlicher Gewebeprobe zu bestimmen. Dies erhebt Forscher in die Position, die Peptidsequezen des humanen Körpers unmittelbar zu identifizieren, nachzuverfolgen und Informationen über deren Eigenschaften zu gewinnen. Nichtsdestotrotz, ist die Erhebung und Auswertung grosser Mengen an Massenspektrometrie-Messungen, sowie die Charakterisierung der HLA-Peptidsequenzen eine hoch komplexe experimentelle sowie Computer-gestützte Herausforderung. Diese Forschungsarbeit hat daher die Evaluierung und Verbesserung der bestehenden Methodik zur Identifizierung von HLA-Peptiden, sowie die Erforschung der Aspekte des Immunopeptidoms von gesunden und Krebs erkrankten humanen Geweben zum Thema. Ein wichtiger Teil dieser Arbeit bestand dabei in der Entwicklung neuer, automatisierter, digitaler Prozessierungs-Pipelines für HLA-Immunopeptidomik Daten. Insbesondere die beiden Pipelines "MHCquant", welche eine höhere Sensitivität im Gegensatz zu existierenden Software Lösungen aufzeigen konnte sowie "DIAproteomics", mit welcher die neuartige Methode der daten-unabhängigen Aufnahme für die Immunopeptidomik exploriert werden konnte, wurden hierfür entwickelt. Weiterhin wurde durch Anwendung der "MHCquant" Pipeline der derzeitig grösste Immunopeptidomik Datensatz an humanen, nicht-erkrankten Geweben erhoben und in der Datenbank "HLA Ligand Atlas" zusammengefasst. Diese gesunde Referenz, welche eine starke Bedeutung für den Vergleich mit erkrankten Gewebeproben hat, wurde vielseitig ausgewertet im Hinblick auf Unterschiede innerhalb der humanen Bevölkerung, Gewebespezifizität und der Präsenz kryptischer Peptide aus nicht-kanonischen genomischen Regionen. Im letzten Teil wurde die Arbeit durch die Analyse eines Satzes klinischer hepatozellulärer Karzinom-Gewebeproben ergänzt, welche in Kombination mit genomischen Sequenzierungs-Methoden der nächsten Generation durch einen Multiomics-Ansatz tiefgehend durchsucht wurden, um Tumor-assoziierte Antigenzielstrukturen für die Krebsimmuntherapie zu definieren. Während hierbei keine mutierten Antigene identifiziert werden konnten, war es möglich einige Tumor-somatische Mutationen zu detektieren, welche durch eine hohe Wahrscheinlichkeit zur HLA-Präsentation herausstachen. Deren fehlende direkte Detektion wurde schliesslich im Hinblick auf die Limitierung der Technologie sowie der niedrigen Mutationslast hepatozellulärer Karzinom Gewebe diskutiert. Die entwickelten Computermethoden sowie die erforschten Datensätze wurden veröffentlicht, um sie der Forschungsgemeinschaft für zukünftige Generationen als Standard für Re- und Neuanalysen bereitzustellen, sowie um das Verständnis immunologischer Prozesse im humanen Körper voranzubringen.

Acknowledgements

To my advisor and main supervisor Prof. Oliver Kohlbacher for providing me the possibility to engage into a bioinformatics PhD program at the University Tübingen, his support, guidance and funding, I would like to express my sincere gratitude.

I would like to thank all my research fellows at the department of computer science – especially Timo Sachsenberg, Oliver Alka, Fabian Aicheler, Christopher Mohr, Andreas Friedrich, Bilge Surun, Julianus Pfeuffer, Eugen Netz, Mark Rurik, Leon Kuchenbecker and Alexander Peltzer – for supporting me scientifically, as well as personally and letting me contribute on the developments of the OpenMS software toolbox for computational mass spectrometry. In particular, I enjoyed participating into the Germany- and Europe-wide community efforts and workshops to advance ongoing software developments as well as teaching of current functionalities to other researchers in the field.

To my co-supervising group leaders Prof. Hans-Georg Rammensee and Prof. Stefan Stevanovic, at the interfaculty department of immunology at the University Tübingen, I would like to thank you for opening my eyes for the research on the human immune system and therapeutic possibilities.

Likewise, I would like to thank my collaborating research colleagues in the department of immunology – especially Ana Marcu, Michael Ghosh, Lena Mühlenbruch, Lena Freudenmann, Jens Bauer, Maren Lübke Markus Löffler, Tatjana Bilich, Annika Nelde and Daniel Kowalewski – for the fruitful and lengthy discussions in numerous meetings, peptide and journal clubs and fun activities. In particular, I would like to thank Ana Marcu and Leon Kuchenbecker for jointly working on the highly challenging effort to fully analyze and complete the HLA Ligand Atlas project.

Last but not least I would like to express my gratitude to Hannes Röst and Shubham Gupta at the University Toronto for their support and effort to work together on the analysis of immunopeptidomics DIA MS data.

Finally, I would like to thank Gisela Gabernet Garriga and my family for continuously motivating and supporting me personally strongly throughout my entire PhD work.

Contents

Notation

Frequently used abbrevations

MHC	The major histocompatibility complex
HLA	The human leukocyte antigen receptor class I or II
MS	Mass spectrometry
LC	Liquid chromatography
RT	Retention time
m/z	Mass-to-charge ratio
SIM	Selected ion monitoring
PRM	Poly reaction monitoring
NGS	Next generation sequencing
DDA	Data dependant acquisition
DIA	Data independant acquisition
CID	Collision induced dissociation
HCD	High-energy collision dissociation
ESI	Electrospray ionisation
PSM	Peptide spectrum match
XIC	Extracted ion chromatogram
ID	Peptide identification
FDR	False discovery rate
DB	Database
LFQ	Label free quantification
SVM	Support vector machine
HPC	High performance computing
$PBMC$	Peripheral blood mononuclear cells
NAT	Normal tissue adjacent to the tumor
TAA	Tumor-associated antigen
GO	Gene ontology
Var	Genetic variant
PNE	Predicted neoepitope
WT	Wildtype
HCC	Hepatocellular carcinoma
Mel	Melanoma
TMB	Tumor mutational burden

CHAPTER 1

Introduction

1.1 *Motivation*

The immune system is one of the most essential components of an organism, defending it against the diversity of invading pathogens. All multicellular but some single cellular organisms too, have developed an immune system that will guarantee survival of its species for a given life span. The two branches of innate and adaptive immunity have evolved subsequently and are shared similarly by humans and a wide span of organisms rendering it a very well conserved and powerful system [1, 2]. All the various kinds of pathogens such as bacteria, viruses, fungi or even physical pathogens are upon discovery attacked by the immune system and have in turn shaped their infection in order to evade the host immune response as well as possible [3]. Medicine and drug development have therefore traditionally focused on supporting the immune system with varieties of antibiotic, -viral and -fungal compounds or biologics. Upon treatment of sick patients with these therapies the bacterial, viral or fungal load in their bodies will be reduced [4]. In this way the exhausted immune system is able to recover and successfully defeat the remaining pathogens. In addition, the immune system has the extraordinary capability to establish a biological memory of previous pathogenic infections and acquire long term immunity. An initially acquired immune response can then prevent a second infection with the same pathogenic organism in the future [5]. The early discovery of this effect by Louis Pasteur and Edward Jenner was the foundation of vaccination technology, which has greatly improved the life expectancy and lead to the nearly complete extinction of several widespread viral infections among the world human population [6].

However, non-infectious diseases for example inherited diseases such as genetic and age-related dysfunction or cancer are difficult to be recognized by the immune system as it usually tolerates self antigens [7]. Genetic mutation-driven cancer is one of the main non-pathogenic disease causes of death in the human population and numbers are expected to rise within the coming decades [8]. While a multitude of different approaches have been developed in order to treat cancer ranging from surgery to radio-, chemo- or targeted pharmaceutical therapies, no single entirely successful therapy exists to treat the various types and modalities of cancer in humans. Yet, it has been shown that even though non-pathogenic, various cancer ontologies are not invisible to the immune system. This has given rise to the modern field of immunotherapy trying to actively stimulate a patient's immunity in the course of a cancer infection or maybe even preventively [9]. In order to achieve this stimulation, different approaches have been proposed [10]. A promising strategy are epitope-based multipeptide vaccines, that achieve immune stimulation through exposure to the peptide biomolecules (epitopes) that are presented through the HLA cell surface receptor to the adaptive immune system [11, 12].

The state-of-the-art approach to identify these immunogenic epitopes is immunoaffinity purification of HLA-peptide receptor complexes, followed by LC-MS/MS measurements and labor-intensive T-cell immunogenicity testing [13]. With the rise of high-throughput biological MS technology, thousands of epitopes can be identified from a measurement and hundreds of MS runs can be taken within 2-3 weeks using a single MS instrument [14]. Hence, this technology has enabled scientists to attempt to measure almost all active biomolecules exposed to the immune system. The exact knowledge of these biomolecules and their sequences, in particular HLA-presented peptides, is of profound importance for basic research in immunology and provides a major advancement for the entire field of immunotherapy far beyond epitope-based vaccination [15].

The increase in the number of complex measurements produced by high-throughput methods such as biological MS also requires scalable automated computational methods to analyze the massive amount of data [16]. Few computational methods are currently able to handle MS data on a large scale, as most were developed in a time when the analysis was carried out manually on single processing computers instead of high-performance computing (HPC) clusters [17, 18]. Moreover, robust storage of the data is a challenge, since one measurement alone can take up to several Giga bytes in size [19]. Evaluating immunopeptidomics MS measurements is especially difficult in contrast to common bioanalytical procedures in protein and peptide MS in other contexts. This is due to the fact, that HLA-presented peptides are unspecifically cleaved and thus span a very large sequence search space when comparing spectra to possible peptide matches of the human proteome. As a consequence, the sensitivity-specificity trade-off between robust false discovery rates and identifying all peptides of interest in every MS run is particularly cumbersome. Hence, at the start of this research work, to the best of my knowledge no method existed that was tailored to deal with immunopeptidomics analysis, but by the end of this research work several methods had been proposed. [20–23]

Along these lines, this thesis focuses in four chapters on the computational analysis of large-scale clinical MS data in the context of cancer immunotherapy. Next to the development of scalable high-throughput computational methods, a unique data set of thousands of measurements of for the immune system highly relevant HLA-presented peptide biomolecules of multiple human tissues of several healthy donors was analysed. Ultimately, this work aims to contribute to the basic understanding of the human immune system and benefit future research on the therapy of a diverse range of diseases, in particular cancer.

The first two chapters focus on the methodological side of the analysis of immunopeptidomics experiments, comparing benefits and disadvantages of existing methodology and establishing novel computational analysis workflows and demonstrating their capabilities. In particular two new robust data analysis pipelines were implemented "MHCquant" and "DIAproteomics". MHCquant enables researchers to pipeline the standard immunopeptidomics MS measurements with higher throughput and higher sensitivity than previously existing methods. DIAproteomics allows to analyse protein and peptide MS measurements acquired using the modern, different technique of data-independent MS acquisition having several advantages in terms of reproducibility and identification depth over the standard approach.

The following two chapters show the results of large studies with clinical patient material. An emphasis is given on the discovery of HLA presented peptides in human tissues and on tumor immunology. Hereby multiple immunological research questions were adressed such as the variability of the immunopeptidome across tissues and individuals within the human population and the feasibility of identifying tumor-associated antigens. Hepatocellular carcinoma, a disease with limited treatment options [24, 25] was chosen as an example and tissue samples of a cohort of cancer patients were investigated using immunopeptidomics analysis. In particular the samples were searched for the presence of tumor-specific mutated neoantigens as suitable targets for personalized cancer immunotherapy.

Finally, all research projects are summarized and related to each other in a last concluding chapter. Ultimately, a concise outlook on future developments in the field is given towards the end.

CHAPTER 2

Background

2.1 *Immuno-Oncology*

This section highlights general principles and concepts of the human immune response and the various cellular components that mediate the corresponding effects. In addition, the cell surface protein complex - the major histocompatibility complex (MHC) is introduced in detail and its central function within the immune system. Ultimately, an outline is given on how therapeutic vaccination can activate the immune system in a clinical context towards recent developments to treat infectious diseases or cancer.

2.1.1 *The human immune system*

Anatomic and physiologic defense

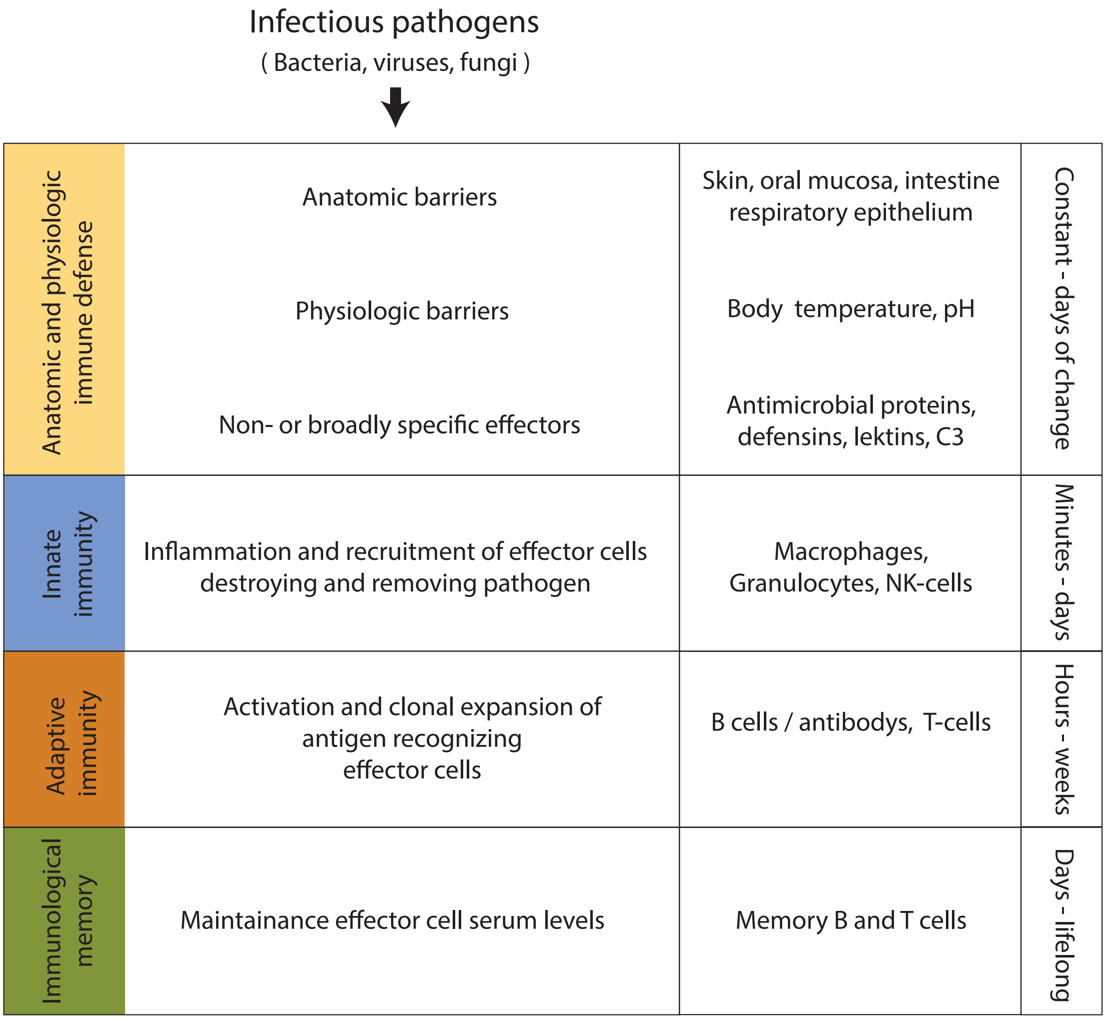
In humans, the immune system is composed of multiple layers of response to pathogens. The most immediate and a constantly active part of the immune system are the anatomic barriers at the periphery of human bodies such as the skin, the oral mucosa, intestine and respiratory epithelium. These physical barriers have evolved to make it difficult for pathogens to penetrate towards the inside of the body. In addition, fluids covering the surfaces of these peripheries contain non- or broadly specific antimicrobial enzymes and proteins that are also referred to as part of the mucosal immune system. Ultimately, physiological defense related reactions of the body to combat an infection such as elevated body temperature (fever) and the acidic pH of the stomach preventing some pathogens from entering the intestine are considered part of the human immune defense [26–28] (Figure 2.1).

Innate immunity

Despite anatomic and physiological defense against infection, the immune responses are grouped into two major branches – innate and adaptive immunity. Innate immunity is considered to be the older evolutionary conserved branch [1]. The first reactions of inflammation in an infected area, phagocytosis and the recruitment of different primary effector cells belong to this category. In particular macrophages or granulocytes that non-specifically incorporate and digest pathogens, are first responses of innate immunity. Moreover, natural killer (NK) cells, when recruited to the inflammation sites, are broadly targeting abnormal cells for example infected or mutated cells of changed phenotype and induce cell death in those cells. While the response typically arises within minutes after infection it can last for multiple days [2, 26–28].

Adaptive immunity

The adaptive immune response is usually triggered within hours – days later than the initial innate inflammation reaction. This is due to the fact that it involves very specialized B- and T-cells that first have to be selected from a pool of vast geneticly diverse immune cells. Upon exposure to the respective antigens, only those cells that are able to specifically recognize antigens of the infecting pathogen may differentiate and expand to form an active immune response. Throughout this process, antigen presenting cells migrate from the infected area to



Infectious pathogens

(Bacteria, viruses, fungi)

Anatomic and physiologic immune defense	Anatomic barriers	Skin, oral mucosa, intestine respiratory epithelium	Constant - days of change
	Physiologic barriers	Body temperature, pH	
	Non- or broadly specific effectors	Antimicrobial proteins, defensins, lektins, C3	
Innate immunity	Inflammation and recruitment of effector cells destroying and removing pathogen	Macrophages, Granulocytes, NK-cells	Minutes - days
Adaptive immunity	Activation and clonal expansion of antigen recognizing effector cells	B cells / antibodys, T-cells	Hours - weeks
Immunological memory	Maintainance effector cell serum levels	Memory B and T cells	Days - lifelong

Figure 2.1: The layers of the immune defense of the human body against infection and corresponding active response mediating cells and biomolecules, as well as time scales of the involved processes are illustrated as table. (Table adapted from [26])

the peripheral organs of the lymphatic system. There they can initiate the clonal expansion of specialized B- and T-lymphocytes that have matching receptors that are able to bind pathogen antigens. In the case of T- cells antigens are presented as peptide fragments of their source protein through the MHC receptor presentation pathway. Once expanded, the specialized B- and T-cells migrate to the infection site and target the infecting pathogens causing them to undergo cell death. Consequently, the adaptive immune response is more complex but very specific. By the activation of B and T memory cells, serum levels of the selected B- and T-cells that were able to target the infecting pathogen can be maintained for long time frames. This thus provides a way for the body to rapidly combat and prevent a second large infection with the same pathogen and result in up to lifelong immunity [26–28].

The lymphatic system

All immune cells involved in the innate or adaptive immune system are distributed over the blood stream and the lymphatic system. Similar to the blood vessels the lymph vessels span the entire human body to provide immune cells to and take up antigens from every possible location. The main central lymphoid organs are the thymus and the bone marrow, as they contain immune progenitor cells for T- and B- cells respectively. These stem cell like progenitors are able to mature into a variety of lymphocytes and possess the ability to develop receptors that can target any kind of antigen. The peripheral lymphoid organs comprise the lymph nodes and other mucosal tissues such as the tonsil, intestine, peyer's patch or bronchus-associated lymphoid tissue (BALT) [29]. Mature naive lymphocytes derived from the progenitor cells recirculate between the peripheral immune tissues and can be activated upon antigen encounter. In order to present antigens from infection sites to the immune system, the lymph fluid, a mixture of antigen presenting cells and circulating lymphocytes from the extracellular matrix of peripheral tissues is continuously drained through the lymphatic system [26–28].

2.1.2 *T cell-mediated adaptive immunity*

The T-cell repertoire

The development of a person's T-cell repertoire and its maturation in order to acquire the capability to recognize a certain antigen takes place in the thymus. It involves the maturation of T-cell clones of innumerable genetic diversity by combinatorial genetic recombination during childhood. Throughout this process, the clones are selected in two stages: (1) positive selection - by recognizing the MHC-I or -II complex it is ensured that only T-cells that are capable of binding to the MHC survive and (2) negative selection - the T-cell clones that bind the MHC presented peptides of the host organism itself too strongly are eliminated, in order to circumvent autoimmunity. The remaining T-cell clones are thus able to recognize only foreign MHC presented antigens and carry a large diversity of T-cell receptors due to their genetic diversity. During homeostasis, the individual T-cell clones exist in low copy numbers. However, upon activation of particular T-cell clones by recognition of antigens with their specific receptors during an infection, their clonal expansion can be triggered to produce millions of cells and elevate their blood serum levels [26–28].

There are two distinct classes of cells that drive T-cell mediated adaptive immunity and many other types of lymphocytes orchestrate the process. Both classes of cells can be distinguished by the CD co-receptor molecules carried on their cell surface.

CD8-positive T-cells

CD8+ T-cells carry the CD8 co-receptor of the T-cell receptor and are mainly involved in the recognition of intracellular antigens presented by the MHC class I complex. Upon activation and recognition of foreign antigens, CD8+ T-cells are in most cases cytotoxic. Consequently, they will trigger a controlled cell death of respective recognized cells such as for example virus-infected cells [26–28] (Figure 2.2 A).

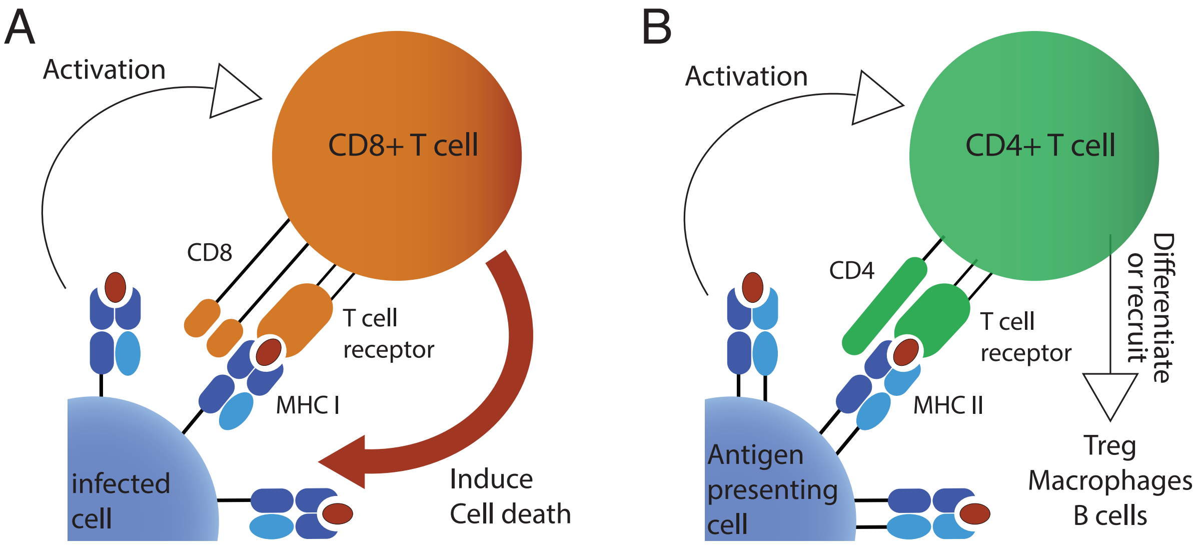


Figure 2.2: A) CD8+ T-cell activation: The CD8 and T-cell receptor co-recognize the MHC class I - peptide complex. Activation upon recognition of a foreign antigen peptide induces cytotoxic activity of the T-cell towards the presenting cell. B) CD4+ T-cell activation: The CD4 and T-cell receptor co-recognize the MHC class II - peptide complex. Activation upon recognition of a foreign antigen peptide can result in the recruitment of regulatory T-cells, cytotoxic B-cells or macrophages.

CD4-positive T-cells

CD4+ T-cells carry the CD4 co-receptor of the T-cell receptor on their surface and are mainly involved in the recognition of extracellular antigens presented by the MHC class II receptor complex. Activation upon antigen recognition in contrast to CD8+ T-cells in most cases does not lead to cytotoxic activity. In contrast, activation of CD4+ T-cells can stimulate their differentiation into various different types of T helper cells, depending on different factors such as the presence of certain cytokines. T helper cells are mainly involved in recruiting other cells to an infection site for example B-cells, macrophages or other phagocytes. In addition, regulatory T-cells (T_{reg}) are a class of T helper cells that are able to down regulate cytotoxicity of CD8+ T-cells preventing for example autoimmunity [26–28](Figure 2.2 B).

Antigen presenting cells

Ultimately, T-cell activation relies on another important class of lymphocytes, which are referred to as professional antigen presenting cells (APCs) - for example macrophages and dendritic cells (DCs). Both are able to take up antigens or even entire cells, digest them and present peptide fragments of those antigens via their surface MHC receptors. By this they are able to prime and activate T-cells and simulate their differentiation into the respective subtype [26–28].

2.1.3 *The MHC receptor complex and antigen presentation*

MHC protein function

The major histocompatibility complex (MHC) is a cell surface protein complex and is referred to as human leukocyte antigen (HLA) complex in humans. Its function to present peptide fragments of antigens to the surveilling T-cells is a central component of the adaptive immune response. The two different classes of receptors (HLA class I and HLA class II) are inherently linked to the T-cell populations (CD8+ and CD4+) that interact with them via their respective T-cell receptors. Moreover, peptides are loaded onto the receptors of the respective HLA class via two distinct molecular pathways. Peptides that stem from intracellular proteins are mainly presented through HLA class I receptors and peptides that stem from extracellular proteins are mainly presented by HLA class II receptors. Both HLA I and II share the structural feature of forming a concave peptide binding pocket and being anchored into the membrane via lipophilic membrane-spanning domains [28].

Genetic origin and diversity

The HLA I complex is encoded by either the HLA-A, B or C gene locus and the invariant $\beta 2$ microglobulin gene giving rise to at most six different HLA I gene products in a single human if all genes are heterozygous. In contrast, the HLA II complex is encoded by the DR, DP and DQ gene loci. It functions as a dimer of two homogenous protein chains out of these loci, creating a high number of possible combinatorically assemblies. The HLA genomic region is the most polymorphic region in the human genome and thus within the human population there exists a vast genetic diversity for HLA I and II molecules that vary in single nucleotide polymorphisms at different locations. Consequently, in the diverse human populations across the world different genetic traits for the HLA gene are over- or underrepresented [30].

HLA I protein structure characteristics

The HLA I complex is composed of a large α and the small $\beta 2$ microglobulin protein chains froming four different structural domains ($\alpha_{1\text{-}3}$ and $\beta\ _1$). (Figure 2.3 A) Its binding pocket is formed exclusively by the α chain and requires peptides to fit tightly within the pocket and restricts their length to a short range of about 8-12 amino acids [28] (Figure 2.3 B). Across the diversity of HLA-I receptors, most peptides are intensively bound via specific interaction with two defined anchor residues at the beginning (position 2) and the C-terminal end of the peptide sequence. Sequence motifs derived from bound peptides clearly display the specificity of the dominantly interacting two anchor residues [31–33] (Figure 2.3 C).

HLA II protein structure characteristics

The HLA II complex assembles as a dimer of nearly equally sized α and β chains forming four structural domains, too ($\alpha\ _{1\text{-}2}$ and $\beta\ _{1\text{-}2}$). (Figure 2.3 A) The binding pocket is formed at the interface of the two protein chains. The pocket is less tight and allows also longer peptides of about 8-25 amino acids to bind into the groove with their ends bulging out of the pocket at both sides [28] (Figure 2.3 B). Most interactions occur therefore between the receptor and a binding core located towards the center of HLA-II bound peptide sequences. Binding motifs of these peptide sequence cores are less clearly defined by specific residues and positions and

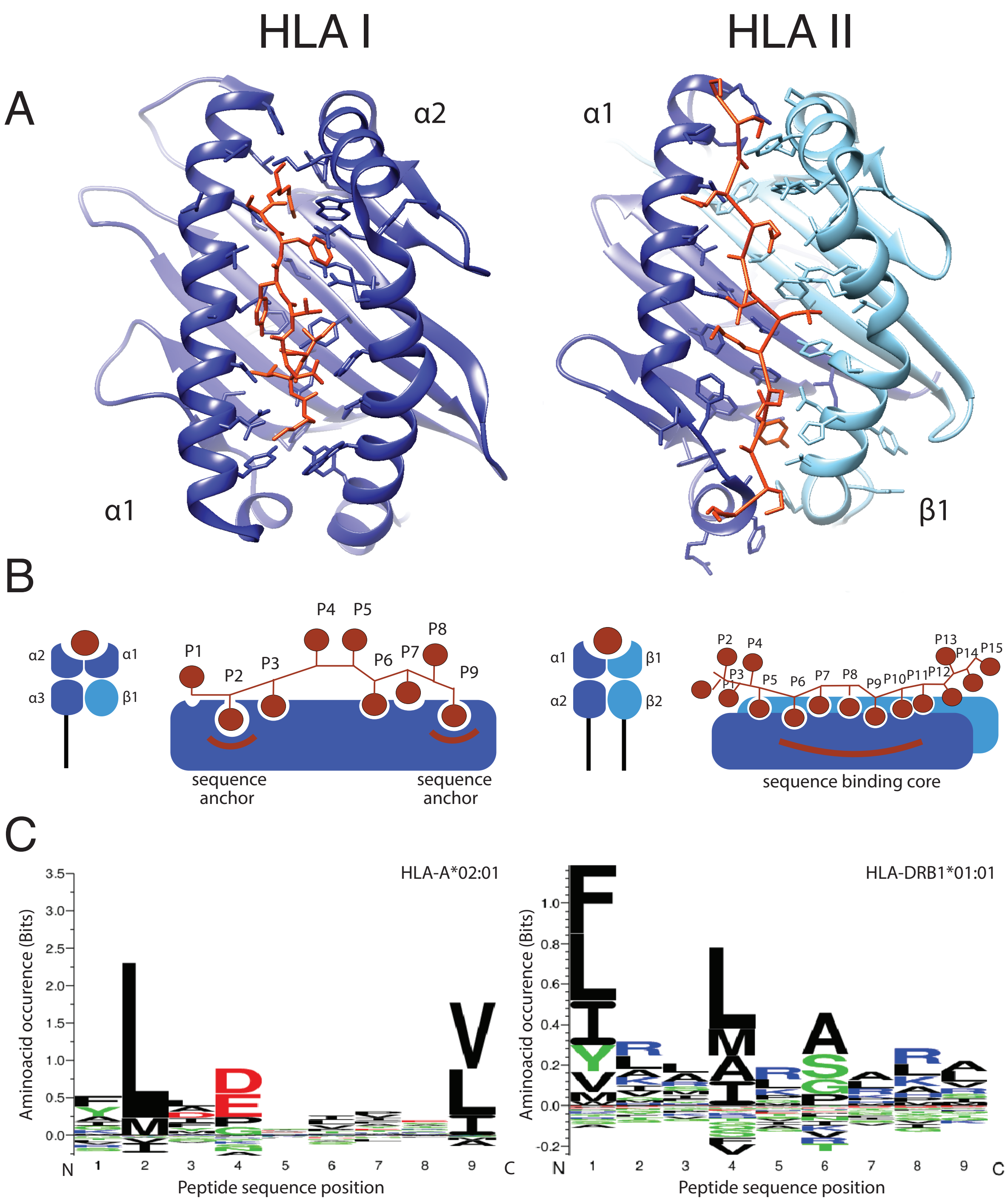


*Figure 2.3: A) Protein structures of the HLA class I (left) and II (right) peptide complexes (HLA-A*03:01, resolution: 2.6 A, PDB: 2XPG [34] and HLA-DR1, resolution: 2.1 A, PDB: 3L6F [35], visualized using UCSF Chimera [36]). B) Abstract cartoon of the tight peptide binding groove of the HLA class I receptor (left), composed of the domains α 1-3 and β 1 and the loose open peptide binding groove of the HLA class II receptor (right), composed of the domains α 1-2 and β 1-2. Peptide residue sequence positions are indicated as P_n. C) Sequence motifs of HLA class I peptides (left) bound to HLA-A*02:01 and HLA class II peptides (right) bound to HLA-DRB*01:01 (taken from the netMHCpan 4.0 [37] and netMHCIIpan 4.0 [38] motif viewer webserver.)*

their investigation remains a challenge [39, 40] (Figure 2.3 C).

Antigen presentation pathways

As mentioned above, peptides presented by the HLA-I and -II receptors stem from distinct molecular pathways in order to be presented on the cell surface. Oversimplified, the two pathways can be described as follows: The HLA-I antigen presentation pathway involves proteolytic cleavage of intracellular, cytosolic proteins by the proteasome and immunoproteasome followed by transmembrane transportation through the TAP transporter into the ER. In the ER, the HLA-I and β_2-microglobulin chain are bound to the peptides and folded into a protein complex by the help of chaperones. The resulting HLA –peptide complex is then transportet via vesicles to the cell surface. In contrast, HLA-II presented peptides are generated from proteins that are taken up from the extracellular environment via endo- or phagocytosis. Proteins are digested into peptides by lysosomal proteases of the endosomal pathway and the late endosome is fused with HLA-II loaded vesicles that were generated in the ER. While the HLA-II binding groove is blocked by the invariant chain protein li upon assembly in the ER, li is digested and replaced upon fusion with the late endosome allowing to reload the HLA-II molecule with other peptides. Finally, resulting HLA-II-peptide complexes, too are transported to the cell surface via vesicles [41].

2.1.4 *Vaccination and cancer immunotherapy*

Vaccine design

Traditionally, vaccines have been created against infectious diseases from injections of whole inactivated or live attenuated pathogens, which would stimulate the immune system to develop active and memory B-and T-cells against their antigens. As it may prove difficult to establish cell cultures of various pathogens in vitro, the production of the traditional whole organism vaccines can result in being a cumbersome, slow and impractical process. In addition, some organisms or inactivated whole viral particles might also lead to detrimental immune responses or could lead to unforeseen host responses that would rather be avoided. Therefore today many more approaches of creating vaccines have been established such as vector-, nucleic acid- or protein-based vaccines of which some reduce the vaccine cocktail to selected antigens [42]. Moreover, cancer immunotherapeutic approaches have been proposed, to stimulate a patients immune system, even retroperspectively after onset of the disease with vaccine cocktails containing tumor-associated antigens [11]. In particular, in this context, focusing the vaccine design on a limited set of antigens instead of whole cells or lysates is anticipated to be a therapeutically beneficial strategy [43].

Epitope-based vaccines

Peptides that are presented by the HLA receptor and are capable of eliciting an immune response via T-cell recognition (epitopes) are interesting targets for actively stimulating the immunesystem. Identifying specific epitopes of pathogens or diseased cells that are HLA matched to the majority of the population and combining them into a multi-peptide cocktail hence provides an efficient way for rational vaccine design. In contrast to vaccination with the

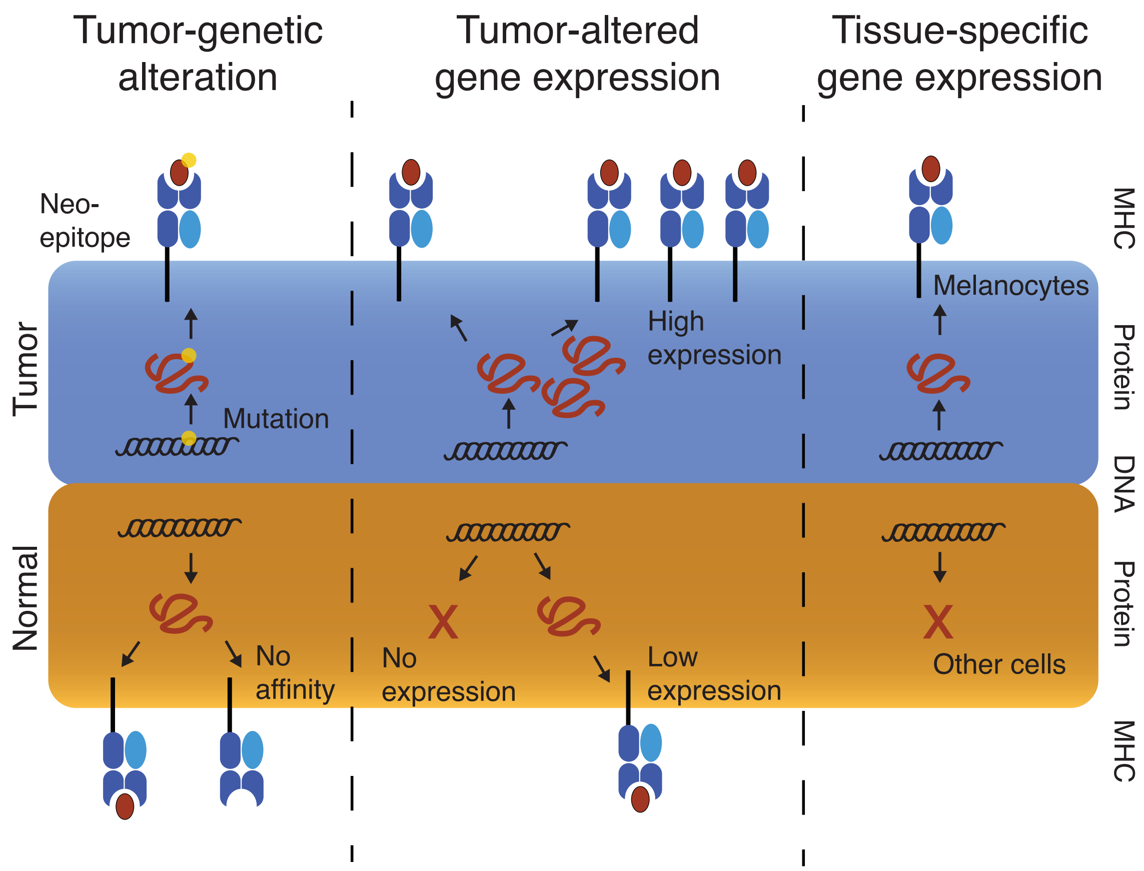


Figure 2.4: Three different types of cancer associated antigens: 1) genetic alterations such as somatic tumor mutations giving rise to the presentation of mutated neoepitopes 2) altered gene expression leading to the presentation of non-mutated antigens not or only to small amounts expressed in healthy cells or 3) tissue specific gene expression leading to the presentation of non-mutated proteins only present in specific tumor tissue cells such as for example melanocytes in Melanoma.

entire pathogen or cells, epitopes would also provide a way to enable easier and low priced large scale production of vaccines [43]. Thus, within the last decades epitope-based vaccines have gained interest in order to attempt treatment of a variety of diseases ranging from viral infections to various types of cancer. In fact, it has been proposed and taken to practice to generate off-the-shelf warehouse peptide vaccine cockatils for given HLA alleles and tumor entities [44, 45].

The origins of cancer associated antigens

Whereas viral peptides are likely to be immunogenic as they don't participate in the T-cell maturation process, it is far more difficult to identify immunogenic epitopes presented by various tumor cells. Cancer specific immune responses can be elicited by either tumor-specific antigens (TSAs) that are uniquely present in tumor cells or tumor-associated antigens (TAAs) that can be present in normal cells as well but are more abundant in tumor cells. Several scenarios have been proposed for the formation of TSAs and TAAs: (1) tumor-specific genetic alterations such as somatic mutations [46], frameshifts [47] or alternative splicing [48], (2) tumor-specific altered gene expression [49] and (3) tissue specific gene expression [9]. Genetic alterations can give rise to TSAs for example mutated peptides presented by the HLA complexes, so called "neoepitopes". Altered gene expression of tumors can result in the expression of non-mutated proteins that are not expressed or underrepresented in normal cells and the corresponding presentation of their epitopes. Ultimately tissue specific expression such as for example in melanocytes might give rise to non-mutated epitopes that are almost exclusively found in melanoma [9] (Figure 2.4).

Challenges for the translation into clinical practice

Despite significant advances in the development of immunotherapies by the use of epitope-based therapeutic vaccines, in particular the discovery of cancer-associated epitopes remains challenging. Unfortunately, until today cancer vaccines have to a large part failed to prove efficacy [49, 50]. However, it has been argued that the potency of cancer vaccines could increase strongly in combination with more effective adjuvants, prioritized selection of epitopes for vaccine cocktails, integration with multi-omics data or in combination with adoptive T-cell transfer [51, 52]. In addition, the discovery of checkpoint inhibitors that can be used to stimulate activity of cytotoxic T-cells could also be used in combination therapies to increase effectivity of cancer therapies. However, boosting efficacy of immunotherapies has to be approached with care as this can eventually cause severe autoimmune reactions as adverse side effect [53, 54]. Yet, given that cancer causes one of the highest death tolls world wide [8], the successful establishment of protein therapeutic vaccines against cervical cancer such as Gardasil or Cervavix [55, 56] and promising results in personalized cancer neoepitope vaccines [57], epitope-based cancer vaccines still hold great potential [58].

2.2 *Mass spectrometry of HLA-presented peptides: Immunopeptidomics*

This section introduces concepts of the mass spectrometry (MS) technology using the example of HLA-presented peptides (immunopeptidomics). A broad overview of the sample preparation and purification using liquid chromatography prior to MS is given, as well as the description of the key components of MS instruments. Finally, the operation modes of measurement and the recorded signals that correspond to peptides are explained in detail.

2.2.1 *Sample preparation*

In order to prepare tissue samples for an immunopeptidomics analysis using MS, it is necessary to purify the HLA-bound peptides in a complex procedure of multiple steps: (1) The tissue samples need to be homogenized using physical forces using for example a scalpel, potter and sonicator. At all times samples are treated in a cold room and kept in a lysis buffer solution, containing detergent to lyse the cellular membrane and a protease inhibitor that prevents degradation of peptides and proteins. (2) The homogenized tissue solution needs to be centrifuged and filtrated under sterile conditions separating the soluble protein solution containing the HLA-peptide complexes from other insoluble parts. (3) The HLA-peptide complexes are immunoaffinity purified using HLA-I or HLA-II pan-specific antibody-coated beads. (4) Subsequently, mild acid is used to dissociate the HLA-peptide complexes and elute them off the antibody-coated beads. (5) Using ultrafiltration peptides can then be separated of the HLA complexes. (6) Finally, the received peptide solution is then purified a few more times by filtering it with a hydrophobic column and the resulting solution of peptides can be then directly injected into the LC-MS/MS system [14] (Figure 2.5).

2.2.2 *Liquid chromatography*

MS-based analytical chemistry is most commonly coupled to liquid chromatography (LC) prior to injection of analytes into the instrument. The application of chromatography reduces the complexity of the sample for the MS measurement, by separting peptides according to their polar/hydrophobic properties along a chromatographic column. Various types of solvents (mobile phase) and column materials (stationary phase) have been applied for peptide chromatography, however one of the most commonly applied methods is termed "reversed phase" and will be explained here further. It involves a polar solvent (ethanol/acetonitrile and water) and a hydrophobic C18 column. The C18 hydrocarbon chains are attached to a solid silica base and create strong hydrophobic interactions with molecules that are transported across its surface through the chromatography procedure. The polar solvent on the other hand solubilizes the peptides through polar interactions and transports the sample analyte through the column, accelerated by a pump machinery. In addition, a gradient can be dynamically applied throughout the chromatography increasing the ratio of the solvent-to-water mixture to decrease its polarity and solubilize rather hydrophobic peptides as well towards the end of the procedure [59] (Figure 2.6).

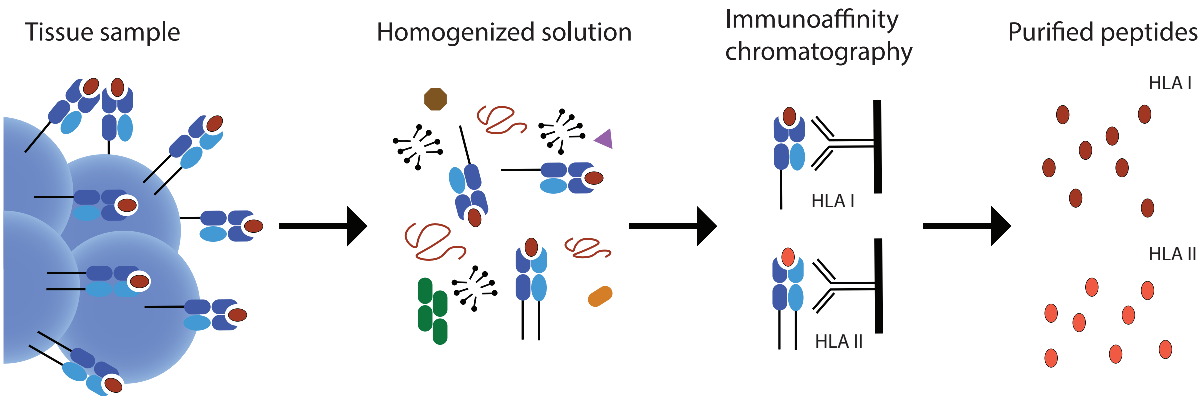


Figure 2.5: Experimental workflow to extract HLA-I and -II bound peptides: Biological samples such as from cells or tissues are homogenized, followed by highly specific immunoaffinity chromatography and further purification to yield only the presented peptide ligands without the receptor complex.

2.2.3 *Electrospray ionisation*

In order to enter the low-pressure gas phase inside the MS instrument, analytes need to undergo a phase-transition from the liquid chromatography solubilized mobile phase. In addition peptides have to be ionized, to be accelerated by an electromagnetic field steering analyte molecules through the instrument. Many methods have been established to achieve this, but the most common and sensitive method earning their creators the Nobel prize, is refered to as "soft" electrospray ionization (ESI) [60]. In this method the solute is pumped through a capillary that is subjected to a strong electric field. Cone shaped drops at the tip of the capillary (Taylor cone) diffuse into an aerosol of highly positively charged droplets that is accelerated by the electric field pointing into the MS instrument. Throughout their trajectory, the volume of the droplets continuously shrinks until reaching the Rayleigh limit, which causes complete dissociation (Coulomb explosion) of the droplet into solute and analyte ions because of the extreme charge repulsion. The resulting analyte ions are non-fragmented and carry the protonation charge, which allows to assess their mass-to-charge ratio inside the MS instrument.

2.2.4 *Mass spectrometry*

MS instruments are composed of four basic units: (1) ion source, (2) mass filter, (3) mass analyzer and (4) detector. Different systems of these four components exists and here we focus on the setup of the Thermo Lumos Orbitrap Fusion mass spectrometer, applied in all studies addressed by this doctoral thesis. (Figure 2.7)

Ion source

The ion source comprises the components that contribute to the ionization of the analyte at the entry point of the MS. In the case of ESI, analytes exiting the injection capillary enter

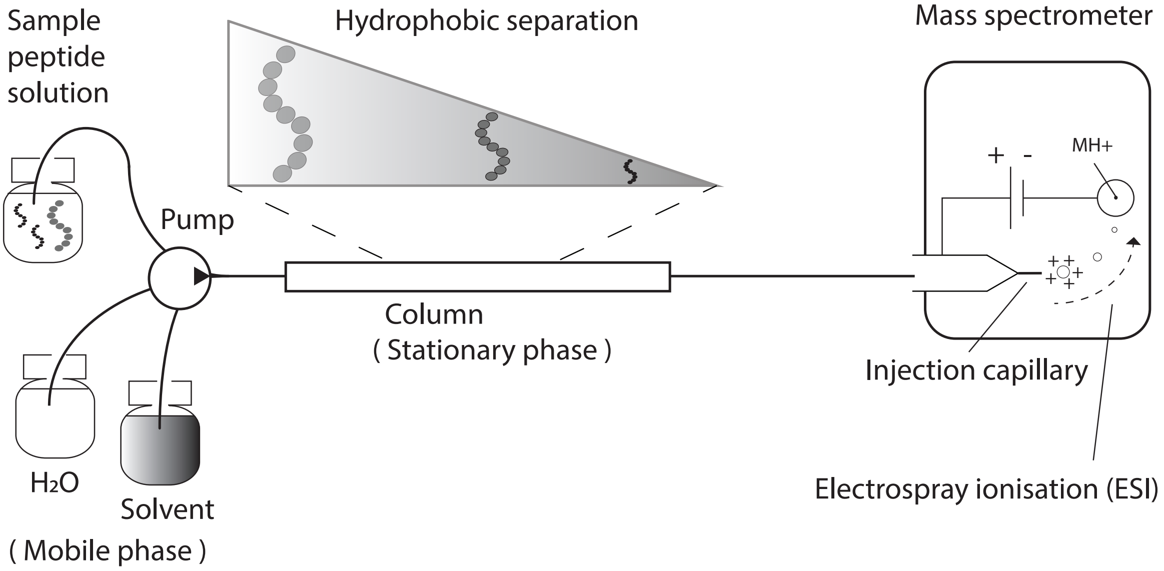


Figure 2.6: Coupled liquid chromatography: Analytes are separated by hydrophobic interactions with the stationary phase. Injection into the MS instrument occurs via electrospray ionisation (ESI) through a capillary subjected to a strong electric field.

a small inlet hole, followed by heating elements and a low pressure cell inside the first MS chamber. Thereby the droplets leaving the injection capillary can completely ionize and an ion beam will be created. In addition, commonly an electrodynamic ion funnel counter-acts the Joule expansion and focuses the beam. Finally ions bypass an active beam guide, to filter out neutrally charged particles that have entered through the inlet to ensure that only charged ions enter the following mass filter [61, 62].

Quadrupole mass filter

A quadrupole mass filter consists of four parallel metal rods creating an oscillating electrical field due to a mix of direct and alternative current. Through variation of the oscillation frequence, only ions of a certain mass-to-charge range are capable of passing the quadrupole with a stable trajectory and are filtered out from the rest. Hence, after the mass filter only ions of a certain mass-to-charge range continue its path through the MS instrument to the mass analyzers. Commonly, a C-trap follows the quadrupole mass filter in order to bring packages of ions into the same phase before passing them on to the orbitrap mass analyzer [63].

Orbitrap mass analyzer

In a first stage (MS1), ions are trapped in a multipole and sent to the orbitrap mass analyzer. The ions measured in the orbitrap stem from the intact, non-dissociated peptide analytes injected into the MS instrument (precursor ions). The ions are trapped in a circulating trajectory "orbit" around an electrical field surrounding the core rod. By measuring the oscillation frequency spectrum and deconvoluting the signal into spectral components using Fourier transformation, the ion mass-to-charge ratios trapped in the field can be determined in high resolution (MS1 precursor spectrum). The amplitude of each frequency component then corresponds to the intensity of a given peptide analyte which is frequently used to determine its

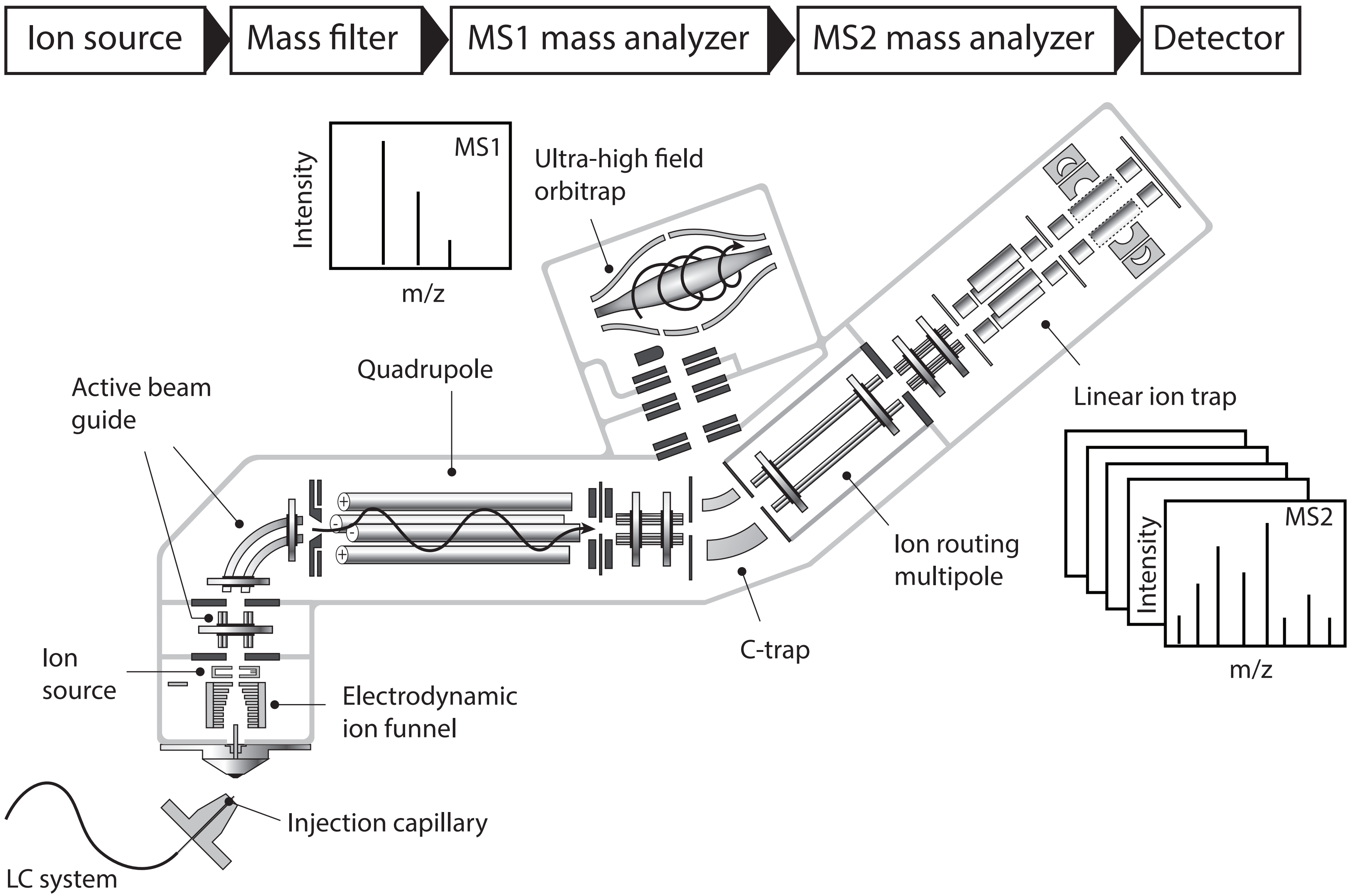


Figure 2.7: Technical setup of the Thermo Lumos Fusion mass spectrometer: Ions are guided from the LC-system to the mass analyzer and detector resulting in the acquisition of precursor MS1 and fragment MS2 spectra. (Figure with permission adapted from "https://planetorbitrap.com/orbitrap-fusion-lumos")

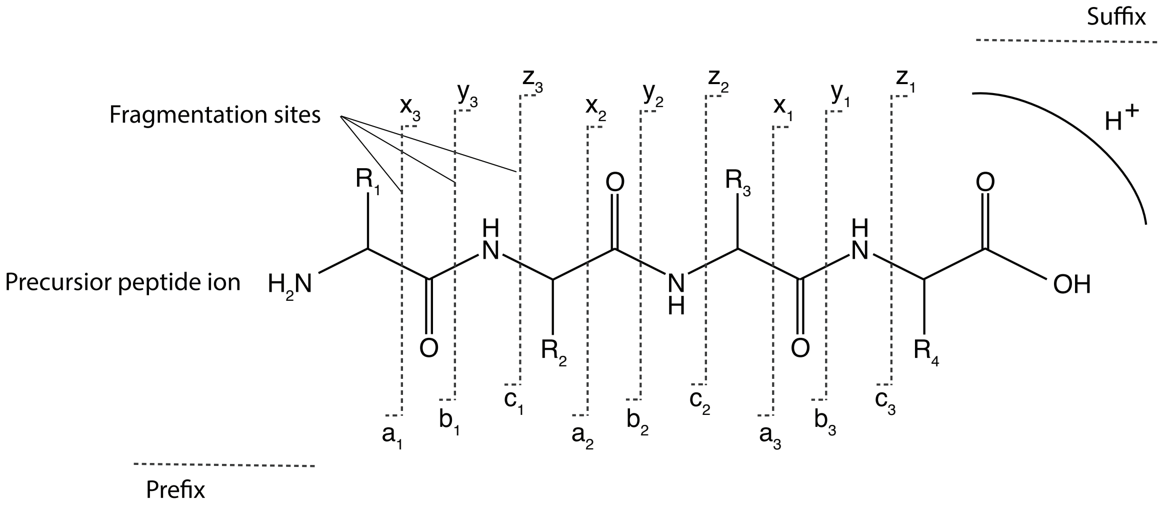


Figure 2.8: The characteristic a,b,c and x,y,z fragmentation sites of a peptide precursor ion [66].

abundance in the injected sample at a later stage [64].

Collision-induced fragmentation

Oftentimes the determined MS1 precursor mass is highly ambiguous, as multiple different peptide species could result in isobaric observed masses within an instruments detection tolerance. Therefore, to unambigously identify the peptide sequence of an analyte entirely, the ions are physically fragmented into smaller pieces. Hence, in a second stage (MS2) all ions kept in the multipole ion route are subjected to a fragmentation method such as collision-induced dissociation (CID) or high-energy collision dissociation (HCD) [65].

Throughout the fragmentation procedure, the precursor ions are bombarded with neutral noble gas atoms (He, Ne, Ar) that transmit their kinetic energy (E_{Kin}) and cause the precursor ions to undergo intramolecular dissociation into fragments. Among the various fragmentation products, in particular characteristic prefix and suffix ions of peptides are created. The most abundant fragments observed after this type of dissociation are commonly b and y fragment ions that split the peptide at its amide bond connecting two distinct amino acids. (Figure 2.8) [66] :

$$[MH]^{+}_{Precursor} + E_{Kin} \rightarrow [MH]^{+*}_{Precursor} \rightarrow [MH]^{+}_{Fragment(a|b|c|x|y|z)} + [M]_{Fragment(a|b|c|x|y|z)} \qquad (2.1)$$

Linear ion trap mass analyzer

The mass-to-charge ratio of the resulting peptide fragment ions can be measured using a linear ion trap. Similar to a quadrupole a linear ion trap is composed of four metal rods that create an oscillating electric field of high frequency. Consequently, it is able to trap ions in of selected mass-to-charge ranges its inner volume by varying the oscillation frequency and thereby selectively passing them on to a detector (MS2 fragment spectrum).

Detector

While the orbitrap mass analyzer is able to detect the mass-to-charge ratio and their intensity simultaneously, the linear ion trap needs an additional detector to analyze the intensity of a given ion. The ions are thus forwarded to additional elements such as an electron multiplier or faraday cup that register the incidence of a charged particle hitting the detector surface. The resulting signal is then massively amplified turning the weak detection into a strong current that is recorded [63].

MS signal components of peptide ions

The signals of a peptide that are recorded using a MS instrument span in the RT and mass-to-charge dimension. As peptides are composed of atoms and are subjected to natural isotope abundances, the same peptide species is usually represented by multiple isotope variants that have slightly different masses according to isotope mass shifts. The average abundances of the different isotope species can be accurately estimated by the so called "Averagine model" [67]. Hence, the entire MS1 signal recorded from a single peptide is not mono-isotopic and instead is regarded as the collective group of isotope mass peaks and their corresponding RT elution profiles. The terminology for this group of signals varies in literature and scientific discussions but is commonly named MS1 feature, peak group or mass trace. (Figure 2.9 A).

MS2-level fragment mass signals for a given precursor ion are commonly termed "transitions". Depending on the acquisition mode used during the measurement, transitions of the same precursor are acquired few times to only once or highly redundant to multiple times spanning the RT range. Thus the resulting MS2 signal of a peptide differs among these approaches and is the group of mass peaks corresponding to all analyte fragments in one specifically triggered MS2 spectra (Figure 2.9 B) or all fragment mass peaks in combination with their RT elution profiles (Figure 2.9 C).

Due to fluctuations of the background signal, chemical impurifications from the mobile phase or buffer and contaminants MS signals are also influenced by noise signals that do not stem from peptide analytes [68]. Extracting the RT elution profile of a specific mass peak group only in contrast to all other peaks is referred to as an extracted ion chromatogram (XIC) and can be advantageous to focus on the signals of interest. XICs of fragment ions are only available when measuring in targeted acquisition modes such as reaction-monitoring or data-independent acquisition. Mass peaks belonging to the same peptide ion in an XIC should follow the same RT elution profile, which is a property used by targeted approaches to accurately identify peak groups.

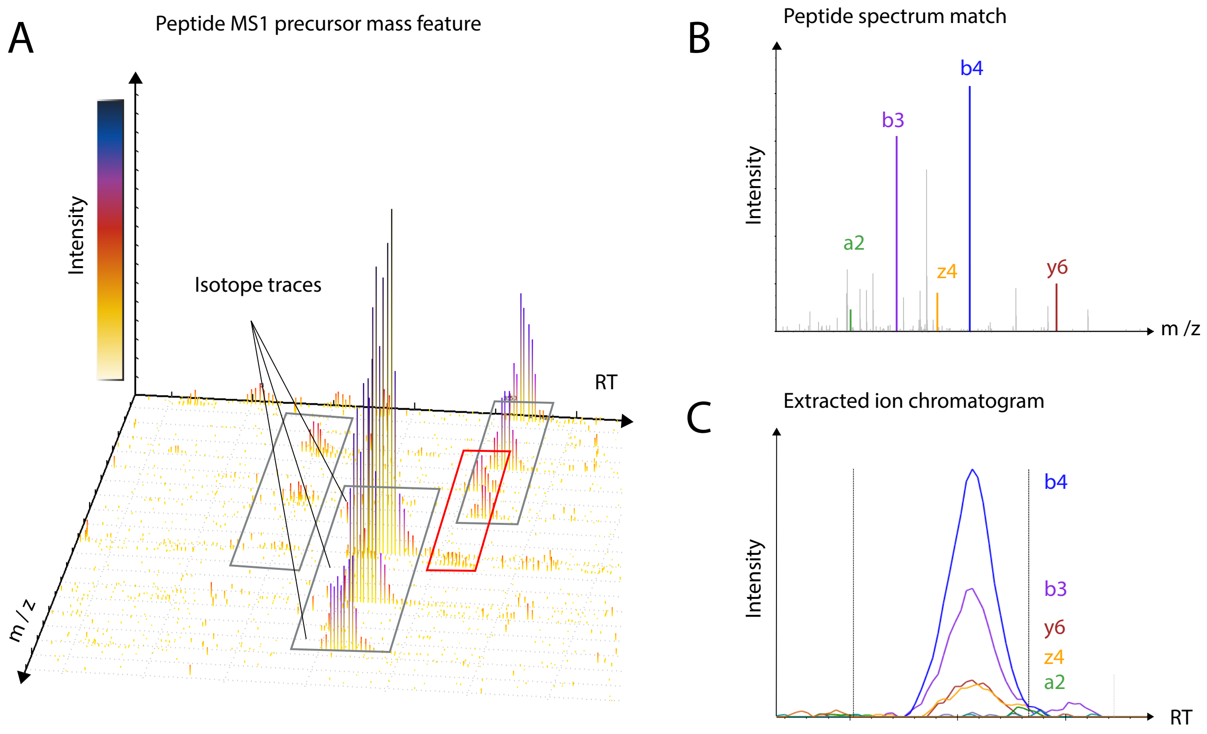


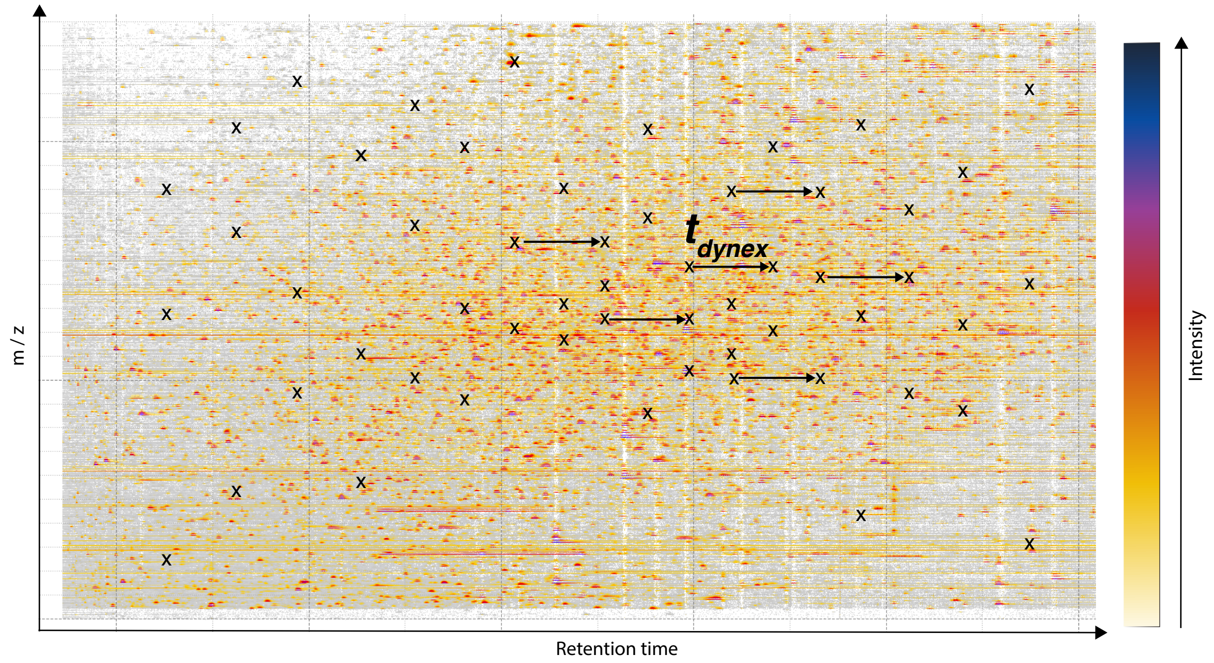
Figure 2.9: The components of signals measured in a MS instrument originating from a peptide ion: A) MS1 precursor mass spectrum of peptides and their corresponding isotopic peak groups. B) MS2 spectrum of a fragmented peptide encountered when measuring in DDA mode, matching fragment ions are highlighted. C) Extracted ion chromatogram of selected MS2 peptide fragment transitions as retrieved when measuring in DIA or targeted mode.

2.2.5 *Data acquisition modes*

The simultanious acquisition of both, MS1 level precursor and MS2 level fragment spectra after peptide dissociation requires complex timely coordination of the two levels of spectra acquisition. In most acquisition techniques the MS instrument triggers an MS1 spectrum at regular time intervals, measuring the ions that are sequentially eluting from the LC system. However, regarding the MS2 level, the MS instrument needs to select at a given timepoint, which among the multitude of co-occurring precursor ions are sent for fragmentation. A number of different acquisition techniques have been established that can be set during an MS measurement to influence the way MS1 level and MS2 level spectra are concurrently acquired [69].

Data-dependent acquisition mode (DDA)

In practice the most common way of operating an MS instrument in discovery proteomics has been achieved using the DDA mode. For each acquired MS1 spectrum the top n most intense precursor ions are selected for fragmentation. Moreover, it is possible to specify a "dynamic exclusion time", that will prevent the MS instrument from acquiring fragment spectra of the same precursor mass repeatedly if it is very abundant. This is achieved by briefly appending precursor masses that were already triggered for MS2 fragmentation to a dynamic exclusion list (L_{dynex}) which is subsequently excluded from fragmentation for the specified time (t_{dynex}). The DDA mode of spectral acquisition results in high quality spectra for most peptides in the sample. Ideally, each precursor analyte recorded on MS1 level is fragmented once with a cor-



Algorithm 1 Data-dependent acquisition mode

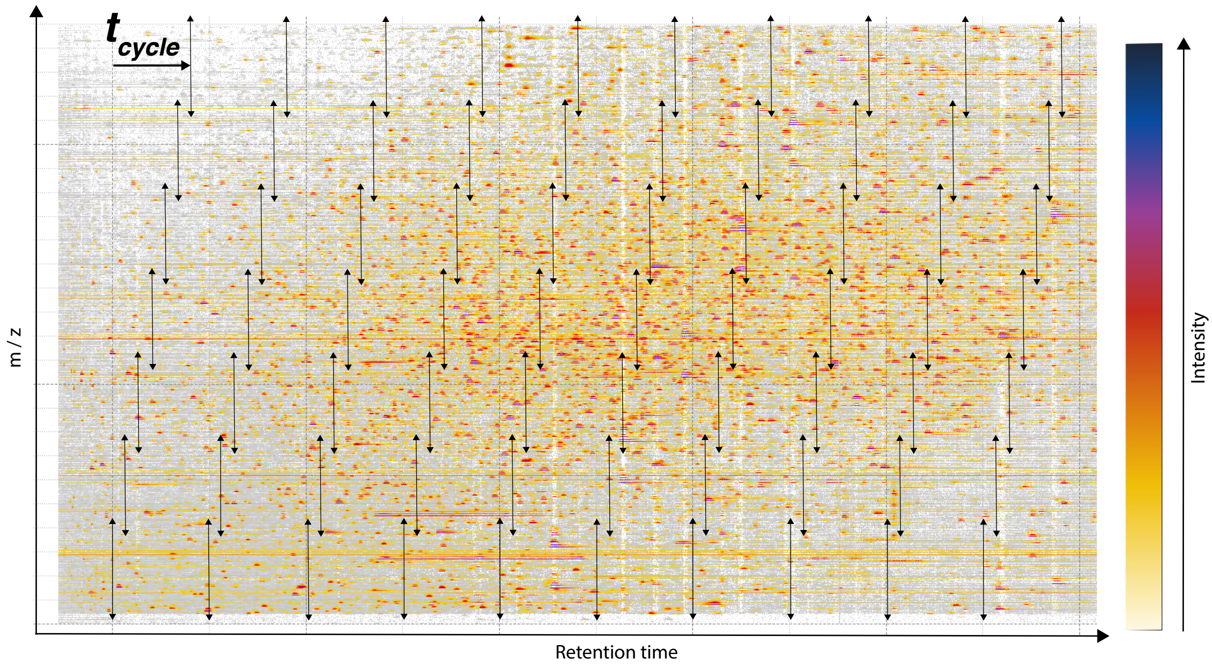
```
while $t_x \leq t_{max}$ do
    $MS_1 \leftarrow$ acquire $MS_1$ $Spectrum$ at $t_x$
    $ions_{sorted} \leftarrow$ sort $ions$ in $MS_1$ by intensity
    $ions_{included} \leftarrow$ select $ions_{sorted}$ if $masses$ not in $L_{dynex}$
    $top_n \leftarrow$ select n most intense from $ions_{included}$
    for $i$ in $top_n$ do
        $MS_2 \leftarrow$ acquire $MS_2$ $Spectrum$ of $i$
        add $mass_i$ to $L_{dynex}$ until $t_x + t_{dynex}$
    end for
end while
```

Figure 2.10: Schematic for the data-dependent acquisition mode (DDA): MS2 fragment spectra (black crosses) are triggered dynamically based on the most abundant intensities of MS1 precursors (top 3) in the LC-MS peak map (illustration generated from in-house immunopeptidomics MS measurement of lung tissue). Below the algorithm for the DDA approach that is carried out by the MS instrument is sketched in pseudo code.

responding MS2 spectra. As a result from this approach, fragmentation events are randomly distributed over the highly abundant peaks in the LC-MS peak map. However, sections of the LC-MS peak map that contain less intense ions are not covered by this approach. (Figure 2.10)

Data-independent acquisition mode (DIA)

When operating the MS instrument in DIA mode, there is no dependence of the MS2 fragmentation event on the intensity of the MS1 precursor ion. In contrast, the LC-MS peak map is divided into a regular grid of windows and at each timepoint all windows are sequentially triggered for fragmentation. The cycle time (t_{cycle}) refers to the time how long one cycle of fragmentation of all windows lasts. Hence, it is possible to trigger fragmentation events in a much more reproducible way, covering nearly all including low abundant peaks in the LC-MS map. However, due to the cofragmentation of multiple precursor ions occuring in the



Algorithm 2 Data-independent acquisition mode

```
while $t_x \leq t_{max}$ do
    $MS_1 \leftarrow$ acquire $MS_1$ $Spectrum$ at $t_x$
    for $window$ in $grid$ do
        $ions_{window} \leftarrow$ select $ions$ in $MS_1$ by $window$
        $MS_2 \leftarrow$ acquire $MS_2$ $Spectrum$ of $ions_{window}$
    end for
end while
```

Figure 2.11: Schematic for the data-independent acquisition mode (DIA): MS2 Fragment spectra are triggered in a reproducible grid of windows (black arrows) covering the entire LC-MS peak map (illustration generated from in-house immunopeptidomics MS measurement of lung tissue) Below the algorithm for the DIA approach that is carried out by the MS instrument is sketched in pseudo code.

same window, the resulting spectra are more complex, highly redundant and require deconvolution to be interpreted by the human eye. Nevertheless, advanced computational methods have been developed to confidently and automatically analyze DIA MS data [70] (Figure 2.11).

2.3 *Computational approaches in immunology & mass spectrometry*

This section summarizes computational approaches that are applied for immunology and general MS data interpretation. An overview over HLA binding affinity prediction algorithms is given and the steps of peptide identification from MS raw data are explained. Finally, the open-source software toolbox OpenMS to analyse MS data is introduced and concepts of computational workflow systems, continuous integration and containerization for good practice in programming and data analysis are introduced.

2.3.1 *HLA binding affinity prediction*

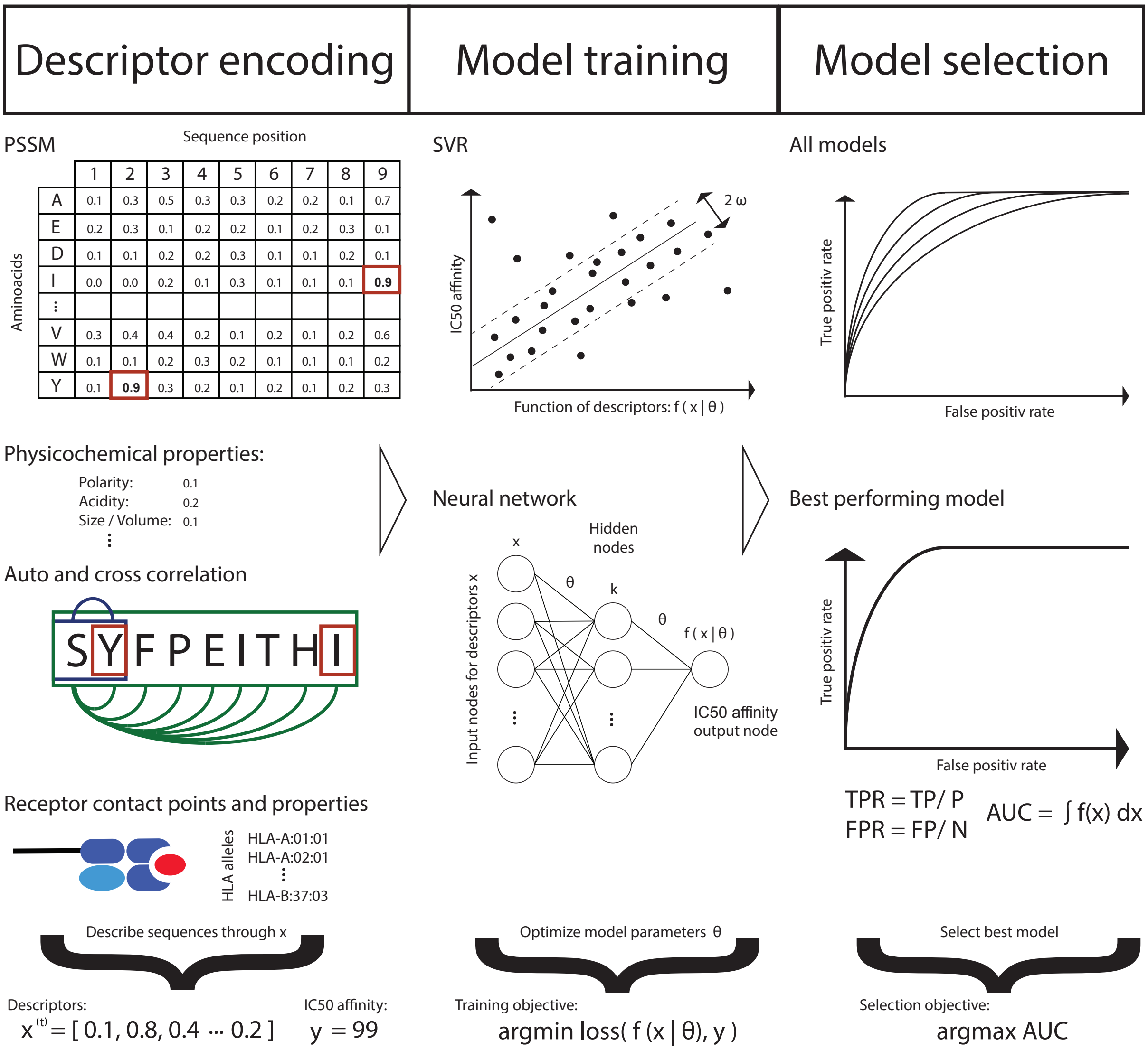
A milestone achievement in computational immunology has been to predict from sequence, which partial peptide sequences of a given protein or antigen are likely to be presented by the HLA receptor of a certain allotype. Therefore a great variety of computational approaches have been developed to achieve this goal [71].

Accordingly, the discovery and gathering of knowledge on peptide sequence motifs corresponding to the various HLA types has been used to train predictive models to recognize sequence patterns and transfer them to any unknown input sequence. Among these, existing methods vary concerning the data they have been trained with in the way peptide sequences are encoded by descriptors, the mathematical, predictive models chosen and how the prediction error is minimized. (Figure 2.12) Attempts to benchmark the multitude of existing methods for HLA class I predictions, reveal superiority of particular approaches for given alleles and test data, but indicate rather similarly good overall performance levels [72].

Early methods are based on position specific scoring matrices (PSSM) [32, 73] that derive a scoring matrix for a given residue in a given position of the sequence. These methods were improved using a large branch of methods that have been based on modern machine learning approaches such as support vector machines [74], artificial neural networks [75, 76] and most recently deep learning [77].

Exceptional among these are pan-allele prediction methods that do not only encode the peptide sequence to feed to the model but also the HLA receptor sequence [78]. In this way, peptide affinities can be successfully predicted for HLA alleles, which have only been scarcely studied and little is known about the actual peptide sequences bound to it. While some methods have been trained with binding affinity data to derive a regression model, other methods have focused on the prediction of presentation probability based on the separation of random sequences from naturally HLA-presented peptides identified using MS. Most recently, these two approaches have been combined as well [37].

While HLA class I peptide binding affinity prediction has proven to be quite accurate, many HLA class II sequence motifs have yet to be identified. This is due to the greater combinatorial receptor variety as well as to the sequence binding core that is more difficult to identify in advance and feed into a predictive model. Ultimately, recent developments on sequence



	1	2	3	4	5	6	7	8	9
A	0.1	0.3	0.5	0.3	0.3	0.2	0.2	0.1	0.7
E	0.2	0.3	0.1	0.2	0.2	0.1	0.2	0.3	0.1
D	0.1	0.1	0.2	0.2	0.3	0.1	0.1	0.2	0.1
I	0.0	0.0	0.2	0.1	0.3	0.1	0.1	0.1	**0.9**
⋮									
V	0.3	0.4	0.4	0.2	0.1	0.2	0.1	0.2	0.6
W	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0.1	0.2
Y	0.1	**0.9**	0.3	0.2	0.1	0.2	0.1	0.2	0.3

Figure 2.12: Machine learning based peptide HLA affinity predictions: Most approaches encompass and differ in the stages of (1) descriptor encoding, (2) model training and (3) model selection. A peptide can be numerically encoded in various ways for example via position specific scoring matrices (PSSM), physicochemical properties and auto- or cross-correlations of these values. Its binding affinity to a given HLA allele (some methods encode the receptor sequence in addition) serves as label for the subsequent training step. During the training procedure a chosen algorithm such as a support vector regression (SVR) or neuronal network (NN) optimizes its internal parameters with the objective to minimize the prediction error for the given label values of all training instances. Finally in a model selection step different trained models can be compared based on a general accuracy measure evaluated for each model. A common measure is for example the area under the receiver operator curve (AUC) that takes into consideration the number of false positive (FP), true positive (TP), false negative (FN) and true negative (TN) events when evaluating the predictions on a test data set.

deconvolution methods and complex representations of all possible sequence cores have been derived in order to derive more accurate prediction models for HLA class II [40, 79].

2.3.2 *Mass spectrometry database search*

Peptide identification

High-throughput dentification of peptide sequences from raw MS spectra is a complex computational task. The multitude of search strategies can be divided into database search and *de novo* peptide identification. Database search relies on a database of protein sequences, that are *in silico* digested into peptides and matched against mass spectra, whereas *de novo* identification attempts to reconstruct the peptide sequence only from the contained fragment masses in the spectrum. For database search of HLA peptides, a protein database human proteome is unspecifically cleaved into all possible peptides of a given length range. For a given peptide spectrum match, candidates are first preselected according to their precursor mass. Next, the resulting candidates are in silico fragmented into possible fragment ions (commonly only b- and y-ions are considered) [66]. The masses of the theoretical fragment ions are then used to build a theoretical fragment spectrum assigning a simplified binary intensity values of 1 or 0 if a corresponding fragment mass is present or not. This simplification step is taken because of the difficulty to accurately predict continuous theoretical fragment intensities, which has only recently been achieved [80, 81]. Ultimately, theoretical fragment mass spectra of all peptide candidates are compared with measured MS2 spectra and ranked according to a scoring scheme. The best ranking peptide spectrum match (PSM) is then considered the most likely candidate for a given spectrum. (Figure 2.13)

Peptide spectrum match scoring

The entire process of matching a peptide to a given spectrum for all measured spectra against the entire database of proteins is computationally demanding and depends on the size of the used database. Over the last decades multiple "search engines" that carry out the process have been developed. They differ by the scoring scheme and preselection of peptide candidates. One very common, simple scoring function for comparing two vectors of the theoretical spectrum x and experimental spectrum y that is employed by the sequest [82] or Comet search engine [83] is the cross-correlation function:

$$C_{XY}(\tau) = \int_{-\infty}^{\infty} x(t)y(t+\tau)dt \tag{2.2}$$

The similarity score (*xcorr*) for discrete discrete signals such as encountered in mass spectrometry is then calculated as:

$$xcorr = x_0 y_0 - \left(\frac{\sum_{\tau=-75}^{\tau=75} x_0 y_\tau}{151} \right) \tag{2.3}$$

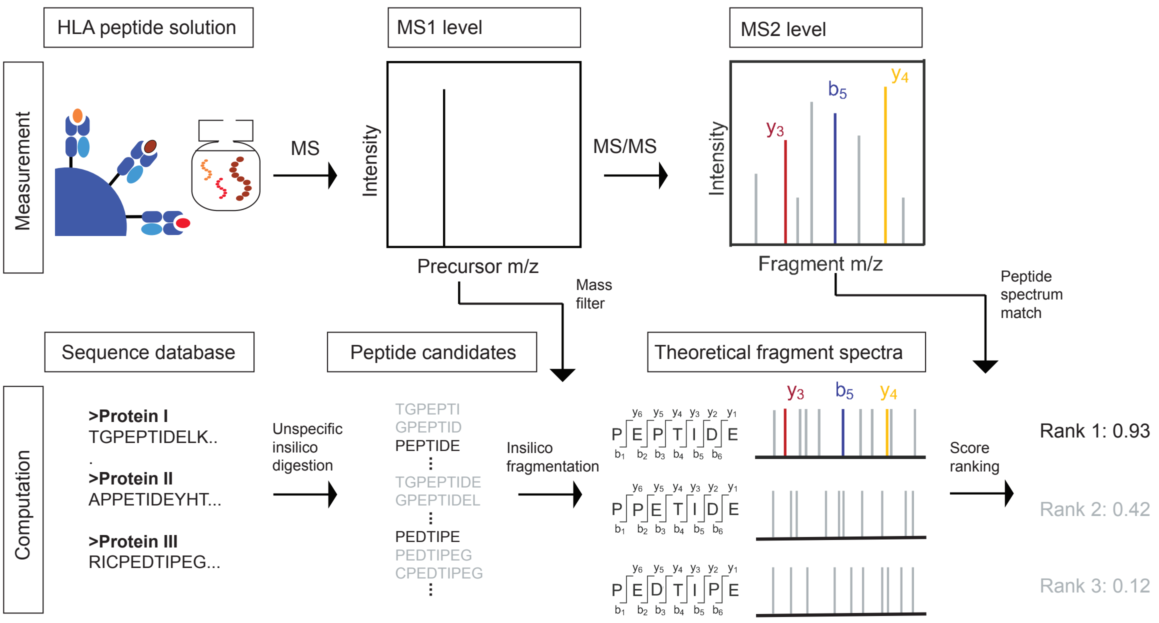


Figure 2.13: MS database search of MHC peptides: A database of protein sequences is in silico cleaved unspecifically into all possible candidate peptide sequences of a given length range. Among those only candidates are considered that match to the experimental MS1 spectrum within the precursor mass tolerance. Next the peptide sequence candidates are in silico fragmented into b and y ions to produce a theoretical spectrum for each candidate. Finally, the best peptide spectrum match (PSM) for a given fragment MS2 spectrum is determined by a specific scoring scheme.

Here, *xcorr* represents the cross product of the discrete input signals of the candidate theoretical spectrum with its experimental acquired spectrum that is translated by windows of -75 – 75 Da relative to its origin (the weight of the lightest amino acid glycin equals 75 Da). It implies that the score of the cross-correlation at $\tau = 0$ is substracted by the average cross correlation of random matches when shifting by a non-proteinogenic mass. Rewriting equation 2.3 as:

$$xcorr = x_0 \left(y_0 - \frac{\sum_{-75}^{75} y_\tau}{151} \right) \tag{2.4}$$

allows to rapidly compute spectral comparisons with preprocessed input spectra averaged over the translation window. Finally, in addition sparse vector data structures to efficiently represent spectra are used to reduce memory use [83].

False discovery rate estimation

Taking the top scoring rank of all peptide candidate to a given spectrum is not enough. To correctly identify most sequences in high-throughput proteomics, it is important to define what is a good scoring match or not. This is due to the fact that many spectra are of insufficient quality to be correctly identified by any match and some spectra might stem from contaminants in the instrument or peptides not included in the database that should not be identified by a false match. To separate good PSMs from insufficient quality ones, it is therefore necessary to derive a score threshold that determines the rate of correct peptide identifications.

In database search this is achieved by matching in addition to the correct "target" peptide sequences the same amount of random "decoy" peptide sequences that are not supposed to exist in the searched sample. Decoy sequences should be as similar in their sequence properties as possible to target sequences in order to avoid any bias. In practice this is for example achieved by reversing or shuffling of all target sequences.

By evaluating the PSM score distributions of target and decoy sequences, it is then possible to compute a score threshold, that if used for filtering, will result in PSMs that contain only a certain percentage of "decoy" PSMs (Figure 2.14 A). The application of the score threshold is commonly used in proteomics studies to approximate the expected false discovery rate: [84, 85]

$$E\{FDR\} = \frac{b}{a+b} = \frac{|\{d_i > t; 1, ..., m_d\}|}{|\{f_i > t; 1, ..., m_f\}|} \tag{2.5}$$

Here d_i represent the decoy PSM scores and f_i all PSM scores and their respective total numbers (m_d and m_f) passing the FDR threshold t.

The FDR can then be calculated on various levels of abstraction such as on the level of PSMs, Peptides or Proteins [86] (Figure 2.14 B). However, the search engine score is not the only criterion to differentiate good "target" from random "decoy" PSMs. In fact many other factors give hints about the quality of a given PSM such as for example the retention time, precursor mass deviation, or score difference to the second best ranking peptide candidate (dCn). Therefore a number other approaches have been developed that achieve a better discrimintation by multivariate "target-decoy" separation [87]. For instance, the Percolator algorithm [88] is an iterative machine learning strategy to separate targets and decoys to attempt a more accurate approximation of the FDR. This often results in more or different peptide identifications than with an univariate FDR estimation and is therefore at this time applied in most proteomic workflows.

Finally, there are two more important considerations when estimating the FDR: (1) It needs to be considered that the FDR threshold is influenced by the size of the database search space. This is because finding good separations between targets and decoys gets more difficult due to a greater variance in their score distributions. Hence, less targets will be found at high confidence in large search spaces. (2) When considering the estimation of a FDR on large data sets, error accumulation can lead to drastic underestimation when comparing locally and globally evaluated computed FDR results (Figure 2.14 B) [89]. This is because the target identifications for example of a set of human proteins are likely rediscovered in various measurements but the randomly observed decoy identifications do not reoccur. Thus, the discovery of decoy sequence identifications will accumulate in a large set of measurements and lead to an imbalance when assessing the FDR [90]. Especially on the protein or peptide level of abstraction errors accumulate quickly over many measurements repeatedly adding different false discovered proteins or peptides with every sample.

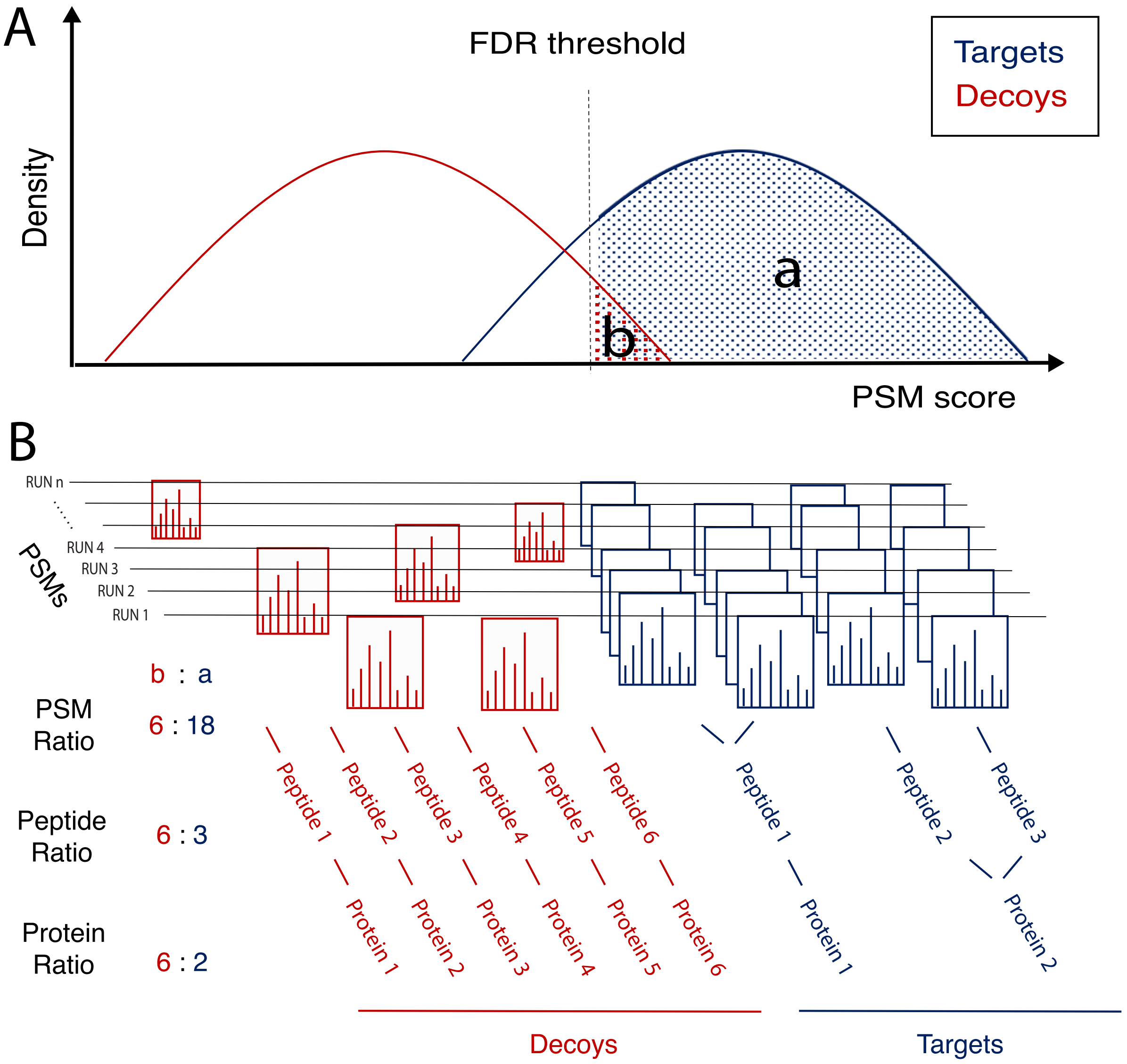


Figure 2.14: A) The score distributions of target and random decoy PSMs are illustrated, among which targets on average achieve better scores than decoy peptide identifications. Thus, it is possible to set a score threshold to accept only a certain percentage of decoy discoveries as an estimate of the FDR from all peptide identifications. B) Effects on target/decoy identification ratios for FDR estimation based on the level of PSMs, peptides and proteins and the number of MS runs. In contrast to decoy PSMs, targets present in a sample are typically discovered by several PSMs from multiple MS runs and can be often associated to a shared set of peptides and proteins. This illustrates the problematic of an increasing imbalance of decoy and target identifications for FDR estimation, due to error accumulation with higher numbers of MS measurements.

One approach to estimate the FDR of very large data sets more accurately has been achieved by modelling the probability of false positive identifications as a hypergeometric distribution (P_{hg}) [91] :

$$
\begin{aligned}
E\{FDR\} &= \frac{E[h_{fp}|h_t, h_{cf}, \theta_{exp}(N)]}{h_t} \\
&= \frac{1}{h_t}\sum_{h_{fp}} h_{fp} \cdot P_{hg}(h_{fp}|h_{tp}, h_{cf}, \theta_{exp}(N)) \frac{P(h_{tp}|h_{cf}, \theta_{exp}(N))}{P(h_t|h_{cf}, \theta_{exp}(N))} \\
\\
&with : \\
P_{hg}(h_{fp}|h_{tp}, h_{cf}, \theta_{exp}(N)) &= \frac{\binom{N-h_{tp}}{h_{fp}}\binom{h_{tp}}{h_{cf}-h_{fp}}}{\binom{N}{h_{cf}}} \\
&and : \\
\frac{P(h_{tp}|h_{cf}, \theta_{exp}(N))}{P(h_t|h_{cf}, \theta_{exp}(N))} &\overset{h_{tp} = h_t - h_{fp}}{=} \frac{N - h_{cf} + 1}{N + 1}
\end{aligned}
\tag{2.6}
$$

Here h_{fp} is modelled as a random variable representing all false positive identifications, h_t all identifications, h_{tp} all true positive identifications, h_{cf} all target identifications being matched by at least one false positive PSM and $\theta_{exp}(N)$ the parameters of the proteomics experiment including the database size N.

2.3.3 *Computational frameworks for mass spectrometry*

Diversity of analysis software

The entire computational workflow for the identification and possibly quantification of peptides or proteins from MS raw measurements requires very specialized analysis processes. It involves multiple individual steps such as “decoy” sequence generation, “database search” and “FDR estimation". All of these steps can be highly customized and parametrized and can make use of a multitude of existing software solutions. Several software frameworks for example the OpenMS framework for computational MS [17, 92, 93] or the Trans Proteomic Pipeline (TPP) [94] exist, that allow executing individual MS data processing steps. In particular, the OpenMS framework has been applied and extended throughout this thesis work. OpenMS is implemented in C++ and is an open-source library that provides unrestricted access to a multitude of MS-related functions and maintained by a community of developers.

The OpenMS toolbox

The OpenMS software allows interaction through a multi-layer software architecture. The fundamental functions of the C++ library are not only available as part of a code library, but they are also directly applicable by more than 180 executable task-specific C++ tools that were built from the code base and read and write results in standardized formats [17, 93]. In addition bindings to easy-to-use scripting languages such as Python and R are provided that allow rapid prototyping of desired applications. Moreover, it is possible to wrap existing soft-

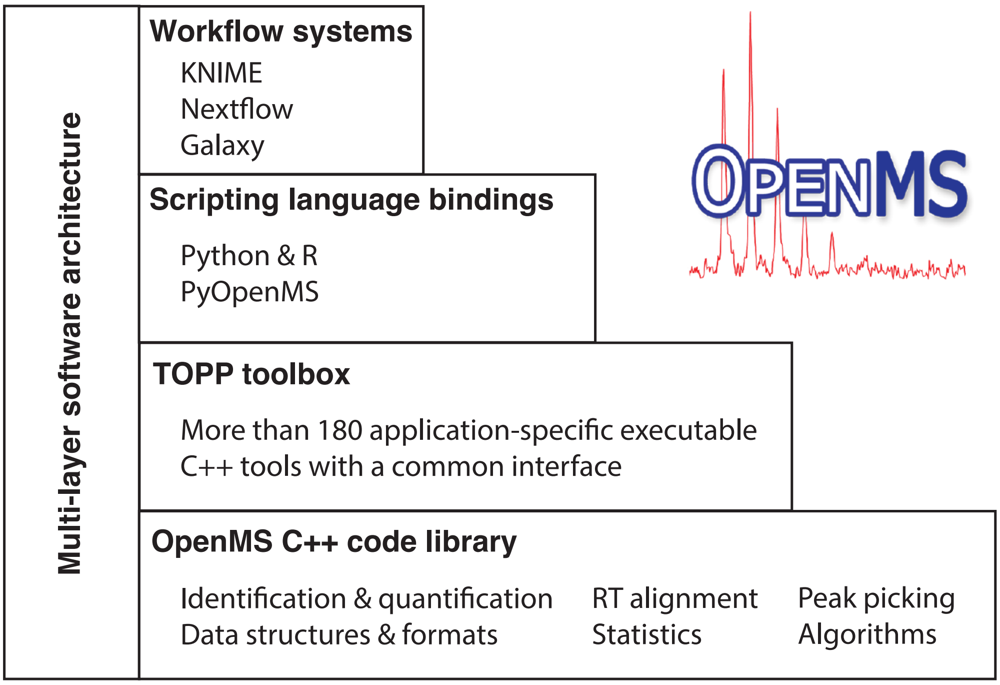

Multi-layer software architecture
Workflow systems
KNIME
Nextflow
Galaxy
Scripting language bindings
Python & R
PyOpenMS
TOPP toolbox
More than 180 application-specific executable
C++ tools with a common interface
OpenMS C++ code library
Identification & quantification
Data structures & formats
RT alignment
Statistics
Peak picking
Algorithms
OpenMS

Figure 2.15: A) The multi-layer software architecture of the OpenMS toolbox: The foundation of the software is based on a comprehensive C++ code library that includes many functions and algorithms for basic tasks regarding mass spectrometry. From this more than 180 executable application-specific C++ tools have been generated and the code is also accessible via bindings to scripting languages such as Python or R. Finally workflow systems can be build around tools and scripts to connect the building blocks into functional systems [17, 92, 93, 95].

ware tools that are not yet integrated into the OpenMS code by creating "third-party adapter tools". In this way it is possible to create whole workflows combining different processing steps in the desired order with a given parametrization. (Figure 2.15) In contrast to "black-box" closed-source MS software [96] this gives an enormous flexibility to the development of analysis workflows such that they can be tailored to specific problems. Finally, many of the available OpenMS tools are parallelizable - meaning that they allow scalable execution times on large computer systems. This is another major advantage to many software solutions as large amounts of data can be processed in parallel on high-performance computing systems.

2.3.4 *Workflow systems*

Diversity of worklow systems

A workflow describes the orderly consecutive execution of a chain of processes. Hereby, the linkage between each process can range from simple linear to complex multi-connected dependencies. Several workflow systems such as KNIME [97], Nextflow [98] or Snakemake [99] exist that can be used for the arrangement of steps in a given workflow. For instance, the KNIME workflow system is graphical, allowing to arrange execution steps symbolized as nodes via directed edges in a graphical user interface (GUI) [95]. Nextflow on the other hand for example does not provide a GUI but allows to define clearly structured processes in the java based groovy language that are executed in the respective order. Hence, these systems are powerful as they allow easy understanding, sharing and reproducibility of complex workflows such as required for the analysis of MS data. As bioinformatics pipelines should be designed according to the FAIR (findable, accessible, interoperable and reusable) data principle [100], workflow systems and publishing them in online available hubs or repositorys provide a feasible solution to attain to the FAIR principles [101] (Figure 2.16).

Continuous integration and deployment

Continuous integration and deployment are modern techniques for agile software development. It involves the continuous testing and installtion of the software on various systems throughout the development. Only when passing all tests an introduced change should be accepted. Today continuous integration and deployment have become a state-of-the-art programming practice, in order to ensure reliable performance when changing features and processes in high-quality computational software [102]. Continuous build systems allow automated testing of rudimentary functions, classes, integrated tools and entire workflows [103]. Hence, when developers commit a change to the code base, the new code and its effect on a given tool or workflow will be first automatically tested by various predefined unit tests, small or real sized test data sets. Once executed the test results will be communicated back to the developer establishing a feedback loop. The continuous integration and deployment therefore ensures the production of the anticipated results and prevent unforeseen changes. For these reasons, all workflows established as part of this thesis and the OpenMS computational toolbox itself are developed using continuous integration systems.

Containerization

Ultimately, in order to achieve reproducibility and interoperability of workflows on the various available computer systems and their respective environments, containerization has been

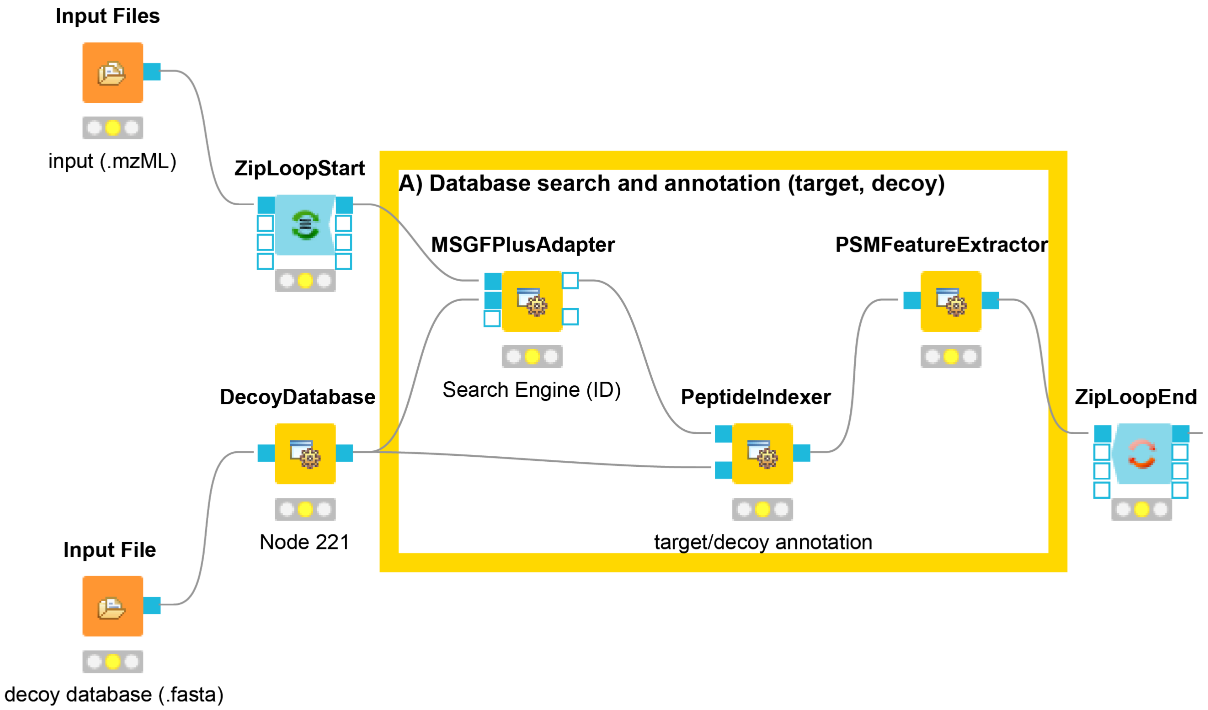


Figure 2.16: An example of an OpenMS data analysis workflow implemented through the graphical workflow system KNIME. Nodes correspond to individual computational steps in the analysis pipeline and edges indicate data input and output flow between the steps.

proven to be a successful solution. Containers such as provided through docker [104] or singularity [105] are lightweight virtual machines that are independent of the used computer system and work as a portable environments that can be set up to contain all dependencies required by a given program [106]. Hence, using containers to store and execute a given workflow allows to execute the workflow on any system with the exact environment it requires for successful execution.

CHAPTER 3

MHCquant

"Automated and reproducible data analysis for immunopeptidomics"

This chapter includes partially identical or adapted content with permission from:

MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics

L. Bichmann, A. Nelde, M. Ghosh, L. Heumos, C. Mohr, A. Peltzer, L. Kuchenbecker, T. Sachsenberg, J. S. Walz, S. Stevanović, H.-G. Rammensee, O. Kohlbacher.

J. Proteome Res. 18, 3876–3884 (2019)

A detailed description of the contributions to the project by coauthors is provided in the Appendix D

3.1 *Introduction*

Numerous experimental methods have been developed and optimized to efficiently purify and extract HLA ligands from biological samples [14, 107, 108], yet computational approaches in immunoinformatics have only scarcely targeted the processing of MS raw data, specifically regarding the HLA-presented peptidome [71, 109, 110].

A major drawback is that the immunopeptidome arises from rather unspecific cleavage steps, hence the database search space is much larger than that of tryptic peptides from a typical proteomics experiment [20, 21]. The fact that a significant proportion of the immunopeptidome may be derived from post-translational splicing and non-canonical reading frames renders the problem even more complex [111–115]. The Human Immunopeptidome Project (HIPP) has thus identified the lack of tailored MS data processing strategies as a major hurdle to improving current analysis protocols [113, 116]. Dealing with these large search spaces and big data sets has led to false discoveries and therefore also requires a robust estimate and rigorous control of false discovery rates (FDRs) to prevent accumulation of false positive identifications [84, 114, 117]. In addition, accurate quantification of epitope abundance with few missing values remains a challenge and could lead to new immunological insights [118].

With the increasing availability of suitable clinical samples there is now a high demand for automated, reproducible pipelines allowing quick processing of experimental data, as well as the large-scale re-analysis of the growing wealth of public data, for example from repositories such as PRIDE [119] or SysteMHC [120]. Previously, immunopeptidomics data has been processed using various other tools and pipelines [112, 114, 120, 121], however at the time of this work there existed no containerized, version-controlled workflow solution specifically tailored to immunopeptidomics data. Recently, the field has attracted more researchers in this area and a similar new approach has been developed [22] and others might follow.

This chapter introduces MHCquant - a novel open-source computational pipeline to identify and quantify HLA-presented peptides from large-scale HPLC-MS raw data. The processing steps encompass database search, FDR estimation, label-free quantification, and HLA-binding affinity prediction. The FDR can be evaluated on multiple levels (PSM, peptide, or protein) and an optional setting allows to refine the FDR on the subset of confident and predicted binding PSMs leading to the rescue of high-confidence PSMs below the conventional FDR threshold [117]. Moreover, the pipeline is containerized, versioned, and available as part of the nf-core initiative for reproducible bioinformatics workflows [101]. Container-based virtualization allows portability, numerical stability, and efficient parallel execution on HPC environments enabling reproducible results for high-throughput computational analysis. In addition, the KNIME integration platform [97] offers easy to use and customizable workflows within a data analysis environment [122]. The workflow has been integrated into the biomedical data management platform qPortal of the Quantitative Biology Center Tübingen as a web-based option [19]. In contrast to other available proteomics software, MHCquant applies targeted feature extraction as a method for label-free quantification and achieves nearly complete quantification of all identified peptides, which can be transferred across runs [16]. At the heart of the identification workflow, we use well-stablished tools (i.e., Comet and Percolator), which are both computationally efficient and highly sensitive ensuring sensitive discovery of neoepitopes.

While we compared MHCquant's performance on a benchmark data set of HLA-I immunopeptidomics samples, where it yields a superior identification rate compared to other solutions, all settings can be adapted for HLA-II data as well. Application of the MHCquant workflow to a previously published melanoma data set (PRIDE: PXD004894 and SysteMHC: SYSMHC00023) revealed fourteen instead of eleven identified mutated neoepitopes reported in the initial publication (27% increase). Experimental validation with spectra of synthetic peptides confirmed these identifications.

3.2 *Materials and Methods*

Biological samples

In order to generate a comprehensive benchmark data set for MHC ligand identification, we generated LC-MS/MS runs from both healthy donors and established cancer cell lines. A data set of 38 HLA immunopeptidomics raw files from peripheral blood mononuclear cells (PBMCs) obtained from nine healthy donors as well as four JY cell line samples were used to assess the performance of the pipeline at each processing step. PBMCs from tissue samples were collected at the University Hospital of Tübingen after written informed consent of donors in accordance with the Declaration of Helsinki. Four-digit HLA typing was carried out by the Department of Hematology and Oncology, University Hospital Tübingen, Germany using sequence-based typing (Luminex). HLA typing characteristics of samples are provided in (Table C.1). JY cells were cultured to obtain $2x10^{7)}$ cells, centrifuged at 1,500 rpm for 15 min at 4 °C, washed twice with cold PBS and aliquoted in four vials containing $75x10^6$ cells frozen at -80 °C. For the assessment of the quantification performance, four separate aliquots were spiked with 66 isotope-labeled peptides with known HLA restrictions yielding concentrations of 0.1, 1, 10, and 100 fmol respectively. A detailed table of the exact sequences utilized, as well as heavy isotope labels is contained in the original publication bichmann et al. [123] Supporting Material Table S3.

Peptide synthesis

Peptides were synthesized using the automated peptide synthesizer Liberty Blue (CEM) using the 9-fluorenylmethyl-oxycarbonyl/tert-butyl (Fmoc/tBu) strategy. In case of isotope-labeled peptides V($^{13}C_5$, ^{15}N), L($^{13}C_6$, ^{15}N) or P($^{13}C_5$, ^{15}N) was used. For the carbamidomethylation of synthetic peptides 1 mM dithiothreitol was added, after 1 h of incubation at room temperature 5.5 mM iodoacetamide was added and incubated in the dark for 1 h.

Analysis of HLA ligands by LC-MS/MS HLA complexes were isolated by standard immunoaffinity purification [124, 125] using the pan-HLA-I specific W6/32 antibody (produced inhouse) as previously described [126]. Sample amounts and applied antibody amounts are specified in Supporting Material Table S1. HLA-peptide extracts were analyzed in two to five technical replicates. Peptides were loaded on a 75 µm x 2 cm C18 Nano Trap Column and separated by nanoflow high-performance liquid chromatography (RSLCnano, Thermo Fisher Scientific) using a 50 µm x 25 cm PepMap rapid separation liquid chromatography column (Thermo Fisher Scientific) 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 Fusion Lumos mass spectrometer (Thermo Fisher Scientific, Resolution: 120,000 for MS1 acquisition and 30,000 for MS2, both at 200 m/z) equipped with a nano electrospray ion source using data dependent acquisition mode (DDA) employing a top speed collision-induced dissociation fragmentation method (CID, normalized collision energy 35%). Survey scans were performed in the orbitrap at a resolution of 120,000. MS/MS scans were detected in the orbitrap at a resolution of 30,000. Dynamic exclusion was set to 7 s. Mass range was set to 400-650 m/z with charge states 2+ and 3+ selected for fragmentation.

Validation of synthetic peptide spectra

All potentially novel identified mutated neoepitopes were synthesized as synthetic peptides and analyzed using the identical method and instrumental settings as described in a previous publication [46]. Measurements were carried out at the Proteome Center Tübingen on an online-coupled QExactive mass spectrometer. Spectra were inspected and annotated using the OpenMS tool TOPPView (version 2.4.0).

Bioinformatic pipeline construction

MHCquant is an integrated data processing pipeline (Figure 3.1) implemented in Nextflow and KNIME. It includes a range of different tools from the OpenMS software package (version 2.4) [17] and other scientific software from the proteomics and immunoinformatics domain (Table 3.1). Some of the steps are optional and will only be carried out if it was specified when executing the workflow. The processing steps of MHCquant are discussed in the following in detail:

PIPELINE INPUT The input to the pipeline is four-fold and specified in a sample sheet: a set of LC-MS/MS raw files (mzML or raw format), a protein database (in FASTA format), a variant calling file (VCF format) containing putative neoantigenic mutations and a file specifying the HLA allotypes (in TSV format).

RAW FILE CONVERSION In a first, optional step provided DDA raw MS measurements (Thermo Raw vendor format) are converted to the open, XML-based mzML format [127].

DECOY-GENERATION In order to compute a FDR based on the target-decoy approach random decoy sequences need to be generated and appended to the input fasta database. For this all sequences in the input fasta database are reversed in a first step and appended using the OpenMS-DecoyGenerator tool.

ADDITION OF VARIANTS TO THE PROTEIN DATABASE If specified annotated genomic variants from the input VCF file are translated into mutated protein sequences and appended to the provided FASTA database using the Fred 2.0 Immunoinformatics framework [128]. Within this procedure transcripts that contain the annotated variant are fetched from the online availble Ensembl database [129] using the BioMart API [130]. The corresponding single nucleotide changes caused by the variant are introduced to the transcripts and their respective translated protein sequences are appended to the protein database.

DATABASE SEARCH The database search is carried out applying the proteomic search engine Comet [83] and unspecific enzymatic restriction allowing to search the mass spectrometry data for all possible peptides of a defined mass, length and charge range. It is wrapped into the OpenMS adapter tool CometAdapter that was implemented in C++ and added to the OpenMS software code base for this purpose. In this way it is possible to execute the database search in a standardized way with well defined parameters and an XML-based output format (IdXML) that can be read by follow-up OpenMS processing tools. A multitude of parameters are available to tune the search engine to match the respective instrumental requirements for example the precursor and fragment mass tolerances. In addition, it can be specified whether to search for fixed or variable mass shifts caused by post-translational modifications that can be selected from all available pre-defined OpenMS modifications in the MS data and which fragment types should be included in the PSM scoring.

identification-based retention time alignment MHCquant corrects for retention time shifts across replicate MS runs that serve as input to the pipeline. For this a linear retention time alignment is computed based on shared identifications across runs passing a given q-value threshold. This functionality is integrated by applying the OpenMS MapAligner-Identification tool.

fdr estimation The FDR is computed on the merged set of all identifications using the post-scoring tool Percolator (version 3.0.1) [88]. Merging the identifications of a given sample group is beneficial for the FDR scoring as it reduces the error accumulation to each sample-group instead of for each individual MS run. Moreover, the semi-supervised machine learning strategy of the Percolator algorithm gets access to more training data for its FDR scoring model, which should improve its accuracy. Percolator is executed by using the OpenMS PSMFeatureExtractor that computes multiple scores for each PSM, excluding enzymatic restriction specificity information for the scoring model and passes them to the OpenMS PercolatorAdapter tool to train the scoring model.

subset fdr estimation In order to obtain a better FDR estimate, MHCquant has the option to re-evaluate the FDR on the subset of PSMs passing either a conventional q-value or a given predicted affinity threshold (subset FDR mode). Corresponding reversed decoy or target counterparts of identified sequences in this subset are kept for the rescoring as well, even when not fullfilling the q-value or affinity score thresholds. A smiliar variant of this technique had previously been applied successfully to similar problems in metaproteomics yielding superior results [117].

label-free quantification Targeted, identification-based, label-free quantification is achieved through the OpenMS tool FeatureFinderIdentification [131]. In this step the areas of MS1 chromatograms in proximity to the PSM of a given peptide identification are integrated. Similar to the matching-between-runs procedure, PSMs that were identified in another MS runs of the same sample group can be transferred and their MS1 area integrated in a targeted manner in other MS runs as well to reduce missing values in the peptide quantification.

feature linking Quantified peptides of each MS run from the same sample group are finally linked across runs to create consensus features. This functionality is provided through the application of the OpenMS tool FeatureLinkerUnlabeledKD. Conflicting identifications mapping to the same feature are removed by choosing the best scoring identification.

text and mztab export Finally, results are exported in plain tab separated text and the community standard format mzTab [132].

mhc affinity prediction If specified, exported peptide sequences can be annotated with affinity prediction results based on MHCflurry, [133] MHCnugget [77] in the Nextflow implementation and predictors available in the ImmunoNodes toolbox (SYFPEITHI, NetMHC and PickPocket) in the KNIME implementation [122]. All MHC binding predictions within the workflow are carried out on corresponding unmodified peptide sequences.

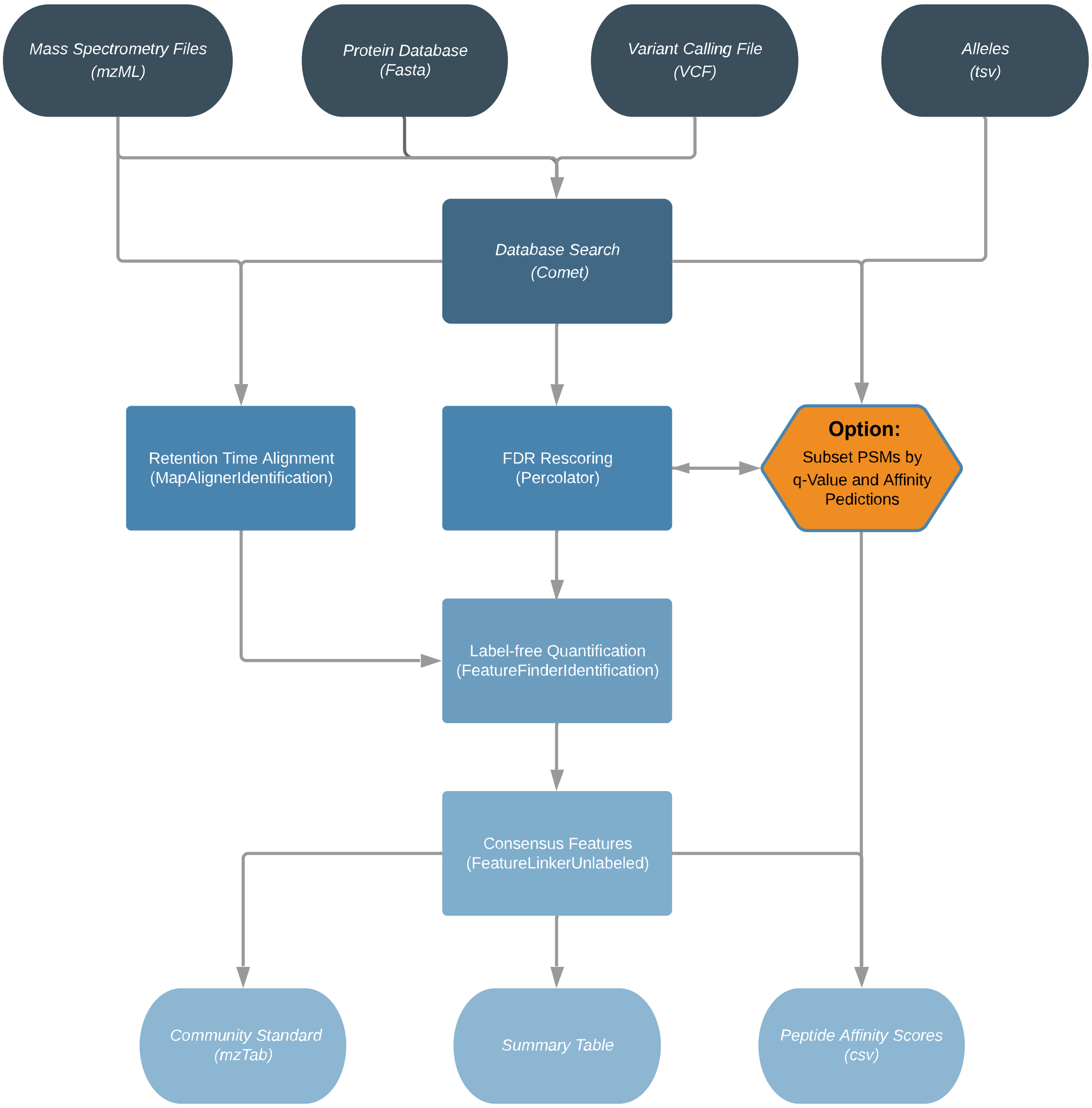


Figure 3.1: Simplified scheme of the MHCquant peptide identification workflow: MS raw files (mzML), a protein database (FASTA), a variant calling file (VCF) and an HLA allele table serve as input. Database matching is carried out using the Comet search engine. Identifications from each sample are used to carry out a retention time alignment using MapAlignerIdentification (OpenMS). The false discovery rate (FDR) is calculated based on reversed decoy sequence hits using Percolator. Optionally, the FDR can be re-evaluated on the subset of PSMs passing either a conventional q-value or a binding affinity threshold, yielding increased identification rates. The resulting identifications are quantified using FeatureFinderIdentification (OpenMS). Search results are exported as community standard format (mzTab), as a summary table and scored by available binding predictions.

Step	Requirement	Name	Employed software tool	
1	(optional)	Generate proteins from variants	FRED2.0 Python library	Data Preprocessing
2	(optional)	Generate decoy-database	OpenMS-DecoyGenerator	
3	(optional)	DDA raw file conversion	ThermoRawFileParser	
4	(optional)	Peak picking	OpenMS-PeakPickerHiRes	
5	required	Database search	OpenMS-CometAdapter	Database search
6	required	Peptide indexing	OpenMS-PeptideIndexer	
7	required	Calculate FDR for ID-based RT alignment	OpenMS-FalseDiscoveryRate	Identification-based RT alignment
8	required	Select IDs by FDR for RT alignment	OpenMS-IDFilter	
9-11	required	ID-based RT alignment	OpenMS-MapAlignerIdentification	
12	required	Merging of all IDs	OpenMS-IDMerger	
13	required	Extract PSM features for FDR scoring	OpenMS-PSMFeatureExtractor	FDR estimation
14	required	FDR scoring	OpenMS-PercolatorAdapter	
15	required	Filter by q-value	OpenMS-IDFilter	
16-19	(optional)	Predict MHC class I affinities of all PSMs	MHCFlurry	Subset FDR estimation
20	(optional)	Filter by q-value or MHC affinity	OpenMS-IDFilter	
21	(optional)	Rescore SubsetFDR	OpenMS-PercolatorAdapter	
22	(optional)	Filter by SubsetFDR q-value	IDFilter	

Step	Requirement	Name	Employed software tool	
23	required	Quantify IDs targeted	OpenMS-FeatureFinderIdentification	Label-free Quantification
24	required	Link extracted features	OpenMS-FeatureLinkerUnlabeledKD	
25	required	Resolve conflicts	OpenMS-IDConflictResolver	
26	required	Export text	OpenMS-TextExporter	Output summary
27	required	Export mzTab	OpenMS-MzTabExporter	
28	(optional)	Predict MHC class I affinites	MHCFlurry	MHC affinity prediction
29-31	(optional)	Predict MHC class II affinites	MHCNuggets	
32	(optional)	Predict theoretical class I neoepitopes	FRED2.0 Python library	Neoepitope annotation
33	(optional)	Predict theoretical class II neoepitopes	FRED2.0 Python library	
34	(optional)	Resolve class I neoepitopes	Custom Python script	
35	(optional)	Resolve class II neoepitopes	Custom Python script	

Table 3.1: Detailed description of all steps within the Nextflow implementation of the MHCquant workflow version 1.2.6. The steps and their names may vary across versions of MHCquant as well as between the KNIME implementation of the workflow. Some of the steps are grouped together in this table but the implementation involved the creation of several steps due to file-handing, reformatting, pre- or post-processing files of precedent or follow-up steps. All analysis in this chapter was carried out with the Nextflow implementation.

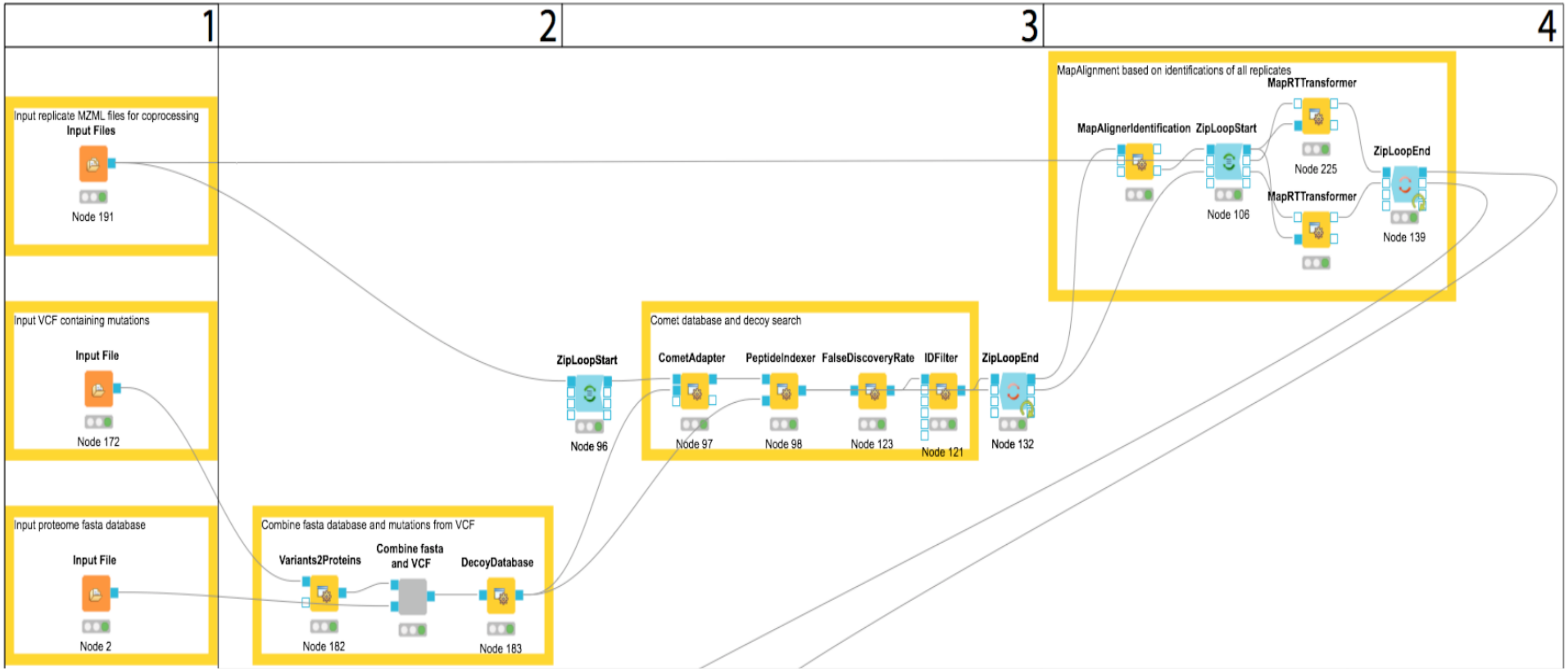


Figure 3.2: MHCquant KNIME graphical workflow depiction. As input serves raw MS data (mzML), a protein reference database (fasta) and optionally a variant calling file (vcf). KNIME nodes are interlinked via loop processes and grouped into the respective workflow steps as follows: 1 - Input files, 2 - Protein database generation (optional, from VCF), 3 - Database Search using Comet, 4 - Identification based retention time alignment)

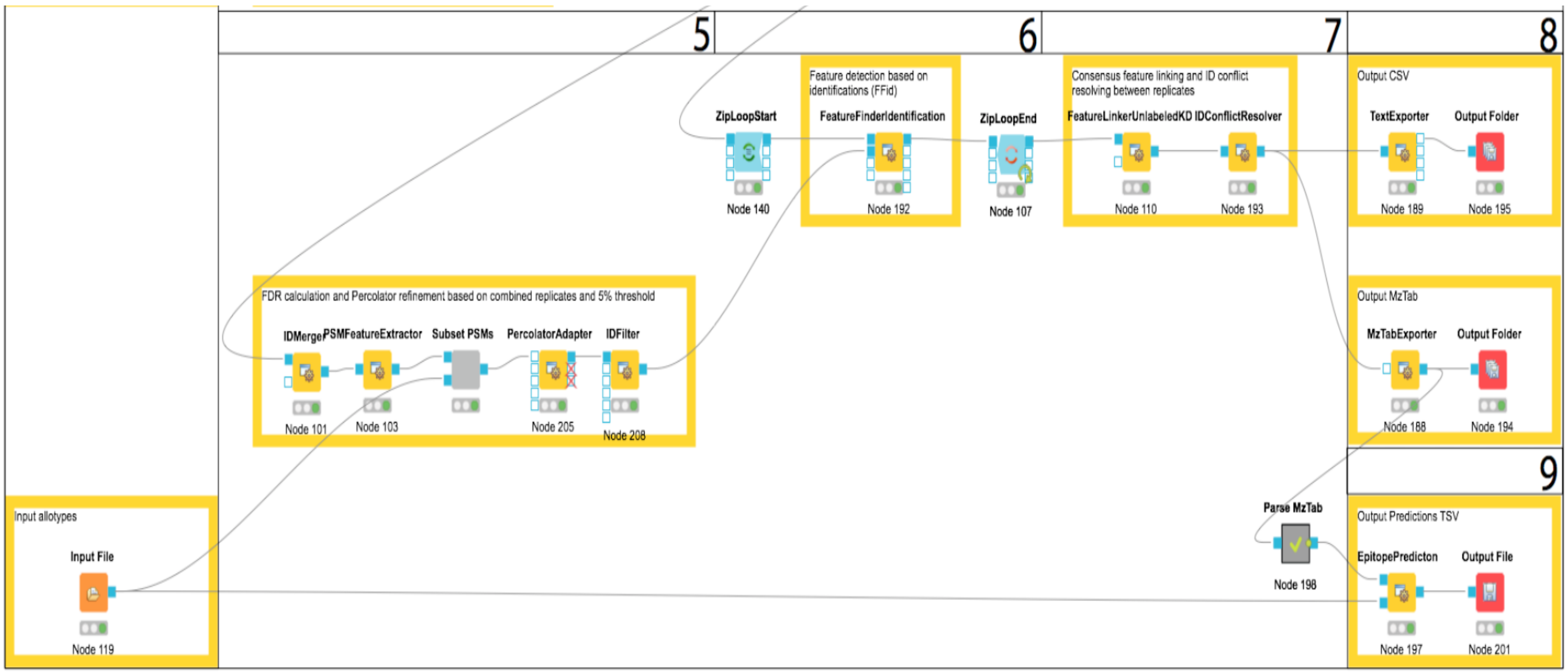


Figure 3.3: MHCquant KNIME graphical workflow depiction, continuation of Figure 3.2 by the workflow steps: 5 - PSM Subset Filtering and FDR calculation, 6 - Targeted Quantification, 7 - Feature Linking, 8 - Text and mzTab Export, 9 - Peptide Binding Prediction (optional)

NEOEPITOPE ANNOTATION In order to make sure that peptide identifications associated with mutated proteins appended to the protein database in a previous step indeed contain the mutation, all theoretical peptides of the specified length range containing the mutation are generated. This set of theoretically possible peptides is finally compared with all peptide identifications and the common peptides, the causing genetic variant and source gene are reported in a csv file.

CONTAINERIZATION AND WORKFLOW SYSTEMS The aforementioned tools are integrated in a pipeline and containerized into an online accessible docker image [104]. Implementation in Nextflow [98] as well as in KNIME allows easy execution on various HPC and compter infrastructures [97]. Implementation in Nextflow based on the nf-core community template for reproducible bioinformatics workflows [101] allows easy execution on various HPC and compute infrastructures such as google cloud or amazon web services. Moreover support for multiple functionalities is provided such as for various container systems (e.g., docker, singularity, podman) and environment management platforms (e.g. Conda).

Reanalysis of existing data and benchmarking

In order to assess the performance of available proteomic search engines we processed the aforementioned data set acquired for this study (PRIDE: PXD011628) as well as a published melanoma data set [46] containing neoepitopes (PRIDE: PXD004894).

MASS SPECTROMETRY DATABASE SEARCH For all search tools unspecific cleavage was set as enzyme specificity and oxidation of methionine residues as the only variable modification (maximum number of modifications per peptide set to three) was selected. In order to achieve a fair comparison, all benchmark search results were assessed with a FDR on the level of PSMs. In contrast, the reanalysis of the malignant melanoma data set [46] required carbamidomethylation of cysteines as additional fixed modification, as iodacetamide was used in the respective purification protocol and the FDR was assessed on peptide-level. The precursor charge was fixed to 2-3 to narrow the search space and avoid susceptibility to the identification of contaminant proteins [134] and the precursor mass tolerance was set to 5 ppm. For MHCquant all search results in this publication were achieved using the Nextflow implementation (revision 1.2.6 - `https://www.openms.de/mhcquant/`). Within MHCquant the Comet search engine (version 2016.01 rev. 3) [83] parameter fragment tolerance bin was set to 0.02 Da and the fragment bin offset to 0 as recommended for high-resolution instruments. Fragmentation was set to CID and neutral losses were enabled for peptide-spectrum matching. The digest mass range was set to 800-2,500.

Where applicable, identical parameters were used for the other search tools and the percolator enzyme specificity was set to no enzyme. In order to benchmark the performance MS-GF+ [135], the Comet search engine step within MHCquant was replaced with the MS-GF+ search engine (version 2017.01.13) and the instrument option was set to high resolution. PEAKS [136] was used within PEAKS Studio (version 8.5) and only database search and no *de novo* results were used for the benchmark. SequestHT [82] and Mascot (version 2.2.04) [137] were used within Proteome Discoverer (version 1.4) including Percolator (version 2.0.4). For MaxQuant (version 1.6.3.3) [96] the additional first search peptide tolerance was set to 20 ppm and the instrument type to Orbitrap. Match between runs was enabled – using a match time window of 0.7 min and an alignment time window of 20 min. In addition, the MaxQuant protein-FDR Parameter was set to 0.9 to have comparable results with the other search engines and filter on PSM FDR only. All search engine results were filtered by a stringent length criterion of 8-12-mers only in order to compare only the most abundant HLA bound peptides. While this is the default setting of MHCquant, these length restrictions can be readily modified by changing the corresponding parameters.

HLA AFFINITY PREDICTION Epitope binding affinity predictions for PSM subset filtering were computed using MHCflurry (version 1.0) and a binding affinity threshold of 500 nM - the default setting of the MHCquant (revision 1.2.6) Nextflow implementation. For the benchmark results, predicted HLA binding peptides were defined as the union of the predicted binders from NetMHCpan (version 4.0), [37] NetMHC (version 4.0), [138] MHCflurry (version 1.0) [133] SYFPEITHI (version 1.0), [32] and PickPocket (version 1.0) [139]. For NetMHC and NetMHCpan rank score thresholds of <2%, for MHCflurry and PickPocket molar affinities <500 nmol/l and for SYFPEITHI half-maximal scores were required for classification as predicted binding peptides. Overlaps of search results and binding predictions were computed using Jvenn [140].

RETENTION TIME PREDICTION Retention time predictions were carried out as an additional post-processing step external to the pipeline. For this the OpenMS tools RTModel and RTPredict based on *nu*-support vector regression (oligo-kernel, ν =0.5, p =0.1, c =1, degree =1, border_length =22, kmer_length =1, Σ=5) [141–143] were employed. Training was performed on the peptide retention times of the 100 most confident consensus identifications (lowest q-value) of each sample without modifications. Subsequently all other peptide identifications of the corresponding samples were predicted using the trained models. The default parametrization was kept and not further optimized using a cross-validation since most retention times were predicted within +/- 12.5min of their experimentally determined retention times. The performance of the search engine results from the benchmark was evaluated applying a linear least squares fit to compute a 99 % prediction interval for the regression of predicted versus observed retention time using the scipy [144] and sklearn Python modules [145].

3.3 *Results*

MHCquant is more sensitive than previous approaches

In order to assess the performance of the MHCquant pipeline, we compared the HLA-I peptide identifications from an in-house benchmark data set of nine PBMC and four JY cell line samples against the open source available proteomic search engine MS-GF+ [135] within OpenMS, the commercially licensed search engines PEAKS within PEAKS Studio, SequestHT [82] and Mascot [137] within Proteome Discoverer and the Andromeda search engine within MaxQuant [96].

We evaluated the performance of each search tool (1) by the number of peptide identifications that are predicted HLA-binders at a q-value threshold of 1%, (2) by the rate of predicted HLA-binding peptides among all peptide identifications. (Figure 3.4 A) To avoid bias of a specific epitope prediction method we applied five different HLA binding prediction tools (NetMHCpan 4.0 [37], NetMHC 4.0 [138], MHCflurry [133], SYFPEITHI [32] and PickPocket [139]).

MHCquant in subset FDR mode identified the highest number of unique, predicted HLA-binding peptide sequences, directly followed by PEAKS and MHCquant in default mode. Peptide hits produced by all search engines follow the expected [41, 146] distribution of peptide lengths ranging from 8-12 amino acids and a maximum at length nine. Regarding the quality of peptide identifications, all search engines reveal rates of 87 % to 99 % HLA binding peptides among their identifications as seen previously for PBMC and JY samples [147]. Moreover, for none of the investigated search engines we observed evidence for any motif bias as all tools performed equally across the various allotypes, non-tryptic and tryptic restriction sequence endings Figure **??** and Figure A.4 in the data set. SequestHT, Mascot, and MaxQuant achieve slightly higher and PEAKS the lowest rates of binders among their identifications in comparison with the other tools. Conversely, SequestHT, Mascot, and MaxQuant (Figure 3.4 A) yield a smaller absolute number of identified predicted HLA-binders.

The overlap of peptide identifications between search engines indicates that most unique identifications are revealed using MHCquant and PEAKS (Figure 3.4 B) and the rate of predicted binders among these unique identifications is greater 70 %. In fact, all additional unique identifications yielded using MHCquant's subset FDR refinement option are predicted binders.

In conclusion, we find that the MHCquant pipeline provides the best trade-off between the number of identified peptides, the rate of predicted HLA-binders among the search engines tested in this benchmark.

Validation of identified peptides using a retention time predictive model

In order to gain more confidence into the uniquely identified peptides of MHCquant applied in our benchmark, we evaluated the retention time prediction error for all the peptides identified at 1% FDR in consensus of all search engines (all) and those that were identified by MHCquant in default and subset mode uniquely (unique). The predictive model was trained on the retention times of the 100 most confident consensus peptide identifications.

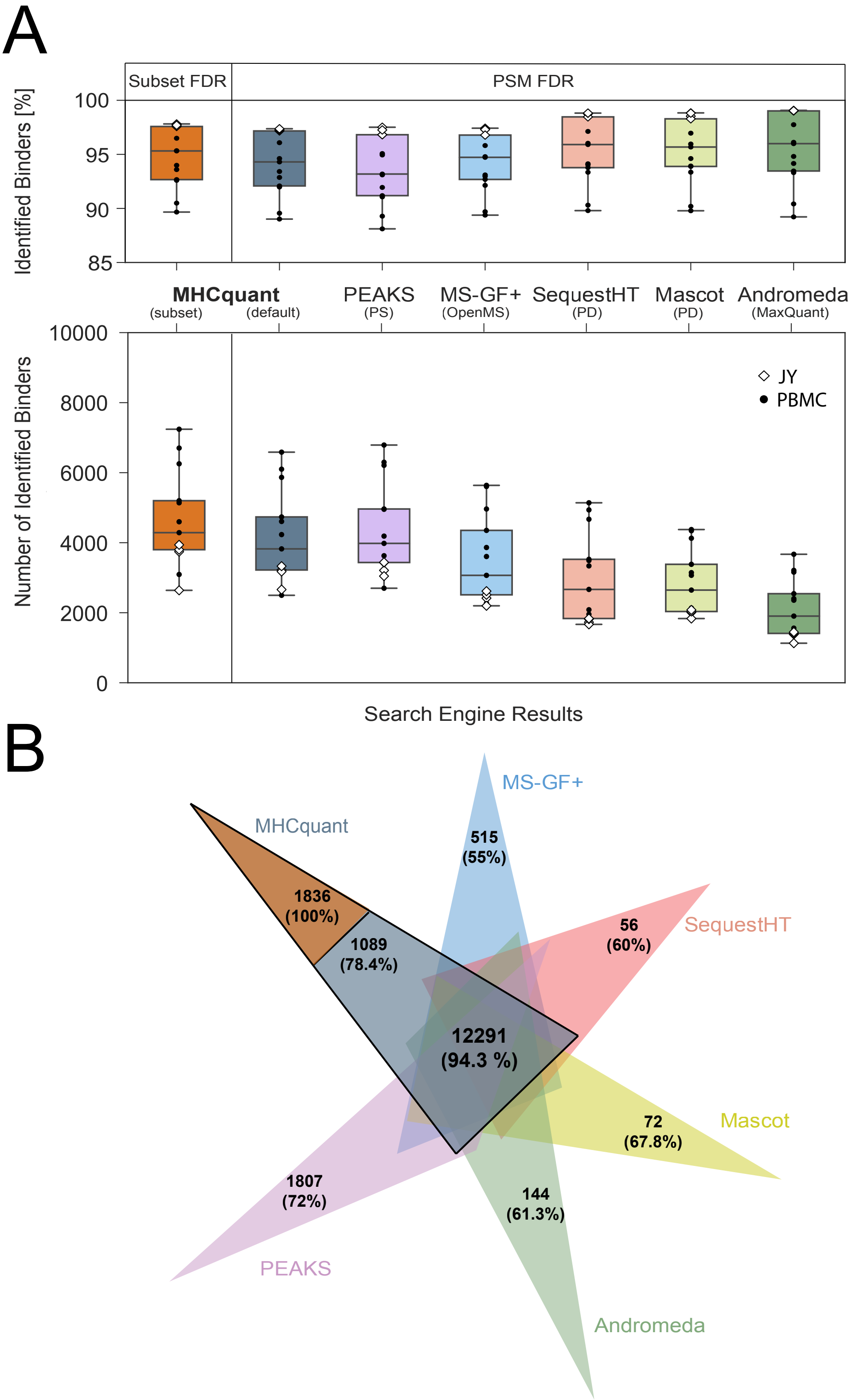


Figure 3.4: Comparison of the pipeline performance. A) The absolute number of identified peptides that are predicted HLA-binders and the rate of predicted HLA-binding peptides among all identifications (median annotated) from 13 samples (9 PBMC and 4 JY) comparing the described MHCquant pipeline at 1% PSM-FDR against PEAKS (PEAKS Studio), MS-GF+ (OpenMS), SequestHT, Mascot (Proteome Discoverer) and Andromeda (MaxQuant). B) Uniquely identified peptide numbers and their corresponding percentage of predicted binders (indicated in brackets) across the different search engines and their overlap.

Finally, the reliability of peptide identifications of MHCquant was assessed by the corresponding percentage of points within the prediction interval comparing (1) all+unique (2) unique (3) unique from a single sample and (4) decoy identifications. In particular the results provide confidence in the additionally identified peptides of MHCquant since 84 - 98% of the additionally identified peptides in contrast to 34 % percent of random decoy peptide hits lie within the 99% prediction interval of retention time predictions that were trained on the search engine consensus identifications. (Figure 3.5) However, it is also evident that there is a greater variance among the increased peptide identifications retrieved using the subset FDR mode. Hence, it indicates that while more peptides can be discovered in this way, running the MHCquant workflow in this mode might be underestimating the actual FDR to some extend.

The influence of subset filtering on the identification performance

In order to better understand the effect of subset FDR re-evaluation, we computed the respective quantities of spectra, PSMs and peptides at each of the consecutive processing steps for MHCquant in both subset and default PSM FDR mode of action.

While the MS samples contain on average around 25,000 MS/MS spectra, the retention through the pipeline shows that these result in around 15,000 target PSMs. Subsequent subselection of PSMs based on predicted HLA-binding affinity or an initial q-value threshold of 1 % PSM-FDR drastically reduces the number of target PSMs to a subset of around 6,000 in subset FDR mode. Re-evaluation of the FDR on the reduced subset of PSMs leads to a median number of 4,000 unique peptides that can be identified at 1 % PSM-FDR control, computed through Percolator (Figure 3.6). As compared to assessing the FDR on the entire set of PSMs in median 12 % more peptides are identified in this way. PSM features that had the highest impact for the iterative refinement procedure of Percolator were mass accuracy-based only and we were unable to identify other features improving this step further.

Using targeted label-free quantification 99 % of these identified peptides can be achieved [131]. Ultimately, depending on the sample characteristics, around 87 % to 99 % of the unique identified and quantified peptides are predicted to be HLA-binders of the respective alleles of a given patient.

Validation of the peptide quantification performance

In an additional experiment to validate the quantification performance of MHCquant, 66 isotope-labelled spike-in peptides in a $\log_{10}$ concentration series from 100 fmol to 0.1 fmol were spiked into JY cells and separately measured. All samples were co-processed treating identifications across runs as internal.

As a result, 58 of these peptides were identified and could be quantified in all runs. When separately processing the MS runs of each concentration step, fewer spiked-in peptides could be identified, especially at concentrations below 10 fmol. Hence, the transfer of IDs across the MS runs that is an intrinsic step of the MHCquant workflow, allowed to quantify these peptides even at concentrations lower than 10 fmol. Consequently, the quantification results reveal a linear trend for the spiked-in peptide concentrations. (Figure 3.7 A) In addition, all predicted HLA-binding peptide LFQ-intensities of the JY cell background were compared in a volcano analysis (Figure 3.7 B), highlighting peptide LFQ-intensity stability in the JY cell

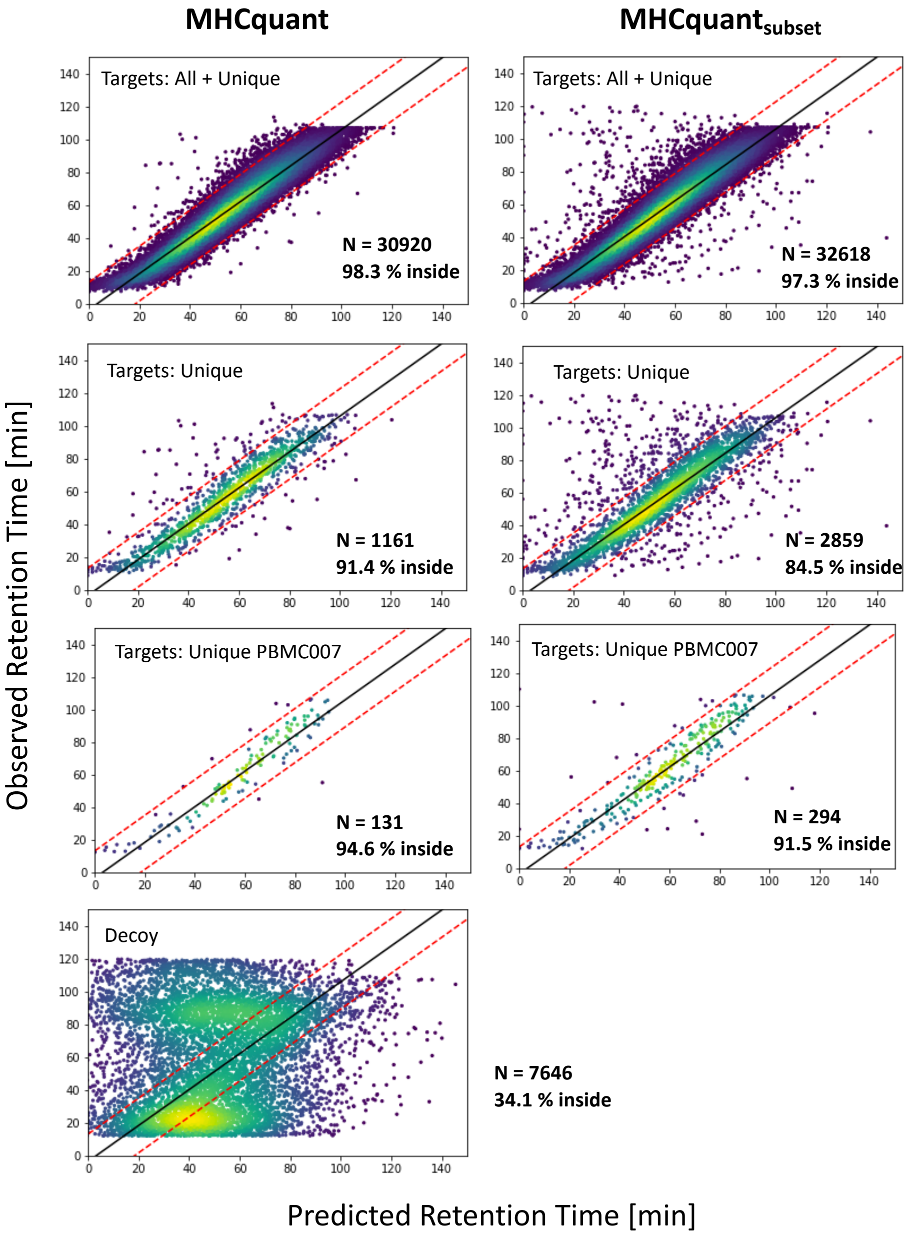


Figure 3.5: Comparison of retention time predictions and observed retention times for search engine consensus within the benchmark (All) and unique identifications of MHCquant (Unique). Each tile in the plot represents the corresponding set of peptide identifications mentioned in the upper left. Points represent peptide identifications in a given set of samples and are colored by its 2d-kernel density computed over the point distribution. N depicts the number of points in each tile and the percentage of points that fall in the prediction interval (red lines) is annotated in the lower right.

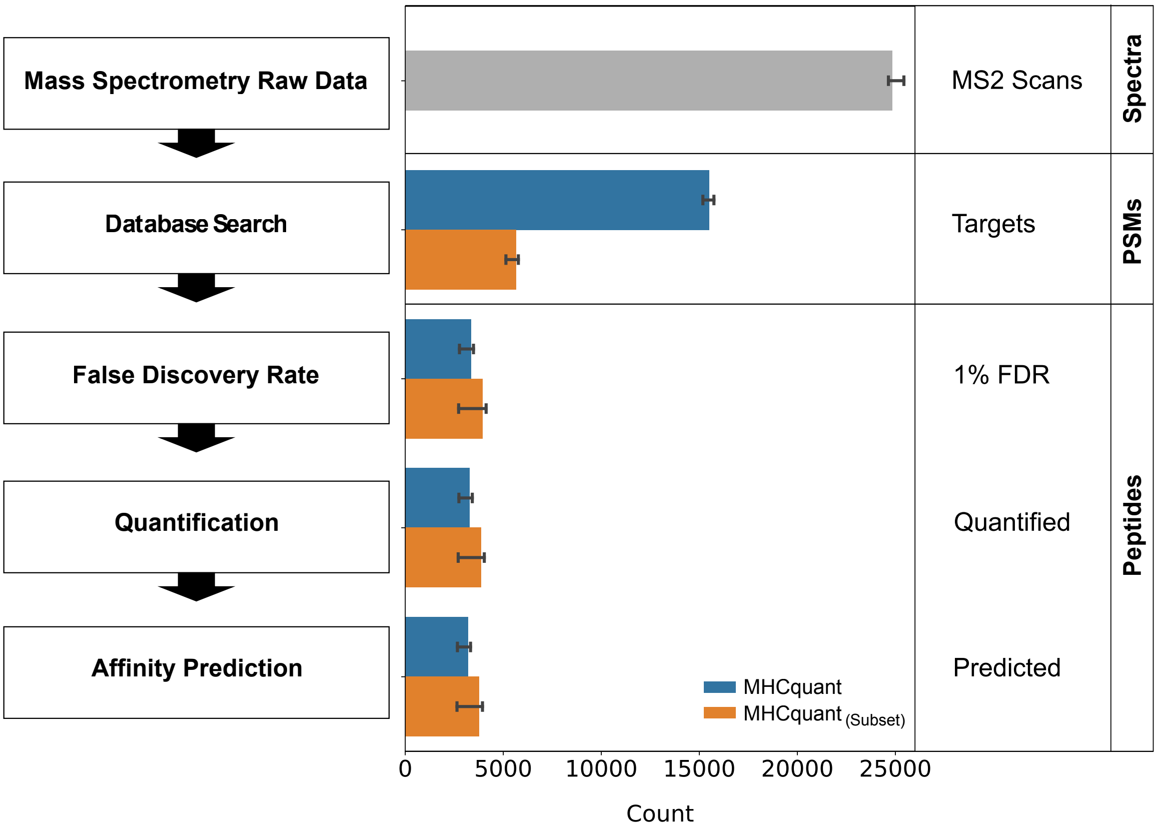


Figure 3.6: Number of spectra or peptides identified in each of the steps (depicted on the right) comparing MHCquant in the default (blue) and subset FDR mode (orange). The initially large number of spectra is drastically reduced to a subset of target PSMs in subset FDR mode. Re-evaluation of the FDR on this subset results in an increase of around 12 % more uniquely identified peptides compared to the default mode. 99 % of the peptides satisfying the FDR condition undergo targeted, label-free quantification and binding affinity predictions estimate that in median 95 % of these final peptides are HLA-binders.

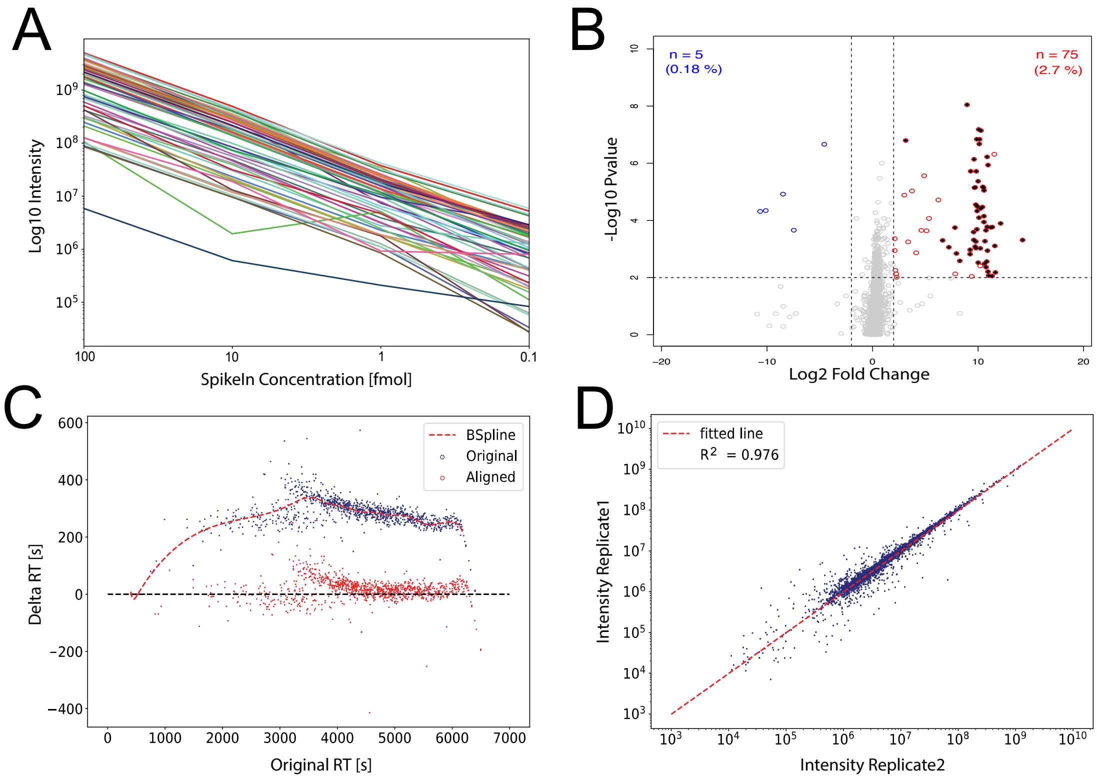


Figure 3.7: A) Quantification of 58 isotope-labelled spike-in peptides in a log_{10} concentration series from 100 fmol to 0.1 fmol. IDs were transferred across measurements achieving a linear trend and the possibility to quantify at 0.1 fmol. Label-free quantification performance. B) Volcanoplot showing all predicted HLA binding peptide LFQ intensity changes in the comparison of 100 fmol with 0,1 fmol spike-in peptide (black dots) concentrations. C) MapAlignmentIdentification retention time alignment based on shared identified peptides across runs (points represent peptides). It achieves a massive reduction in delta RT across two MS runs, drastically differing in their chromatography. D) Quantified intensities of the same peptides across two replicates result in a high correlation.

background in contrast to the spike-in peptides.

The performance of the integral step of ID-based retention time alignment within MHCquant (Figure 3.7 C) was evaluated on two chromatographically distinct samples from our dataset (PBMC001, PBMC008), resulting in a correction of retention time differences across the two runs. Finally, the reproducibility of all peptide intensities measured in two replicate measurements of JY cells (Figure 3.7 D) was compared, resulting in a high correlation across measurements.

Discovery and confirmation of mutation-derived neoepitopes

To check whether the increase in sensitivity of MHCquant over other commonly applied search tools could lead to new discoveries in published immunopeptidomics data sets, we reprocessed the MS raw measurements of a recent malignant melanoma study, containing mutated neoepitopes that were previously identified using the MaxQuant environment 1.5.3. Consequently, we were able to corroborate all of the previously published neoepitope hits using MHCquant (Table 3.2).

Patient	Neoepitope	Mutation	UniProt	Gene	MHC Restriction
Mel8	SPGPVKLEL	P169L	A0A0C4DGU5	NOP16	B*07:02
Mel5	ETSKQVTRW	E161K	Q06546	GABPA	A*25:01
	YIDERFERY	Q125R	Q15019	37500	A*01:01
Mel15	GRTGAGKSFL	S1342F	Q92887	ABCC2	B*27:05
	RLFKGYEGSLIK	P46L	B4E3T4	RBPMS	A*03:01
	GRIAFFLKY	S363F	Q96C24	SYTL4	B*27:05
	RIKQTARK	T4I	Q6NXT2	H3F3C	A*03:01
	ASWVVPIDIK	E689K	P80192	MAP3K9	A*03:01
	KLILWRGLK	P333L	Q86XI2	NCAPG2	A*03:01
	KLKLPIIMK	M1482I	Q13023	AKAP6	A*03:01
	LPIQYEPVL	P52L	Q15436	SEC23A	B*35:03
	FVPPTAISHF	S584F	Q9NR30	DDX21	B*35:03
	ETLKPGTC*VKR	P778L	P49790	NUP153	A*68:01
	SRFLSQLDK	E183K	Q13535	ATR	B*27:05

Table 3.2: Identified neoepitopes from published melanoma data set [46]. Mutated amino acids are colored in red in the peptide sequence. The additionally identified peptides at the bottom of the table are highlighted in bold letters and the cysteine residue marked with an asterisk carries a carbamidomethyl modification. For each peptide the highest scoring MHC restriction is annotated in addition.

Moreover, we discovered three additional hits for novel potentially mutated neoepitopes (Table 3.2) that were not reported in the original publication and verified them by comparison to spectra of their synthetic peptide counterpart (Figure 3.8 and Figure B.1). All three peptide hits show quantities in a similar range as the previously detected neoepitopes that rank in the middle of all intensities in the samples. One of the peptide hits (NUP153 P778L) carries a cysteine carbamidomethyl modification, which may have prevented its detection in the original publication. Yet, carbamidomethyl modified peptides represented only about 2% of all identified peptides in our reanalysis with MHCquant. In conclusion, MHCquant allows searching immunopeptidomics data in a reproducible way and its increased sensitivity can potentially lead to novel discoveries.

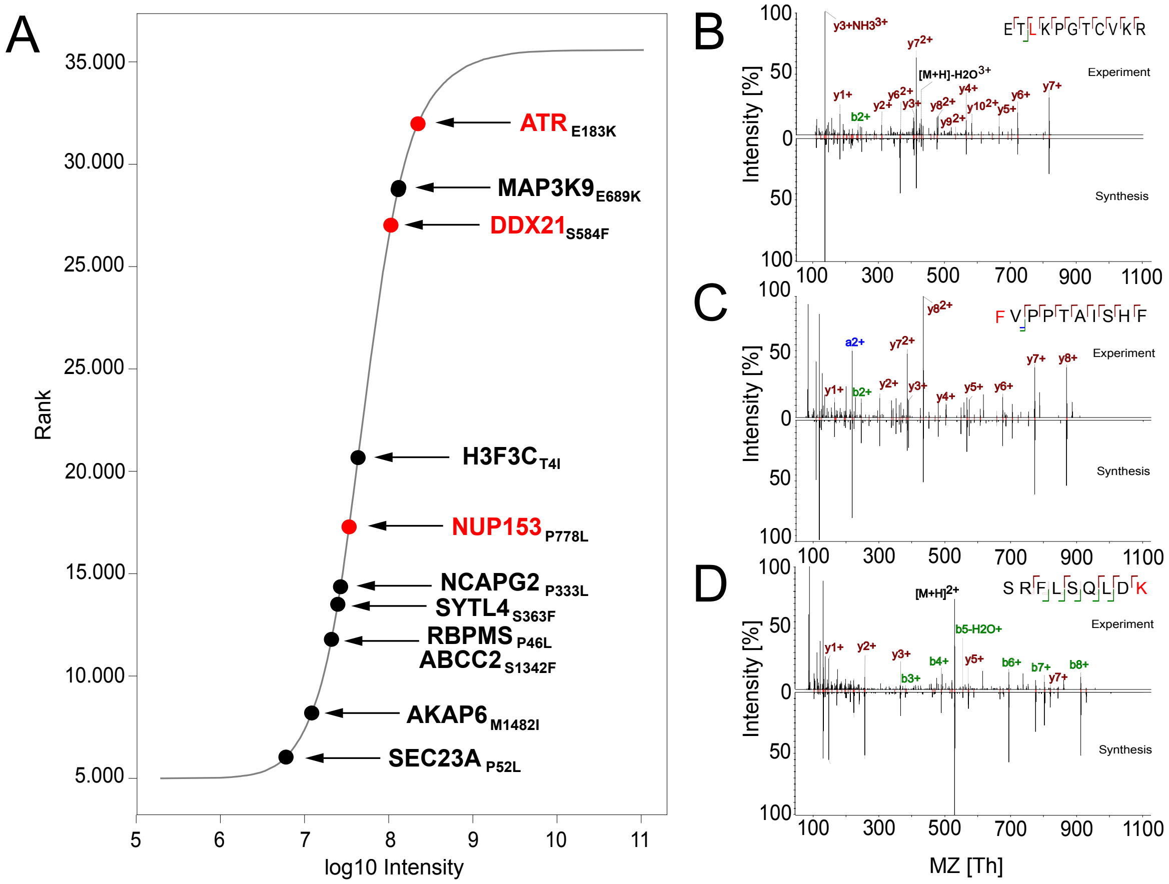


Figure 3.8: Relative LFQ intensities (A) and peptide spectrum matches (B-D) of additional potential neoepitopes - experimentally determined (upper) and synthetic (lower) peptide. The mutated amino acid is highlighted in red. b- (green), y- (brown) and a-ions (blue) are annotated for important fragments.

Computational runtime

The runtime of the workflow was tested by a set of two replicate measurements (PBMC007) on a 3.1 GHz dual core (+HT) Intel Core i7 processor setting 8 GB RAM as maximum memory limit. Spectral batch processing was set to 500 to avoid reaching the memory limit. Subsequently processing of each file took approximately one hour with the comet database search being the most memory and CPU-intensive step followed by the targeted quantification procedure. Parallel execution of the workflow on a 28 core high-performance computing (HPC) node of the de.NBI cloud infrastructure finished processing of three JY replicate measurements in 10 min (Figure A.2). Finally, processing of the entire HLA-Ligand-Atlas dataset (described in Chapter 5) was carried out on an in-house multi-core HPC node, which accounts to 771 and 720 MS runs from 246 HLA-I samples that were processed in 7h 49min (894 CPU-hours) and 240 HLA-II samples in 18h 18min (4756 CPU-hours).

3.4 *Discussion*

In this work the open-source, containerized computational pipeline MHCquant, tailored to automatically process large-scale immunopeptidomics HPLC-MS/MS data sets was presented. We showed that selecting the proteomic search engine Comet combined with the Percolator FDR estimation on the subset of high-confidence or predicted HLA-binding PSMs outperforms other commonly applied search tools and provides the best trade-off between the number of identified peptides and the rate of predicted HLA-binders in our benchmark. Moreover, the identification and quantification performance of MHCquant was validated using orthogonal approaches. At last, the re-analysis of a published melanoma dataset using MHCquant resulted in reproduction of published results and the identification of additional potential neoepitopes that were previously not discovered.

Yet, there are still remaining challenges to be improved upon in the future when analysing immunopeptidomics data. For instance the analysis of non-canonical or cryptic peptides requires the use of very large databases [23, 148]. While this was not tested in this study, most likely MHCquant will share common problems encountered in proteomics when dealing with large search spaces [20]. The subset FDR approach that is an option in MHCquant might help in this regard, but there might be improved methods to achieve similar or better results in the future. In fact, the criterion of how a given FDR threshold for filtering is set itself could be reconsidered, as there might be valid and important peptide spectra for instance mutated neoepitopes recorded in an MS run that are only revealed at a higher FDR threshold. Cost versus benefit tradeoffs might provide a way to do this, however this was not further explored in this work [149].

In addition, recent improvements on the *insilico* prediction of fragmentation spectra intensities will likely contribute to the creation of better search engines in the future [80, 81, 150]. While at this time approaches for prediction using these complex algorithms are very time and computation intensive, once better accessible they are likely to outperform the simple cross correlation scoring function of Comet and those of other common search engines.

Finally, also technological improvements in MS such as ion mobility and data-independent acquisition are likely to improve sensitivity of immunopeptidomics MS measurments but are currently not featured by the MHCquant workflow [151, 152].

However, at this time we figure the MHCquant pipeline is a reliable and sensitive tool to process immunopeptidomics data efficiently and it is provided for free to the science community, with the intention to improve reproducible results for the detection of neoepitopes.

CHAPTER 4

Data-independent acquisition

"Immunopeptidomics using DIA-SWATH mass spectrometry"

This chapter includes partially identical or adapted content with permission from:

DIAproteomics: A multi-functional data analysis pipeline for data-independent acquisition proteomics and peptidomics,

L. Bichmann, S. Gupta, G. Rosenberger, L. Kuchenbecker, T. Sachsenberg, O. Alka, J. Pfeuffer, O. Kohlbacher, H. Röst

bioRxiv 2020.12.08.415844 (2020)

A detailed description of the contributions to the project by coauthors is provided in the Appendix D

4.1 *Introduction*

Recently, data-independent acquisition (DIA) using sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH-MS) has attracted much attention in the field of proteomics due to its ability to overcome shortcomings of the classical data-dependent (DDA) strategy [69, 153, 154]. Moreover, because of its outstanding performance in reproducibility and quantification, DIA is likely to become a state-of-the-art technology in clinical mass spectrometry (MS) [155]. The main advantages are its capacity to (1) acquire fragment spectra in a reproducible grid based fashion over the entire mass and retention time range, (2) sample fragment spectra for nearly all precursor ions present in a sample and (3) enable to trace elution profiles of fragments and integrate their individual quantities at a greater dynamic range [156]. Yet, this comes at the cost of an increased complexity of the acquired mass spectra, due to simultaneous fragmentation of multiple precursor ions [157]. Nonetheless, DIA has the promising potential to achieve a greater identification rate and quantification range, more reproducibility and fewer missing values than DDA.

Consequently, DIA SWATH-MS has been applied in the field of immunopeptidomics [147, 158] as well. However, in order to adapt the DIA methodology from the classic tryptic proteomic experiment for immunopeptidomics, different settings have to be dealt with. Generally, less complex peptide extracts are spiked into the mass spectrometer due to the immunopurification as part of the MHC ligand extraction procedure. In addition, typically a much narrower mass-to-charge (m/z) range is measured and fewer transitions of peptide fragments are observed, as shorter peptides are encountered in particular for MHC class I. Retention time alignment across very heterogeneous MS runs of different MHC composition is hindered by the lack of shared peptide content.

A key step in order to process DIA data is the generation and application of high-quality spectral libraries in order to interpret the complex DIA spectra with more sensitivity. [159] These libraries can be derived from previously acquired DDA measurements in two ways - (1) as a pan master library (ML) uniting multiple samples or (2) as a sample specific library (SSL) from the exact same sample. Ultimately, public repositories such as the SWATHAtlas [160] and SysteMHCAtlas [120] provide collections of previously acquired spectral libraries. In particular MHC allele-specific master libraries provided through the SysteMHC Atlas are promising strategies to process DIA data, as they allow to limit the inflated search space of unspecifically cleaved MHC peptides in DDA immunopeptidomics database search approaches [20, 161]. However, the library should match the instrument and acquisition method settings of the respective DIA experiment in order to be comparable, as different instruments, ionization methods and corresponding parameters such as collision energies produce vastly different fragment spectra pattern.

Previously, it has been attempted to generate a draft map of the murine and human immunopeptidome of various tissues using liquid chromatography coupled to tandem MS (LC-MS/MS) following immunopurification of the respective MHC-bound peptidome [162, 163]. Both studies have employed the classical approach of DDA shotgun MS acquisition. In this chapter the results from an explorative analysis applying DIA SWATH-MS to a large-scale study of the human immunopeptidome of various autopsy body donor tissue samples are described.

For this we employed the open-source software framework OpenMS [17] and specifically the OpenSWATHWorkflow [70] wrapped into a containerized nextflow analysis pipeline as part of the nf-core initiative for reproducible bioinformatics research. In addition, the pipeline makes use of state-of-the art methods such as pyprophet for FDR assessment [164] and dialignR [165] for precise retention time alignment and transfer of identifications across runs. While comparing DDA and DIA processing results in general, we also investigated the difference between the use of SSLs and patient or HLA allele-specific MLs. Ultimately, we demonstrate that using DIA SWATH-MS for immunopeptidomics may increase peptide identifications and reveals similar patterns of immunopeptidome presentation in the same tissues across different individuals.

4.2 *Materials and Methods*

Patient material and sample preparation

The samples were acquired and prepared identically as described in our previous study on the DDA analysis of the HLA-Ligand-Atlas immunopeptidomics ressource (see Chapter 5) [163]. In fact, the same samples used in this study were measured in a pairwise manner acquiring first three DDA and then two DIA replicate measurements directly after each other for each sample. This ensured very similar retention time behaviour and instrumental spectral quality in the respective DDA and DIA measurement pairs. In addition, a dilution series of JY cell immunopeptidome purifications, that were prepared identically as described in Chapter 3, was generated in linear steps from 0.2 to 0.0125 fmol and measured using DIA and DDA.

LC-MS DIA-SWATH Acquisition

DIA methods consisted of tSIM spectra in profile mode at a resolution of 120,000 at m/z 200 with a target value of 1.5x10^5. DIA isolation windows were adjusted to precursor density resulting in different non-overlapping isolation window widths as described in Table C.2. Spectra were acquired in Top20 mode for both tSIM scans, DIA 1 and DIA 2 experiments with 100 ms maximum injection time. Similar to the DDA acquisition method, we limited the precursor mass range to accommodate HLA class I and class II ligands without including contaminating precursor species.

HLA class I ligands were fragmented with CID at a collision energy of 35% and an activation Q of 0.25. The Orbitrap resolution was set to 30,000. The tSIM scan isolation was performed in the quadrupole with an isolation width of m/z 252 centered at m/z 525. Thereby, we obtained isolation windows ranging from m/z 399 – 651. HLA class II ligands were fragmented with HCD with a collision energy set to 30%. The Orbitrap resolution was also set to 30,000. However, the tSIM isolation was centered at m/z 700 with an isolation width of m/z 502 resulting in a mass range of 449 to m/z 951.

Bioinformatic pipeline construction

In order to process the large amounts of immunopeptidomics DIA-SWATH measurements acquired in this project, the bioinformatics pipeline DIAproteomics was constructed. DIAproteomics is an automated analysis pipeline implemented in Nextflow [166] that can be broadly partitioned into the following parts: Optional spectral library and iRT generation from provided DDA data, optional spectral library merging and RT alignment, DIA library search, false discovery rate (FDR) estimation, MS2 chromatogram alignment across runs, and output summary (Figure 1). Each of these parts involves one or more required or optional steps within the workflow (Supplementary Information Table S1 and Figure S1). An experimental design needs to be provided in the form of an input sample sheet specifying DDA and DIA samples, libraries or iRT standards that should be co-processed in one batch. The processing steps of the pipeline are discussed in the following in detail:

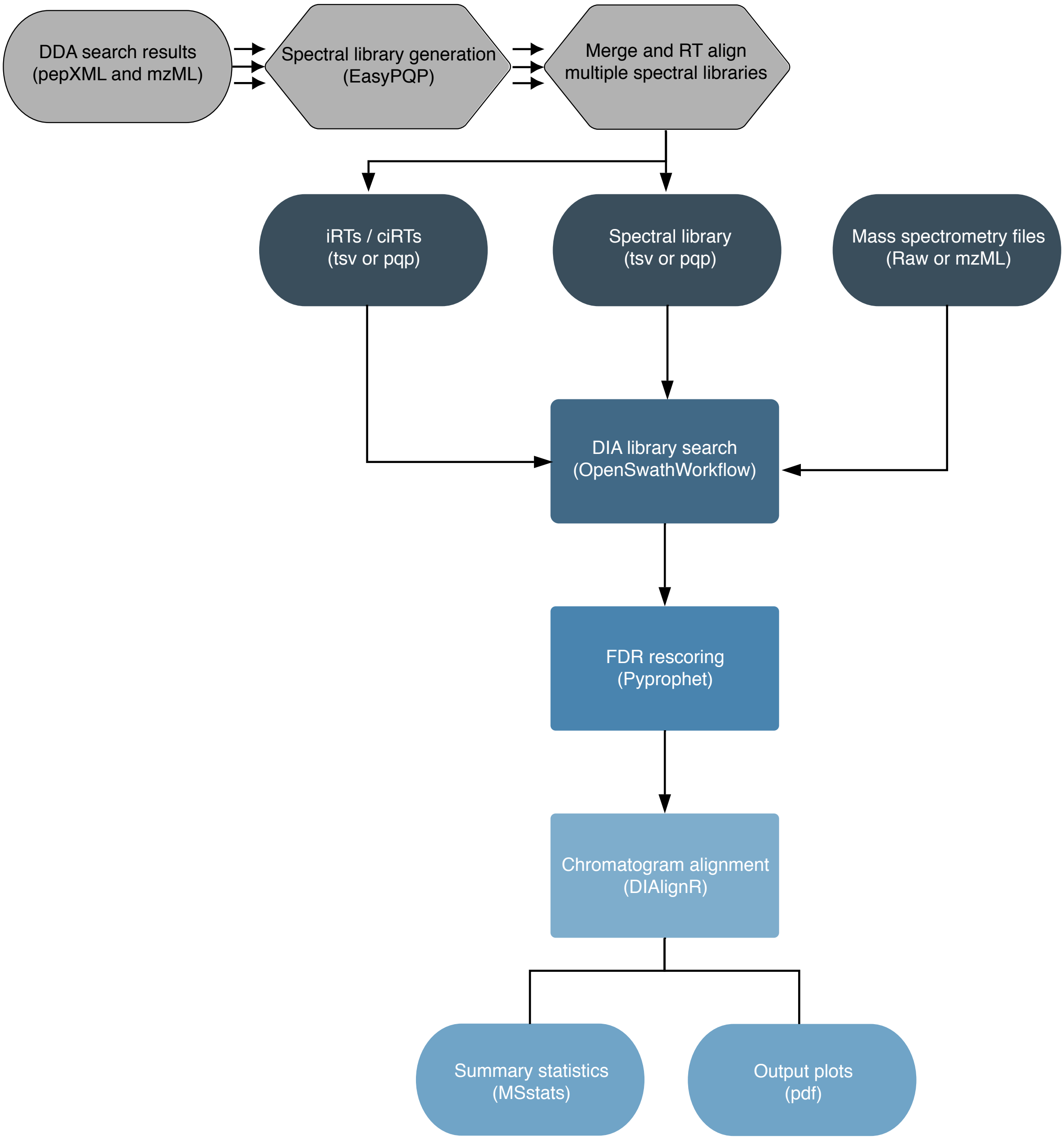


Figure 4.1: Simplified scheme of the DIAproteomics peptide identification workflow: MS raw files (Raw or mzML), internal retention time standards (iRTs / ciRTs) and a spectral library (tsv or pqp) and the corresponding DIA swath window table serve as input. Optionally, the spectral library can be generated from DDA search results (pepXML) and raw files (Raw or mzML) using the EasyPQP software. The library search of the DIA raw data is then carried out using the OpenSwathWorkflow. The false discovery rate (FDR) is calculated based on decoy library hits using Pyprophet. The resulting extracted chromatograms of peptide identifications of all input files are aligned and matched (DIAlignR). Finally, the search results are exported as a summary table (tsv).

PIPELINE INPUT The input to the pipeline is multi-fold and specified in a sample sheet: a set of DIA-SWATH HPLC-MS raw files, a spectral library (in tsv, pqp or TraML format), a set of iRT standards (in tsv, pqp or TraML format). Alternatively, the spectral library and iRTs can be generated from a set of DDA HPLC-MS raw files and corresponding peptide identifications (pepXML format or others).

RAW FILE CONVERSION In a first, optional step provided DDA and DIA raw MS measurements (Thermo Raw vendor format) are converted to the open, XML-based mzML format [127].

SPECTRAL LIBRARY GENERATION If specified the library is generated using EasyPQP (available at https://github.com/grosenberger/easypqp) which matches the provided search results (for example in pepXML format) and the corresponding DDA raw measurements to annotate and store peptide transitions and their properties in a tab-separated table [167]. The library is transformed into an assay containing a specified number of transitions of b- and y-ions falling into a custom mass-to-charge range using the OpenSwathAssayGenerator. Subsequently, decoy transitions that can be generated by OpenMS in multiple ways such as reversed or shuffled are added to the library using the OpenSwathDecoyGenerator. Finally, the generated library will be exported in the peptide query parameter (pqp) sqlite-based data format. Optionally, all steps of the library and decoy generation can be skipped, and an existing library can be used instead.

PSEUDO IRT GENERATION If specified, a given number of highly confident peptide identifications spanning the entire RT range will be selected and exported to serve as iRT-standards in the DIA library search step. This is important, for example, if no iRT-standard kit was spiked into the samples before the DIA measurements. Selected iRTs will be exported in the peptide query parameter (pqp) sqlite-based data format. However, if provided, a set of user-defined iRTs can be used instead.

SPECTRAL LIBRARY MERGING If multiple libraries per sample are provided, for example when stemming from a set of technical replicates, the libraries can be optionally merged and will then undergo a linear RT alignment onto the same reference. When merging is enabled, the best scoring peptide identification is kept in the library omitting a lower scoring duplicate.

SPECTRAL LIBRARY RT ALIGNMENT When RT alignment is enabled, the multiple input spectral libraries will be pairwise aligned onto the same reference. This is achieved by computing a minimum spanning tree connecting all provided libraries by shared peptide overlap. Hence the library having the highest overlap in shared peptides with all other libraries will be the central reference for the other libraries. Importantly, this strategy is also applicable when aligning very distant libraries onto the same reference that share no consensus peptide identifications among all libraries [168]. However, it requires peptides to be shared between all pairs of libraries, resulting in a connected tree.

DIA SPECTRAL LIBRARY SEARCH DIA library search is carried out using the OpenSwathWorkflow, implemented within the OpenMS toolbox. The spectral library and iRT-standards are used to search all input DIA raw measurements individually with a customizable parametrization. The swath windows can be determined from the data. Finally, extracted ion chromatograms (XICs) of the searched peptide transitions (mzML) are exported and the output features and transition properties are stored in OpenSwathWorkflow files (osw).

Step	Requirement	Description	Employed software tool	
1	(optional)	DDA raw file conversion	ThermoRawFileParser	Spectral library and iRT s
2	(optional)	DDA library generation	EasyPQP	
3	(optional)	Assay generation	OpenSwathAssayGenerator	
4	(optional)	Library merging and alignment	Custom Python script using the libraries networkX, scipy	
5	(optional)	Pseudo iRT generation	Custom Python script	
6	(optional)	Decoy generation	OpenSwathDecoyGenerator	
7	(optional)	DIA raw file conversion	ThermoRawFileParser	DIA search
8	required	DIA spectral library search	OpenSwathWorkflow	
9	required	DIA search output merging	Pyprophet	FDR estimation
10	required	Global false discovery rate estimation	Pyprophet	
11	required	Export of scoring results	Pyprophet	
12	required	Chromatogram indexing	OpenMS-FileConverter	Chromatogram alignment
13	required	Chromatogram alignment	DIAlignR	
14	(optional)	Reformatting	Custom Python script	
15	(optional)	Statistical post processing	MSstats	
16	(optional)	Output visualization	Custom R script	

Table 4.1: Detailed description of all steps within the Nextflow implementation of the DIAproteomics workflow version 1.1.0. The steps and their names may vary across versions of DIAproteomics.

false discovery rate estimation The OpenSwathWorkflow output files (osw) are merged sample-wise as defined in the experimental design (sample sheet). The merged file is then scored using the PyProphet target-decoy FDR estimation procedure. Finally, the level of confidence such as local transition- or global peptide or protein level-based can be define [164]. The PyProphet scoring results will then be exported as a tab-separated table per DIA MS run and the results will be visualized in a pdf report.

ms2 chromatogram alignment As the last processing step, the extracted and scored MS2 chromatograms will be aligned using the DIAlignR software. This involves matching chromatograms between runs that can be aligned and integrating their transition areas. The sum of the integrated areas per peptide will be reported as peptide quantities in a TSV file. For this procedure, DIAlignR provides several FDR estimates that can be customized within the workflow to define cut-offs for transitions that should be excluded from matching between runs. This allows to match confidently identified peak groups (FDR < 1%) in one MS run with less confident identifications (FDR > 1%) from other runs [169, 170].

output summarization The output is summarized in a pairwise manner on peptide or protein level using the MSstats post-processing software [171]. In addition, it is possible to export a number of diagnostic plots illustrating peptide and protein identification results, their quantities and properties.

containerization and workflow systems The aforementioned tools are integrated in a pipeline and containerized into an online accessible docker image [104]. Implementation in Nextflow [98] based on the nf-core community template for reproducible bioinformatics workflows [101] allows easy execution on various HPC and compte infrastructures such as google cloud or amazon web services. Moreover support for multiple functionalities is provided such as for various container systems (e.g., docker, singularity, podman) and environment management platforms (e.g. Conda).

Analysis of human tissue immunopeptidomics DIA SWATH data

All spectral libraries were generated from DDA runs deposited in PRIDE as part of the previously analyzed HLA-Ligand-Atlas data set (PXD019643) (see Chapter 5).

dda data processing DDA replicates were processed using the nf-core containerized bioinformatics workflow MHCquant 1.2.6 [123] (see Chapter 3), The nextflow-based workflow implementation comprises tools of the OpenMS toolbox for computational MS applying the database search engine comet and local FDR assessment using Percolator 3.0. MHCquant settings for high-resolution instruments involving a precursor mass tolerance of 5 ppm and a fragment bin tolerance of 0.02 Da were applied. The swissprot database (stand June 2018) and "unspecific cleavage" was set as database and enzymatic restriction setting for the peptide identification procedure. Replicate measurements were co-processed by assessing the FDR on the merged set of identifications and aligning them to a common RT reference.

sample specific spectral libraries (ssl) Generation of the spectral libraries was carried out externally to the pipeline, since at the time of the analysis no containerized version of the respective steps existed within DIAproteomics. Identifications resulting from the DDA data processing were filtered by a 5% peptide level q-value cut off and used for spectral library generation with the SpectraST 5.0 software as part of the Trans-Proteomic-Pipeline. Finally, the generated spectral libraries of replicate runs were merged, keeping the best scoring peptide identification for a set of replicates.

PAN MASTER SPECTRAL LIBRARIES (ML) SSLs were combined into pan patient and pan allele master libraries by applying a pairwise linear RT alignment as no common reference could be constructed across all MS runs of one patient or allele. The libraries were merged by keeping only the best scoring peptide out of each library. As for the pan A*02:01 HLA-allele master library only peptide hits passing the binding affinity prediction threshold were used for generating the library. The pairwise linear retention time alignment was carried out along the edges of a minimum spanning tree linking all samples of a given patient or allele to a central reference (Figure A.8) as described earler. The MST was constructed using functionalities of the Python library networkX [172].

DIA DATA PROCESSING DIA measurements were processed using a developmental version of the nf-core containerized bioinformatics workflow DIAproteomics (state: December 4th, 2019 - `https://www.openms.de/diaproteomics/`). The nextflow-based workflow implementation comprises tools of the OpenSwathWorkflow v.2.4 [70], FDR assessment using pyProphet v.2.1.3 [164] and chromatogram alignment using DIAlignR v. 1.1 [165] (see Figure 4.1 and Figure A.6). The internal retention time alignment between spectral library and DIA run was carried out based on a LOESS alignment using up to 250 high-scoring DDA peptide identifications spanning the entire retention time range. As for the OpenSwathWorkflow, m/z and retention time extraction windows of 20ppm and 600 seconds were chosen respectively. The FDR was assessed on local peptide level applying a q-value threshold of 1%. DIAlignR chromatogram alignment was carried out using an upper FDR limit of 5%, a sampling time of 1.4 and gapQuantile of 0.9, as a set of optimal parameters for high resolution orbitrap data.

Computation of Jaccard coefficients between samples

We investigated the similarity between replicates, samples and tissues by pairwise comparisons of all peptide identification results. When comparing DDA and DIA results the overlap was calculated for common peptide identifications. When comparing tissues across donors only peptide identifications that were retrieved from more than one sample were considered. The Jaccard index was calculated by dividing the set intersection by the set union for all pairwise comparisons:

$$j = \frac{A \cap B}{A \cup B} \tag{4.1}$$

Hierarchical clustering of tissues according to peptide quantities

The matrix of peptide quantities across samples resulting from the DIAlignR output was used for hierarchical clustering approach. Only the four patients (AUT-DN04, -DN12, -DN14 and -DN15) sharing the allele HLA-A*02:01 and among those, only tissues that are shared by at least two individuals were considered for the approach. The clustering was carried out with the Python package seaborn v. 0.8.1 using euclidean distance and average linking as cluster parametrization.

4.3 *Results*

DIA achieves greater reproducibility among replicates than DDA

In order to compare the reproducibility of DIA and DDA we assessed the pairwise set overlap between groups of replicate MS measurements. (Figure 4.2 B-D) DIA results were retrieved applying a sample specific spectral library generated from the respective DDA replicate measurement. In addition, the analysis here was performed without incorporating matching between runs or chromatogram alignment, to see compare the intrinsic capability of both approaches.

On average DIA yields 95% overlap among pairwise comparisons of replicates in contrast to 75% overlap using DDA. The reproducibility of peptide quantities across two replicate MS runs is achieved equally well by DDA and DIA and most peptide quantification results agree well between both approaches (Figure A.7).

When comparing retrieved peptide identifications in a linear concentration series of JY cell immunopeptidome purifications, it is very evident that using DIA it is possible to identify more peptides at low concentrations than with the DDA approach. (Figure 4.2 E) Therefore, the greater reproducibility achieved comes likely due to the greater dynamic range - the fact that DIA is able to all theoretical fragment ions (including low abundance fragments) in a grid-based fashion in contrast to the "shotgun" DDA sampling approach. (Figure 4.2 A).

Pan master spectral libraries increase transfer of peptide identifications among samples

The results of a DIA analysis depend entirely on the spectral library employed, similar to the size and content of the database used in a database search approach for peptide identification in DDA. (Figure 4.2 A) Ideally, a spectral library contains the spectra of all peptides present in the measured sample, in order to identify them correctly by spectral matching. As HLA-presented peptides should be to the largest part similar across tissues and only to a small extent tissue specific (see Chapter 5), we compared the performance of a SSL generated from the respective DDA runs of the sample with a pan patient ML generated from all DDA runs of patient AUT-DN12.

While maintaining a similar percentage of retrieval of the same peptides identified in the DDA runs (ML: 72% and SSL: 77.6%), applying a pan patient ML resulted in an increase of 33% identified peptides. (Figure 4.3 A, C) These additionally discovered peptides correspond to HLA presented peptides found in other tissues of the same patient and do not stem from source proteins that serve a tissue specific role. Moreover, the greatest amount of peptides retrieved using the ML and SSL both (ML: 90.1% and SSL 79.4%) are found within less than 50 seconds from their library retention time. (Figure 4.3 B, D) Hence, applying pan master libraries has retrieved a large fraction of peptides with a global FDR 1% not identified using a "shotgun" DDA approach.

When assessing properties of the additional peptide identifications using the ML in contrast SSL approach, differences in their quantities as well as confidence of identification (p-values)

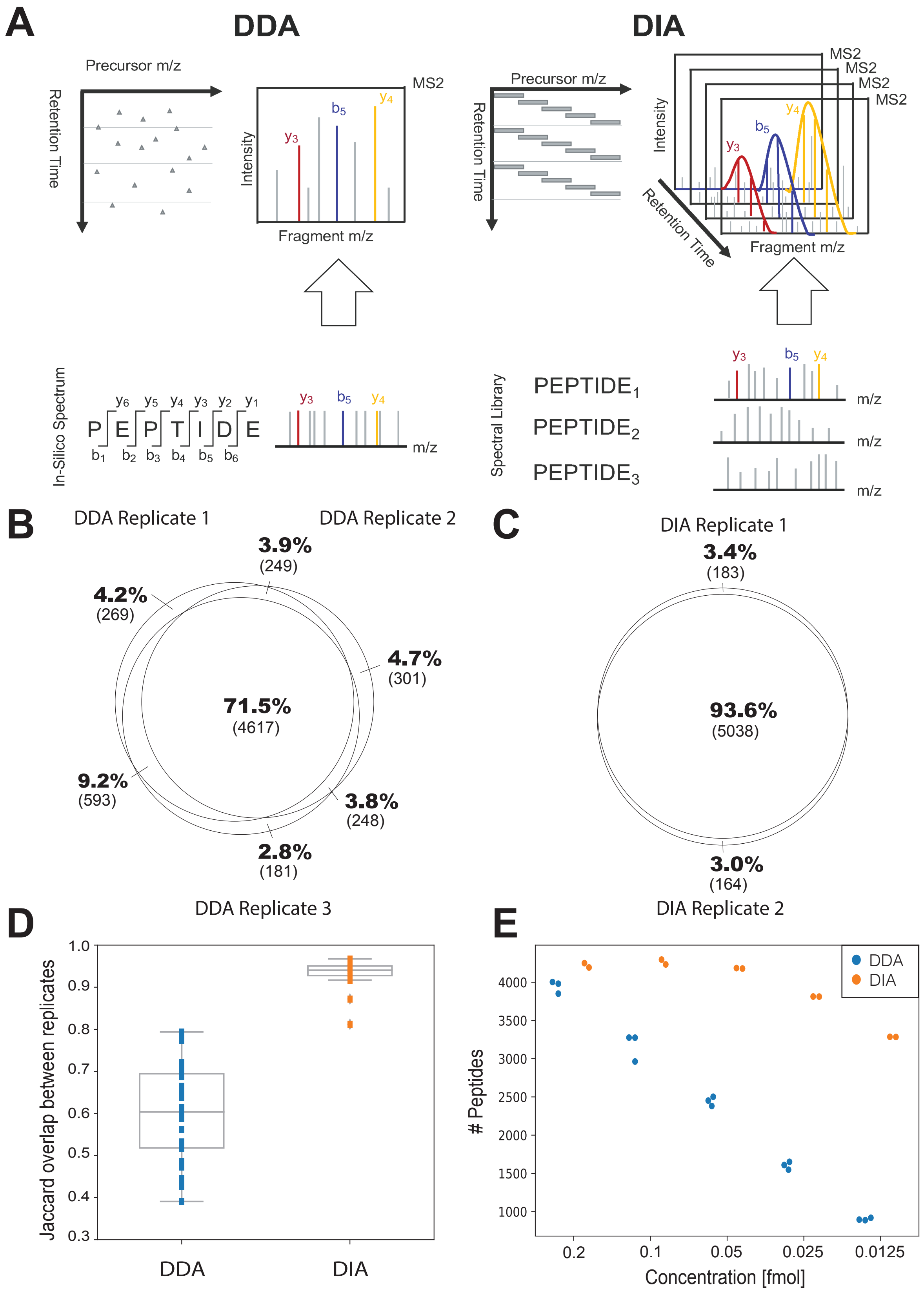


Figure 4.2: A) Overview scheme of the MS data acquisition methods DDA and DIA and the respective different procedure applied for peptide identification. B) Three DDA replicate measurements and the percentage of consensus peptide identifications among them. C) Two DIA replicate measurements and the percentage of consensus peptide identifications among them. D) Multiple pairwise replicate comparisons for DDA and DIA and the Jaccard overlap of peptide identifications among them. E) Comparison of retrieved peptide identifications in a linear concentration series of JY cell immunopeptidome purifications from 0.2 to 0.0125 fmol measured in three DDA and two DIA replicates.

can be found. (1) The additional identifications are distributed at the lower end of the intensity spectrum in contrast to the DDA reconfirming identifications found using the SSL approach (Figure 4.3 E). This is similar to the greater reproducibility in accordance with the fact that the DIA methodology is capable of retrieving less intense peptide ions out of each MS run due to the greater dynamic range. (2) However, the additional peptide identifications have less significant p-values (Figure 4.3 F). It indicates that the peptide-spectrum matches are in general of worse quality, which could be due to low fragment ion intensities or noise.

DIA and DIAlignR increase the Jaccard overlap between different samples

To check whether the increased reproducibility and retrieval of peptide identifications within the DIA results corresponds to a greater overlap across different samples of the same patient, we searched 10 different tissues of the same patient with the same pan patient ML. In addition, the resulting extracted MS2 chromatograms were aligned using the DIAlignR software.

Indeed, the additional peptide identifications yielded from the DIA pan patient ML resulted in a significantly greater share of peptides among all samples than using the classical DDA approach. (Figure 4.4) The manual inspection of the XICs of some of the additional identified peptides in DIA in contrast to DDA confirmed the presence of the transition peak groups in multiple DIA but not all DDA MS runs (Figure B.2). Moreover, in combination with chromatogram alignment using DIAlignR the overlap of peptides shared across samples was increased significantly even further. In particular, the sparsity of the peptide identification matrix across samples was tremendously reduced by 15.1% from 69.1% to 54.2% using DDA and DIAlignR respectively. However, the results also indicate that there is a large biological variety across samples in the HLA-immunopeptidome of different tissues and samples, as the sparsity can't be reduced to less than half of the entire matrix.

Finally, we aimed at comparing the identification increase per sample and relating it to the sample complexity. We observe that the peptide identification increase varies largely across samples and is slightly related to the number of MS1 features detected in a sample. Consequently, a slight trend can be seen that the more complex a sample, the more beneficial spectrum identification using the DIA methodology will be, as MS1 features can give an approximation of the total number of analytes simultaneously present in a given sample.

DIA and DDA reveal tissue similarity of the immunopeptidome across multiple donors

In order to investigate the effect of increased peptide identifications across multiple donors and tissues, a ML for the allele HLA-A*02:01 was generated. The spectra of all peptides predicted to bind HLA-A*02:01 were retrieved from all DDA samples from four patients (AUT-DN04, -DN12, -DN14 and -DN15) carrying this particular allele to build the library. Applying the pan allele ML to all DIA measurements of these patients in combination with the MS2 chromatogram alignment approach using DIAlignR revealed the same effect of increased shared peptide identifications across samples. (Figure 4.5)

In addition, unsupervised hierarchical clustering of the resulting peptide identification matrix yielded multiple clusters of samples of the same or similar tissue across different donors.

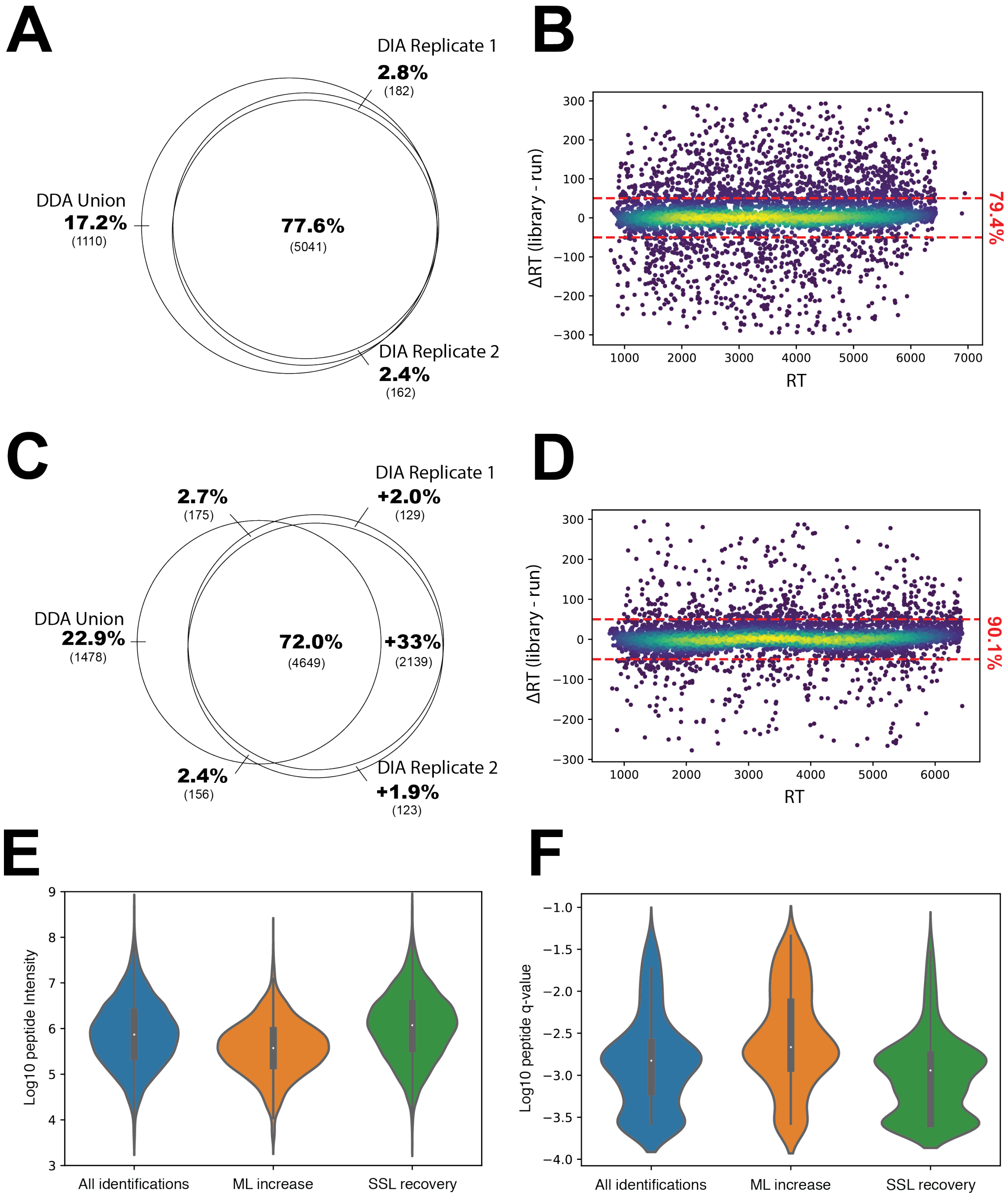


Figure 4.3: Overlap of DDA and DIA peptide identifications using A) a sample specific libray (SSL) or C) a pan master library (ML) of all samples of patient AUT-DN12. The pan master library successfully transfers identifications of other other samples to this sample and thus increases identifications by 33 %. Deviations of the retention times of the DIA MS runs and the corresponding used libary B) SSL and D) ML are shown on the right. Corresponding differences in E) intensity and F) p-value distributions of peptides identified through either both or only one of the approaches (SSL or ML) are depicted.

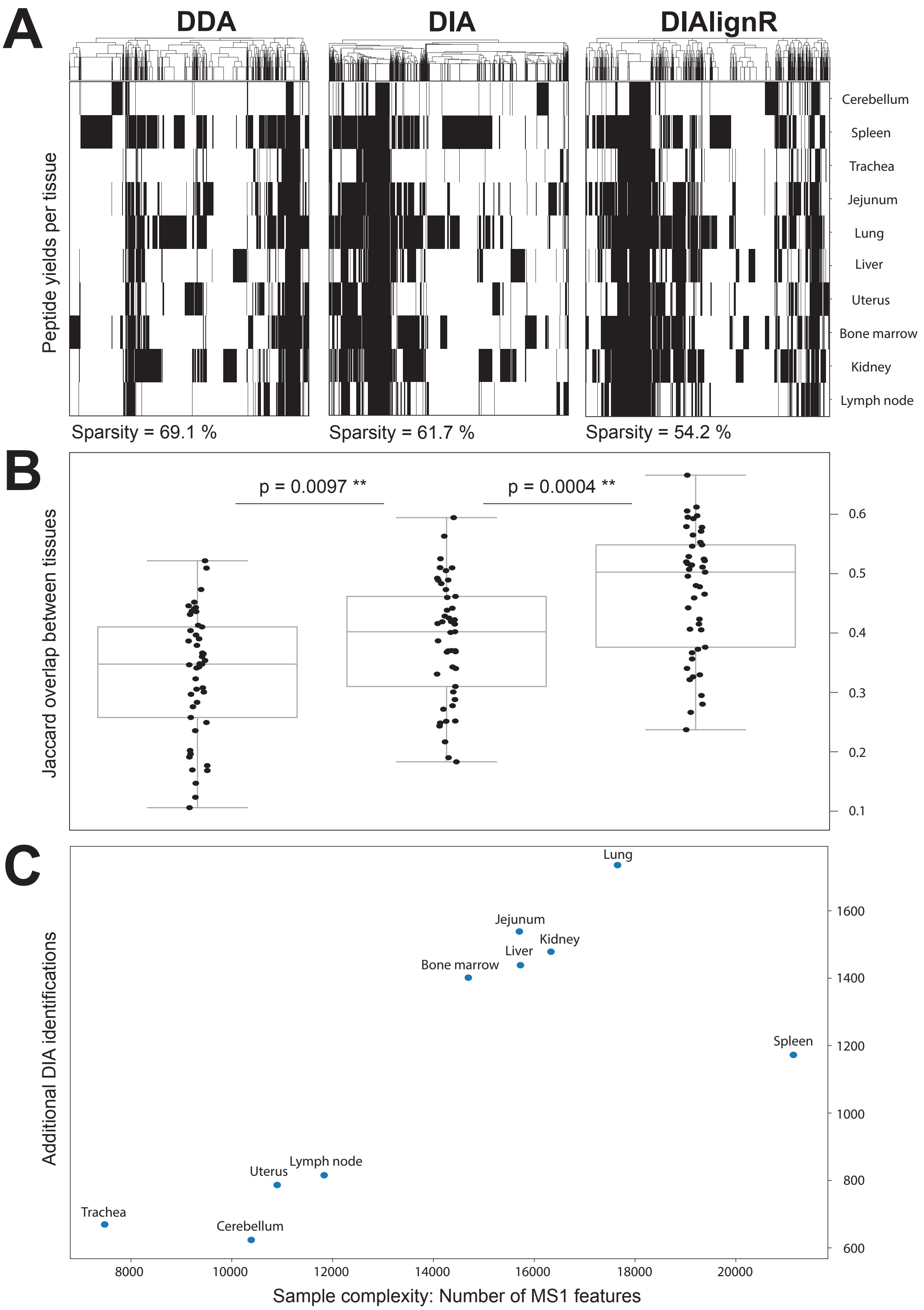


Figure 4.4: Comparison of results from DDA, DIA and DIA in combination with DIAlignR applying a pan patient ML to multiple tissues of patient AUT-DN12. A) Hierarchically clustered heatmaps correspond to present (black) or absent (white) peptide identifications across samples. The sparsity of the entire matrix is indicated at the bottom. B) Boxes indicate pairwise jaccard overlaps of peptide identifications for each method.

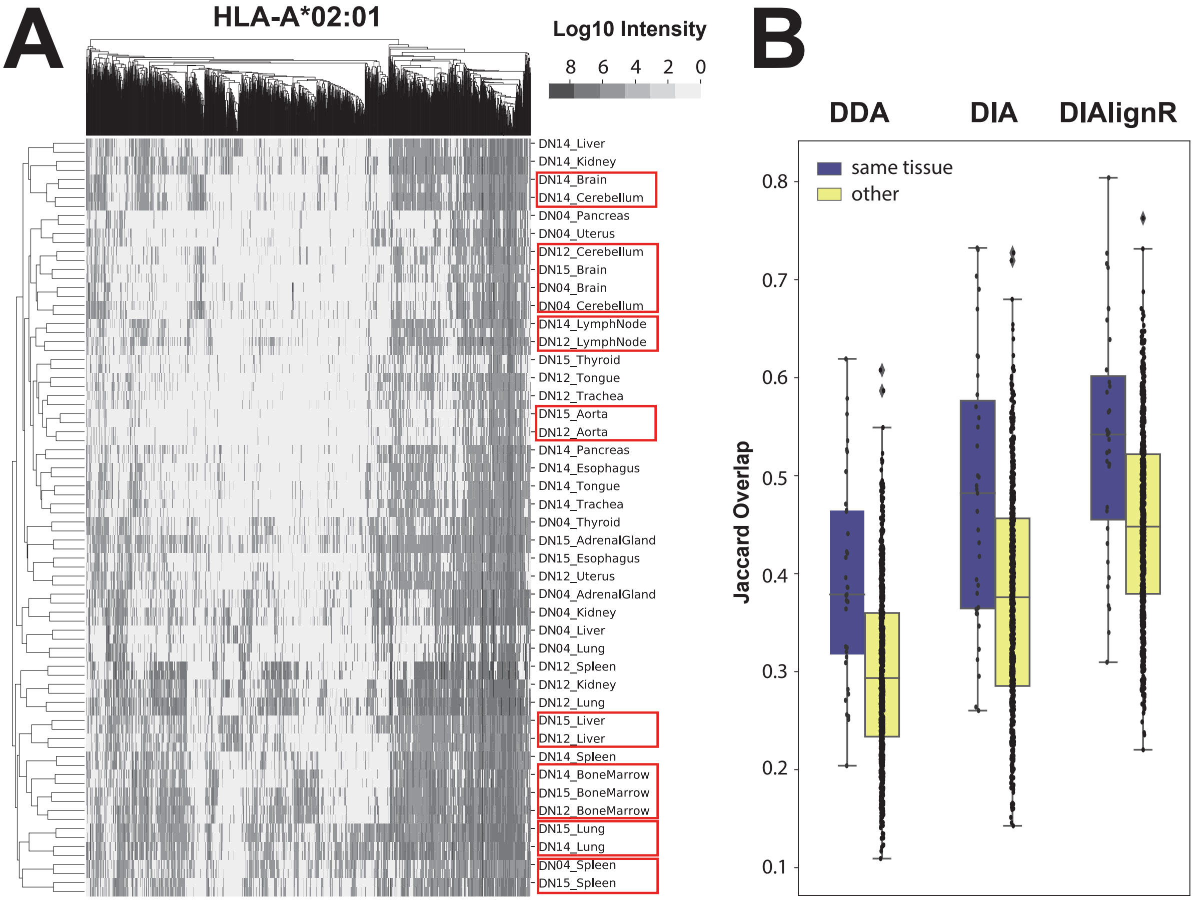


*Figure 4.5: Comparison of peptide identifications resulting from DDA, DIA and DIA in combination with DIAlignR applying a pan allele ML to multiple tissues of different individuals (AUT-DN04, -DN12, -DN14 and -DN15) having the HLA-Allele A*02:01. A) The hierarchically clustered heatmap corresponds to results applying DIA in combination with DIAlignR, that illustrate peptide quantities (grey scale indicates $\log_{10}$ intensity) across all samples. Clusters of the same tissue across individuals are highlighted in red. B) Boxes indicate pairwise Jaccard overlaps of peptide identifications for each method. DIA and DIAlignR result in significantly higher Jaccard overlaps across all samples and same tissues across individuals reveal similar immunopeptidomes.*

When comparing pairwise Jaccard overlaps of all samples, DDA as well as DIA and DIA in combination with DIAlignR yielded higher similarities of the same or similar tissues across donors. These results indicate that there is clear evidence for not only a patient but also a tissue specific bias when analyzing immunopeptidomes.

4.4 *Discussion*

In this study we have investigated the use of cutting edge DIA SWATH-MS in immunopeptidomics in comparison with the standard DDA approach. We have found that DIA leads to greater reproducibility across replicates and the application of MLs can result in a larger number of shared peptide identifications across samples. In particular when moving to clinical settings these two factors - reproducibility and comparability - are crucial and therefore of significant interest to the field. Finally, for the first time, we applied DIA to a large set of human tissue immunopeptidomics samples and were able to demonstrate tissue similarity on immunopeptidome level across multiple donors. In addition, the open-source bioinformatics workflow DIAproteomics that was constructed for this analysis is provided publicly accessible to the science community and is applicable to general proteomic and peptidomic data sets as well.

Having the possibility to fragment all precursor ions, lead to the rediscovery of many more peptides in each DIA run among those previously found in other DDA runs. Moreover, DIA outperformed the DDA database search approach most likely through limiting the large unspecific cleavage search space to a fixed spectral library of interest. In addition, the chromatogram alignment approach DIAlignR [165] was able to increase Jaccard overlaps significantly. Consequently, DIA analysis for immunopeptidomics can be strongly improved by combining it with MS2 chromatogram alignment. In contrast to similar approaches for DDA eg. matching between runs (MBR) [96], DIAlignR can make use of FDR control thresholds computed during the DIA scoring procedure. Hence, alignment across runs provides more confidence in DIA than DDA. In addition, DIAlignR makes use of MS2 chromatograms, which have better signal-to-noise ratios as compared to MS1-chromatograms used by the general MBR approach.

As the sparsity of the peptide identification data matrix still remained at approximately 50%, even when combining DIA with DIAlignR, the variance across samples is intrinsically large in immunopeptidomics. Consequently, even though in our analysis we are able to showcase similarity of tissues on immunopeptidome level it is accompanied by a strong patient and sample specific bias. However, most likely this variance is partially caused by the incomplete sampling of the immunopurification procedure, too.

Ultimately, a major goal in immunopeptidomics would be to tailor personalized therapies to the HLA alleles of a given patient and provide off-the-shelf warehouse peptide vaccine cocktails for various tumor ontologies [44]. The here described results indicate that there might be difficulties achieving this goal, considering the great individual variance of the immunopeptidome across patients and tissues even within a single shared HLA allele. Yet, future research in this area could lead to new insights into which HLA-bound peptides are omni-present in multiple patients or exclusive to particular tissues.

Nevertheless, in agreement with previous studies [147, 158] we conclude that the DIA methodology has key advantages over the DDA approach for immunopeptidomics studies and hope that our provided comprehensive data set will lead to deeper insights in the human immunopeptidome to possibly advance the field of immunotherapy.

CHAPTER 5

The HLA Ligand Atlas

"A database of natural HLA ligands presented on benign tissues"

This chapter includes partially identical or adapted content with permission from:

The HLA Ligand Atlas - A resource of natural HLA ligands presented on benign tissues.

A. Marcu *, L. Bichmann *, L. Kuchenbecker *, D. J. Kowalewski, L. K. Freudenmann, L. Backert, L. Mühlenbruch, A. Szolek, M. Lübke, P. Wagner, T. Engler, S. Matovina, J. Wang, M. Hauri-Hohl, R. Martin, K. Kapolou, J. S. Walz, J. Velz, H. Moch, L. Regli, M. Silginer, M. Weller, M. W. Löffler, F. Erhard, A. Schlosser, O. Kohlbacher, S. Stevanović, H.-G. Rammensee, M. C. Neidert.

bioRxiv, 778944 (2020)

Authors marked with * contributed equally to research work and manuscript

A detailed description of the contributions to the project by coauthors is provided in the Appendix D

5.1 *Introduction*

Advances in biotechnology have lead to recent breakthroughs in biological and medical research such as the sequencing of the human genome (genomics) [173, 174], the entire assessment of human gene expression (transcriptomics) [175] as well as the mapping of the human proteome (proteomics) [176–178]. These discoveries are considered milestones as they enabled to construct for the first time a draft map of all sequences involved at the respective biomolecular layers and lead to an understanding of biological processes at a higher level of complexity. In the context of the immune system, the HLA immunopeptidome represents another independent consecutive layer that has to not been completely mapped so far. Despite the proteome and transcriptome origin of HLA-presented peptides, no quantitative correlation with their precursors on other biologicial levels such as mRNA transcripts and proteins has been achieved [179–181]. Theoretical approaches based on in silico HLA-binding predictions of all possible peptide sequences of a given length in combination with transcriptomics and proteomics data are limited and currently provide the only option to investigate the entire human immunopeptidome [182, 183].

Thus, direct experimental evidence of naturally presented HLA ligands is necessary to possibly gain understanding of the immune system by annotating those that are potential target peptides to T-cells. In particular the field of immunotherapy in the context of cancer treatment benefits from this enhanced insight. Here, approaches have been aiming to identify tumor-specific HLA-presented antigens by comparing benign and malignant immunopeptidomes of diverse cancer entities [46, 184, 185]. Yet, a major impediment often still resides in the lack of a reference of the immunopeptidome of healthy individuals for the comparison [111, 148, 180]. Due to the scarce availability of benign human tissue, common alternative strategies are frequently based on morphologically normal tissue adjacent to the tumor (NAT) as a control reference. However, it has been demonstrated that NATs are suitable only under restrictions, since they may been infiltrated by the disease and have been suggested to represent a rather intermediate state between healthy and malignant tissues, with a pan-cancer-induced inflammatory response [186]. Finally, for some tumor entities such as for example affected regions in the brain, it is surgically impossible to extract NATs without causing permanent damage to a given patient. Moreover, in order to prevent off-target adverse side effects such as in particular severe autoimmunity, it is necessary in all cancer entities to investigate the presence of potential tumor-associated targets (TAAs) on benign tissues when administering immunotherapies to patients [53, 54, 187].

Consequently, this chapter highlights the results of the HLA Ligand Atlas study, throughout which we collected and profiled the immunopeptidome of benign tissues originating from research autopsies. The subjects had not been diagnosed with any malignancy and have deceased of other causes, an approach previously described as a surrogate source of normal tissue [186, 188]. Although these donors cannot be referred to as "healthy" since they may be affected by a range of undiagnosed non-cancerous diseases, we designate their analyzed tissue samples as benign to emphasize morphological normalcy and absence of malignancy. In the same way, the well-stablished Genotype-Tissue Expression Consortium [175, 189] (GTEx) provided an ample resource of transciptome RNA sequencing data of tissues originating from autopsy specimens that have been classified as benign. The tremendous importance of investigating the HLA ligandome in health and disease using mass spectrometry (MS) has been well recognized and outpointed multiple times [14, 15], particularly to improve precision medicine for cancer therapy, [185, 190, 191]. Moreover, insilico HLA-binding prediction algorithms have been shown to be improved by integrating information recovered from MS experiments [33, 40, 192, 193].

The results from the analysis of this dataset might enable us to add to the complexity of our holistic understanding of many biological processes in adaptive immunity and disease. In particular the definition of tumor exclusivity for cancer immunotherapy benefits from the benign human reference, when defining TAAs originating from non-mutated self-peptides [180, 194, 195], MiHAs [196], cryptic peptides [111, 148, 197, 198], and proteasomally-spliced peptides [112, 114, 115, 199]. In addition, the HLA Ligand Atlas hereby offers a further orthogonal level for neoantigen prioritization [22] based on HLA presentation of the native non-mutated counterpart, as suggested by the hotspot hypothesis [185, 200]. Even our understanding of autoimmunity and response to infectious diseases can be advanced by furthering our knowledge of the tissue-wide healthy-state immunopeptidome.

Multiple public repositories maintaining immune-related data have been established, such as the IEDB [201], SysteMHC [120], the 10,000 immunomes project [202], and the DC cell atlas [203]. The results of this work are provided publicly in an user-friendly web interface at `https://HLA-Ligand-Atlas.org` and we envision that the scientific community will integrate this new resource with the existing databases to improve the holistic understanding of the HLA-ligandome and immunology in general.

5.2 *Materials and Methods*

Experimental model and subject details

Human tissue samples were obtained post mortem during autopsy performed for medical reasons at the University Hospital Zürich. The study was approved by the Cantonal Ethics Committee Zürich (KEK) (BASEC-Nr. Req-2016-00604). For none of the included patients a refusal of post mortem contribution to medical research was documented and study procedures are in accordance with applicable Swiss law for research on humans (Bundesgesetz über die Forschung am Menschen, Art. 38). In addition, the study protocol was reviewed by the ethics committee at the University of Tübingen and received a favourable assessment without any objections to the study conduct (Project Nr. 364-2017BO2).

None of the subjects included in this study was diagnosed with any malignant disease. Tissue samples were collected during autopsy, which was performed within 72 hours after death. Tissue organ annotation was performed by a board-certified pathologist. Tissue samples were immediately snap-frozen in liquid nitrogen. Thymus samples were obtained from the University Children Hospital Zürich, Switzerland. Thymus tissue was removed during heart surgery or for other medical reasons. Tissue samples from residual material not required for diagnostic or other medical purposes were obtained after informed consent from the parents of the respective children, in accordance with the principles of the Declaration of Helsinki. The study was approved by the Cantonal Ethics Committee Zürich (KEK) (EC-Nr. 2014-0699, PB-2017-00631) on February 27th 2015.

Furthermore, two benign ovarian tissue samples were collected for the project (OVA-DN278 and OVA-DN281). Both patients were post-menopausal and had a bilateral ovarectomy for cystadenofibromas, which were diagnosed by histopathologic examination of the specimen. The samples were obtained from a normal part of the ovary. The study was approved by the ethical committee of the University of Tuebingen (354-2011BO2).

HLA typing

Multiple HLA typing approaches were performed for the different sources of patient material. Autopsy subject AUT-DN08, AUT-DN16, and two benign ovary samples (OVA-DN278 and OVA-DN281) were typed at the Department of Transfusion Medicine of the University Hospital of Tübingen.

High-resolution HLA typing was performed by next-generation sequencing on a GS Junior Sequencer using the GS GType HLA Primer Sets (both Roche, Basel, Switzerland). HLA typing was successful for HLA-A, -B, and -C alleles. However, HLA-II typing was only reliable for the HLA-DR locus, and incomplete for the HLA-DP and -DQ loci. Therefore, we performed exome sequencing of lung tissue for remaining autopsy subjects. Exome sequencing data was processed and OptiType was employed to extract HLA-I and an in-house modified development version was used to type HLA-II alleles carried out by A. Szolek.

Finally, sequence-based typing was performed for the five thymus samples by sequencing exons 1-8 for HLA-I alleles and exons 2-6 for HLA-II alleles (Histogenetics, Ossining, NY).

HLA immunoaffinity purification

HLA-I and -II molecules were isolated from snap-frozen tissue using standard immunoaffinity chromatography. The antibodies employed were the pan-HLA-I-specific antibody W6/32 [204], and the HLA-DR-specific antibody L243 [126], produced in house (University of Tübingen, Department of Immunology) from HB-95, and HB-55 cells (ATCC, Manassas, VA) respectively. Furthermore, the pan-HLA-II-specific antibody Tü39 was employed and produced in house from a hybridoma clone as previously described [205]. The antibodies were cross-linked to CNBr-activated sepharose (Sigma-Aldrich, St. Louis, MO) at a ratio of 40 mg sepharose to 1 mg antibody for 1 g tissue with 0.5 M NaCl, 0.1 M $NaHCO_3$ at pH 8.3. Free activated CNBr reaction sites were blocked with 0.2 M glycine.

For the purification of HLA-peptide complexes, tissue was minced with a scalpel and further homogenized with the Potter-Elvehjem instrument (VWR, Darmstadt, Germany). The amount of tissue employed for each purification is documented in Supplementary Table 1. This information is not available for seven tissues, annotated as n.d. in said table. Tissue homogenization was performed in lysis buffer consisting of CHAPS (Panreac AppliChem, Darmstadt, Germany), and one cOmpleteTM protease inhibitor cocktail tablet (Roche) in PBS. Thereafter, the lysate was sonicated and cleared by centrifugation for 45 min at 4,000 rpm, interspaced by 1 h incubation periods on a shaker at 4°C. Lysates were further cleared by sterile filtration employing a 5 µm filter unit (Merck Millipore, Darmstadt, Germany). The first column contained 1 mg of W6/32 antibody coupled to sepharose, whereas the second column contained equal amounts of Tü39 and L243 antibody coupled to sepharose. Finally, the lysates were passed through two columns cyclically overnight at 4°C. Affinity columns were then washed for 30 min with PBS and for 1 h with water. Elution of peptides was achieved by incubating four times successively with 100 – 200 µl 0.2% TFA on a shaker. All eluted fractions were subsequently pooled. Peptides were separated from the HLA molecule remnants by ultrafiltration employing 3 kDa and 10 kDa Amicon filter units (Merck Millipore) for HLA-I and HLA-II, respectively. The eluate volume was then reduced to approximately 50 µl by lyophilization or vacuum centrifugation. Finally, the reduced peptide solution was purified five times using ZipTip Pipette Tips with C18 resin and 0.6 µl bed volume (Merck,) and eluted in 32.5% ACN/0.2% TFA. The purified peptide solution was concentrated by vacuum centrifugation and supplemented with 1% ACN/0.05% TFA and stored at -80°C until LC-MS/MS analysis.

Time Series Experiments

We performed time series experiments to assess the suitability of tissues obtained from autopsies as a source of human organs for the characterization of the benign immunopeptidome. We evaluated the degradation profile of the immunopeptidome, when tissues were stored at 4°C for up to 72 h after tissue removal, to mimic the conditions at autopsy. The time series experiment was repeated in three benign tissues from different individuals: one benign liver obtained at autopsy (AUT-DN16 Liver), and two benign ovaries removed surgically (OVA-DN278 and OVA-DN281). The tissues were extracted and incubated at 4°C until a certain time point and flash-frozen in liquid nitrogen until HLA ligand extraction. As more tissue was available form AUT-DN16 Liver, tissue samples were frozen after 8 h, 16 h, 24 h, 48 h, and 72 h. Due to the limited sample amount obtained from OVA-DN278 and OVA-DN281, only three time points could be accounted for: 0 h, 24 h, and 72 h. The HLA immunoaffinity purification was performed as mentioned, with the exception that mass-to-volume ratio in ovary samples was adjusted to the lowest mass across all time points before loading onto sepharose columns.

Mass spectrometric data acquisition

HLA ligand characterization was performed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with a Nanospray FlexTM Ion Source (Thermo Fisher Scientific) coupled to an Ultimate 3000 RSLC Nano UHPLC System (Thermo Fisher Scientific). Peptide samples were loaded with 1% ACN/ 0.05% TFA on a 75 µm x 2 cm Acclaim™ PepMap^{TM} 100 C18 Nanotrap column (Thermo Fisher Scientific) at a flow rate of 4 µl/min for 10 min. Separation was performed on a 50 µm x 25 cm PepMap RSLC C18 (Thermo Fisher Scientific) column, with a particle size of 2 µm. Samples were eluted with a linear gradient from 3% to 40% solvent B (80% ACN, 0.15% FA in water) at a flow rate of 0.3 µl/min over 90 min. The column was subsequently washed by increasing to 95% B within 1 min, and maintaining the gradient for 5 min, followed by reduction to 3% B and equilibration for 23 min.

Data acquisition was performed as technical triplicates in data-dependent mode, with customized top speed (3 s) methods for HLA-I- and HLA-II-eluted peptides. HLA-I peptides have a length distribution ranging mostly between 8 - 12 amino acids, therefore, the scan range was restricted to 400 - 650 m/z and charge states of 2 - 3. MS1 and MS2 spectra were detected in the Orbitrap with a resolution of 120,000 and 30,000 respectively. Furthermore, we set the automatic gain control (AGC) targets to 1.5x10^5 and 7.0x10^4 and the maximum injection time to 50 ms and 150 ms for MS1 and MS2, respectively. The dynamic exclusion was set to 7 s. Peptides were fragmented with collision-induced dissociation (CID) while the collision energy was set to 35%.

HLA-II peptides have a length distribution in the area of 8 - 25 amino acids , thus the scan range was set to 400 -1,000 m/z and the charge states were restricted to 2 - 5. Readout for both MS1 and MS2 were performed in the Orbitrap with the same resolution and maximum injection times as for HLA-I peptides. The dynamic exclusion was set to 10 s and AGC values employed were 5.0x10^5 and 7.0x10^4 for MS1 and MS2, respectively. Higher-energy collisional dissociation (HCD) fragmentation with 30% collision energy was employed for HLA-II peptides.

Database search with MHCquant

MS data obtained from HLA ligand extracts was analyzed using the nf-core (Ewels et al., 2020) containerized, computational pipeline MHCquant [123] (release 1.5.1 - `https://www.openms.de/mhcquant/`) with default settings. The workflow comprises tools to analyze LC-MS/MS data of the open-source software library OpenMS (2.5) [17]. Identification and post-scoring were performed using the OpenMS adapters to Comet 2016.01 rev. 3 [82] and Percolator

3.4 [88] at a local peptide level false discovery rate (FDR) threshold of 1% among the technical replicates per sample. Subsequently, we estimated the global FDR by dividing the sum of expected false positive identifications from each sample (1% peptide level FDR) by the total number of identified peptides in the entire dataset (HLA-I: 4.5% FDR, HLA-II: 3.9% FDR) [91, 206]. The human reference proteome (Swiss-Prot, Proteome ID UP000005640, 20,416 protein sequences) was used as a database reference. Database search was performed without enzymatic restriction, with methionine oxidation as the only variable modification. MHCquant settings for high-resolution instruments involving a precursor mass tolerance of 5 ppm and a fragment bin tolerance of 0.02 Da were applied. The peptide length restriction, digest mass and charge state range were set to 8-12 amino acids, 800-2500 Da and 2-3 for HLA-I and 8-25 amino acids, 800-5000 Da and 2-5 for HLA-II, respectively.

HLA binding prediction

Peptide binding predictions were computed based on the subject's HLA alleles. For HLA-I ligand extracts, we employed SYFPEITHI [32] and NetMHCpan-4.0 [38] in ligand mode (default). The SYFPEITHI score was computed by dividing the sum of amino acid-specific values for each position in the tested peptide by the maximally attainable score for the respective HLA allotype [207]. HLA-II ligand extracts were annotated with NetMHCIIpan-4.0 [38], MixMHC2pred [40] in ligand (default) mode and SYFPEITHI.

Peptides were categorized as strong binders against a given HLA allotype if either netMHCpan-4.0, netMHCIIpan-4.0 or MixMHC2pred reported a percentile-rank score $\leq$ 0.5. Peptides that failed this criterion defined by a reported as weak binders if any of the tools reported a percentile-rank score $\leq$ 2.0. All peptide-HLA allotype associations within these limits were included in the dataset, i.e. a single peptide sequence can be reported as a binder against multile allotypes of the same donor. Unless allele associations are specified, all peptides including those that were not classified as binders against any subject's allotype were included in the analysis.

Quality control thresholds based on binding prediction and length distribution

We defined the fraction of predicted binders of a sample as the ratio of predicted binders divided by the total number of peptide identifications. Technical replicates with a fraction of predicted binders lower than 50% for HLA-I and lower than 10% for HLA-II ligand extracts were excluded from the data set. Furthermore, individual replicates were removed from the data set if the mode of the length distribution differed from nine amino acids for HLA-I and was not in the interval 12-18 for HLA-II (Figure 5.1).

Quantitative time series analysis

Database search of LC-MS/MS data from the three time series experiments was performed with MHCquant 1.5.1 as previously described (Bichmann et al., 2019). Identifications were matched between runs (Tyanova et al., 2016) based on retention time alignment and targeted feature extraction (Weisser and Choudhary, 2017) to integrate respective MS1 areas for all time

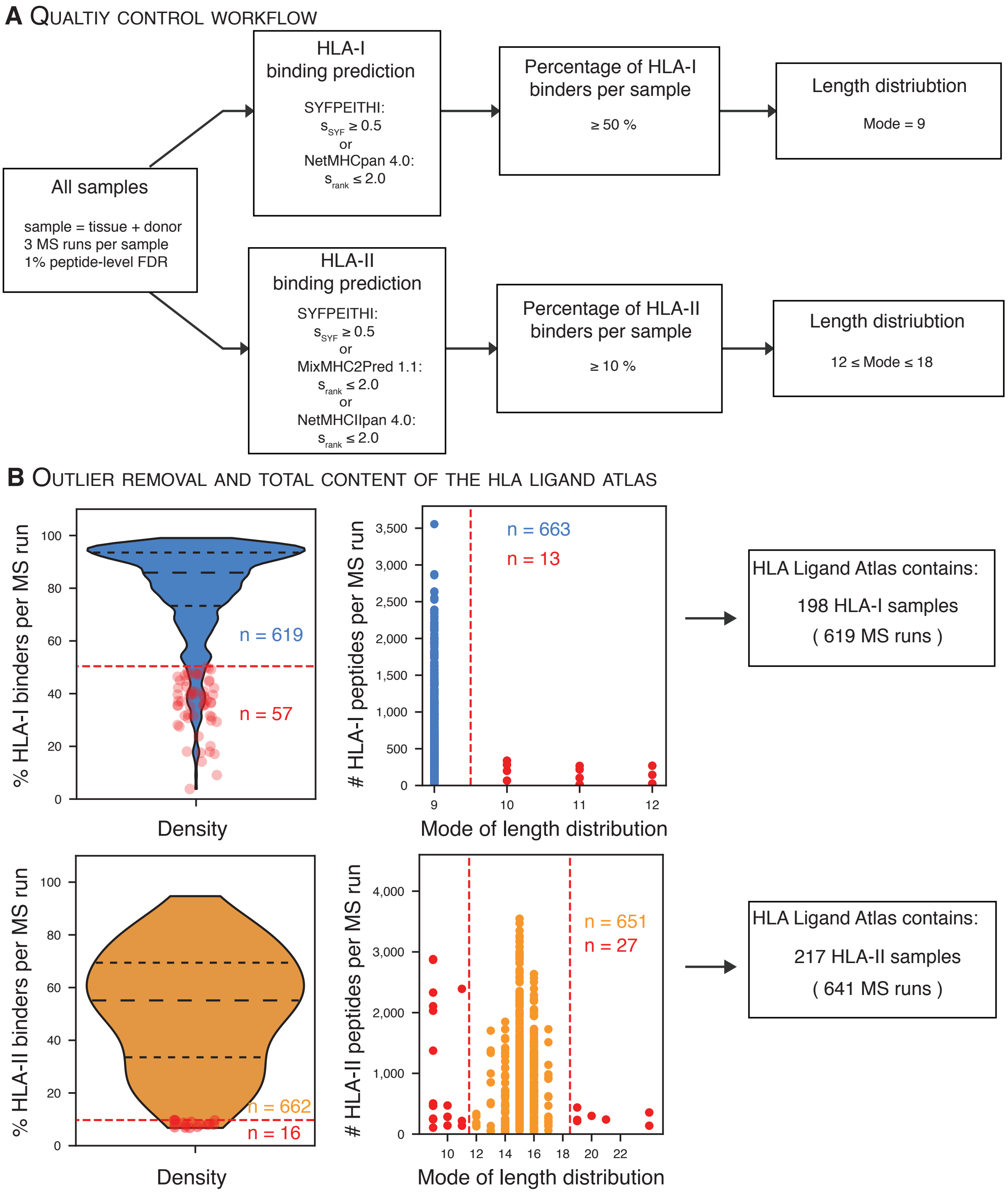


Figure 5.1: Overview of the quality control workflow: A) Data from three LC-MS/MS runs (technical replicates) per sample (one tissue from one subject) were processed with MHCquant with a local peptide-level FDR of 1%. Identified peptides were categorized into peptides predicted as strong, weak, or non-binders. Samples were filtered, employing the specified binding prediction and peptide length mode cut-off thresholds respectively. B) Violin plots (left) depict the percentage of peptides predicted to be HLA-binders per LC-MS/MS run and the quality control cut-off for LC-MS/MS runs employed for HLA-I and -II immunopeptidomes. Dot plots (center) depict the mode of the peptide length distribution encountered per LC-MS/MS run and the quality control cut-off employed for HLA-I and -II. The number of LC-MS/MS runs (HLA-I - blue and HLA-II orange) failing the QC thresholds is indicated by red dots.

points and technical replicates. MS1 areas (x) were normalized to z-scores (standard scores) per MS run by subtracting the mean and dividing by the standard deviation:

$$z = \frac{x - \mu}{\sigma} \tag{5.1}$$

The trajectory of scaled MS1 areas was clustered by k-means unsupervised clustering with 6 seeds using the tslearn (v.0.3.1) Python package. All trajectories are related to the first time point by subtracting its median z-score from all other timepoints in the respective analysis.

Comparison of the HLA-Ligand-Atlas data base with IEDB and SysteMHC

All peptides contained in the HLA Ligand Atlas database were compared with peptides listed in the IEDB and SysteMHC databases for HLA-I and HLA-II ligands separately. The list of peptides stored in the IEDB was obtained by downloading the file "epitope_full_v3.zip" from the "Database Export" page. The obtained table was subsequently filtered for positive MS assays, linear peptides and human origin. Peptides with modifications were removed. Peptides stored in the SysteMHC database were obtained by downloading the file "180409_master_final.tgz" from "Builds_for_download" page. The obtained table was subsequently filtered for human as organism.

Principal component analysis of HLA-I and -II source proteins

To assess the divergence between source proteins of HLA-I and -II peptides, a principal component analysis (PCA) of HLA-I and -II samples was computed. For this a binary matrix was constructed spanning all samples and proteins covered by any HLA-I or -II peptide and annotating its presence (true) or absence (false) in a given sample. The first two principal components of the matrix were then computed using the scikit-learn Python package (v 0.23).

Gene ontology (GO)-term enrichment

GO term enrichment analyses were performed with the Panther 15.0 database (Released 2020-02-21) with the integrated "statistical overrepresentation test" (Release 2019-07-11). Gene identifiers of proteins presented exclusively by either HLA-I or -II allotypes were queried against the "GO cellular component complete" database using the default "Homo sapiens genes" reference list. GO terms were sorted by Fisher's exact raw p-value, and top 10 scoring terms reported as overrepresented and their corresponding p-values were selected for illustration.

Tissue-specific source proteins were defined as HLA-I or -II source proteins identified exclusively in one tissue (Table S5). Gene identifiers of tissue-specific HLA-I and -II source proteins were queried against the "GO biological process complete" database, with the only difference that only the top 5 scoring terms reported as overrepresented were selected for illustration.

Tissue-specific gene set enrichment

Analogously to the GO-term enrichment, tissue-specific HLA-I and -II gene identifiers were separately queried against the GTEx database for gene set enrichment analysis. Gene sets with upregulated gene expression profiles per tissue "GTEx tissue sample gene expression profiles up" were retrieved using the gseapy implementation (v.0.9.15, 2019-08-07) through the enrichr API. All tissues covered in the HLA Ligand Atlas were matched and compared against all tissues in the GTEx database that co-occur in the HLA Ligand Atlas. Fisher's exact raw p-values for the enrichment were computed for each pairwise comparison.

HLA-I peptide yield correlation to expression of immune-related genes

We computed a linear model to compare the median HLA-I peptide yields per tissue with gene expression values (RPKM) of the following genes involved in the HLA-I presentation pathway: HLA-A, HLA-B, HLA-C, immunoproteasome, constitutive proteasome, TAP1, and TAP2. Median HLA-II peptide yields per tissue were correlated to genes involved in the HLA-II presentation pathway: HLA-DRB1, HLA-DRA, HLA-DQB1, HLA-DQA1, HLA-DPB1, HLA-DPA1. The corresponding gene expression values were taken from a previously published study [182].

An ordinary least squares linear model correlating gene expression and log10 median HLA-I/-II peptide yields was computed using R (v.3.5) and the corresponding stats (v.3.5) package reporting R^2, F-statistic p-value, and spearman rho. The cross correlation between all immune related genes and their individual linear models were computed using R (v.3.5) and the corresponding packages corrplot (v. 0.84) and ggplot2 (v.3.2.1). As the expression levels of the investigated genes are highly covariant, the regression would be overfitting when correlating peptide yields to multiple genes involved in the antigen presentation pathway, thus the analysis was limited to a single gene at a time.

Computation of Jaccard coefficients between samples

We investigated the similarity between organs and subjects by pairwise comparison of all samples in the HLA Ligand Atlas. Comparisons were performed both on HLA-I and -II peptide level and HLA-I and -II source protein level. The Jaccard index was calculated by dividing the set intersection by the set union for all pairwise comparisons:

$$j = \frac{A \cap B}{A \cup B} \tag{5.2}$$

Identification of cryptic peptides with Peptide-PRISM

Identification of cryptic HLA-I peptides from MS data was performed as recently described in detail [23]. Briefly, *de novo* sequencing was performed with PEAKS Studio X ([136]) (Bioinformatics Solutions Inc., Canada). Top10 sequence candidates were exported for each fragment ion spectrum. Database matching of all sequence candidates and stratified FDR-filtering was performed with Peptide-PRISM using the 6-frame translation of the human genome (GRCh38) and the 3-frame translation of the human transcriptome (Ensembl 90). Matched peptides were filtered to 10% FDR and peptides were predicted as binder to the corresponding HLA alleles by NetMHCpan-4.0 [37].

Retention time model for cryptic peptide validation

Retention time predictions were carried out using the OpenMS (2.5.0) RTModel based on the oligo-kernel ν-support vector regression (ν =0.5, p =0.1, c =1, degree =1, border_length =22, kmer_length =1, Σ=5) [141–143]. The model was trained on all peptide identifications of canonical peptides identified with MHCquant and applied to all cryptic peptide identifications resulting from peptide-PRISM. Predictions were evaluated by applying a linear least square fit to compute the 99% prediction interval around the predicted versus measured retention times using the statsmodels (v.0.11) function wls_prediction_std.

Synthesis of isotope labeled peptides

Peptides were synthesized using the Liberty Blue Automated Peptide Synthesizer (CEM) following the standard 9-fluorenylmethyl-oxycarbonyl/tert-butyl strategy. After removal from the resin by treatment with trifluoroacetic acid/triisopropylsilane/water (95/2.5/2.5 by vol.) for 1 h, peptides were precipitated from diethyl ether, washed three times with diethyl ether and resuspended in water prior to lyophilization. Purity and identity of the synthesis products were determined by C18-HPLC (Thermo Fisher Scientific, Darmstadt, Germany) and LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific), respectively. A table listing the exact peptide sequences and their heavy isotope labels is contained in the original publication of Marcu, Bichmann and Kuchenbecker et al. [208] in the corresponding supplementary material tables S3

Spectrum validation

We selected 36 cryptic peptides, identified with 1% FDR for spectral validation with isotope-labelled synthetic peptides. Selected peptides were strong binders to the corresponding HLA alleles of the respective subject, with a netMHCpan-4.0 binding % rank <0.5. Isotope-labeled synthetic peptides were spiked into a sample matrix of native HLA eluted peptides from a JY cell line at a concentration of 20 fmol/µl, with the purpose of showing spectrum identity between the native and synthetic peptides. The spectral similarity was computed analogous to the normalized spectral contrast angle [209] between eluted peptide spectra and synthetic isotope-labelled peptide spectra:

$$\lambda(S_1, S_2) = 1 - \frac{2cos^{-1}(S_1 \cdot S_2)}{\pi} \tag{5.3}$$

where the spectra were encoded as intensity vectors (S1 and S2) based on their theoretical b- and y-fragment ions by using the mzR (v2.16.2), msdata (v0.20.0) and protViz (v0.4) R packages. Intensities of matching y- and b-ion pairs as encoded in the intensity vectors were compared, thereby avoiding the necessity to correct for the mass shift caused by the isotope.

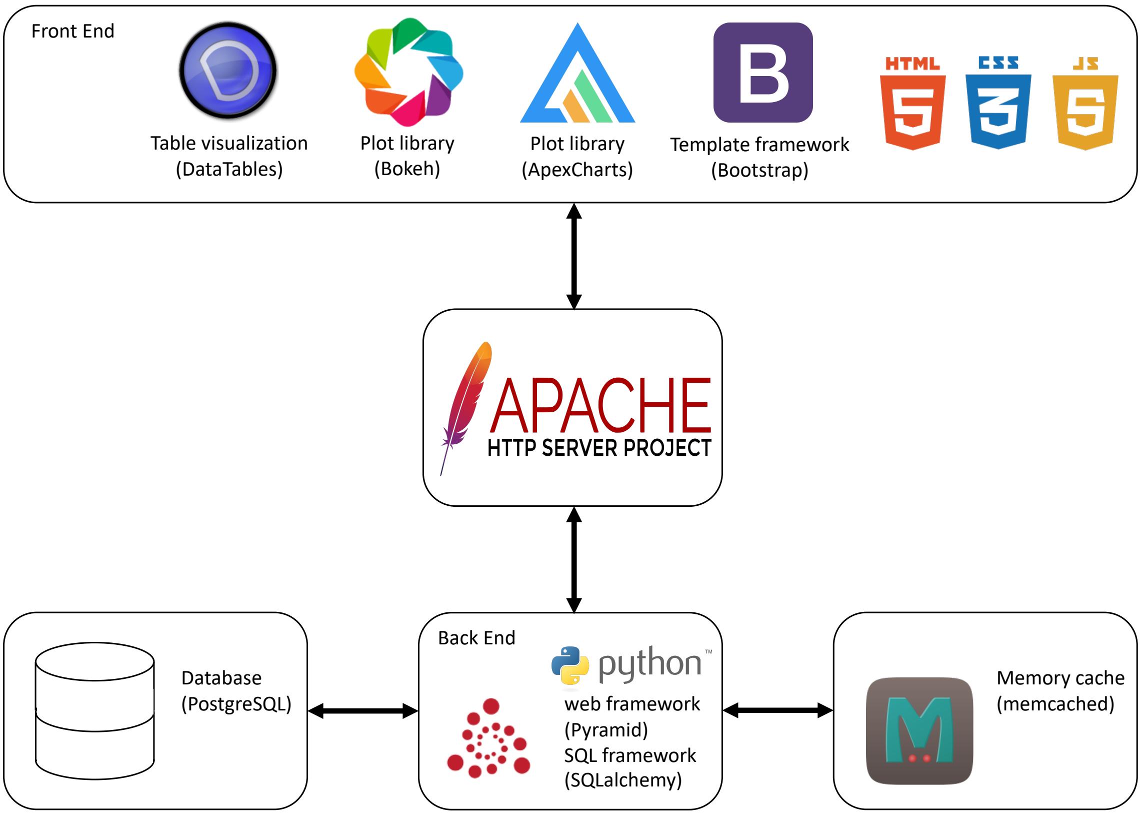


Figure 5.2: Architecture of the public webserver: The webserver is build on a Python back end using the libraries pyramid and sqlalchemy that accesses a SQL database storing all information on identified HLA peptides their respective meta-information. The front end was build using the HTML, CSS and JavaScript template framework Bootstrap and multiple javascript libraries eg. DataTables, Bokeh and ApexCharts. The Figure was adapted from a version drafted by Leon Kuchenbecker who developed the web application.

In order to compare the score distribution of the selected cryptic peptide identifications with a negative control distribution, the same score was additionally computed for the comparison of the same cryptic spectra paired with 1000 randomly sampled spectra containing fragment peaks matching to the same peptide sequence label. Peaks present in at least one of the spectra were considered for the cross product ($S1 \cdot S2$). Intensities of missing peaks in the one spectrum compared to the other were set to zero.

Data storage and web interface

Data was stored and managed using the biomedical data-management platform qPortal [19] (Project ID: MNF_ELK_QHIPP). HLA-I and -II peptides were complemented with their tissue and HLA allotype association and stored in an SQL database. The database is versionized and all data analysis was based on the release 2020.06. A visualization of all tables in an entity relationship database scheme of the SQL database is shown in Figure 5.3.

In addition, a public web server was implemented that allows users to formulate queries against the database, visualize results and allows data export for further analysis. A scheme illustrating the architecture and implementation of the webserver is shown in Figure 5.2. The web front-end was implemented in HTML, CSS and JavaScript based on the front-end

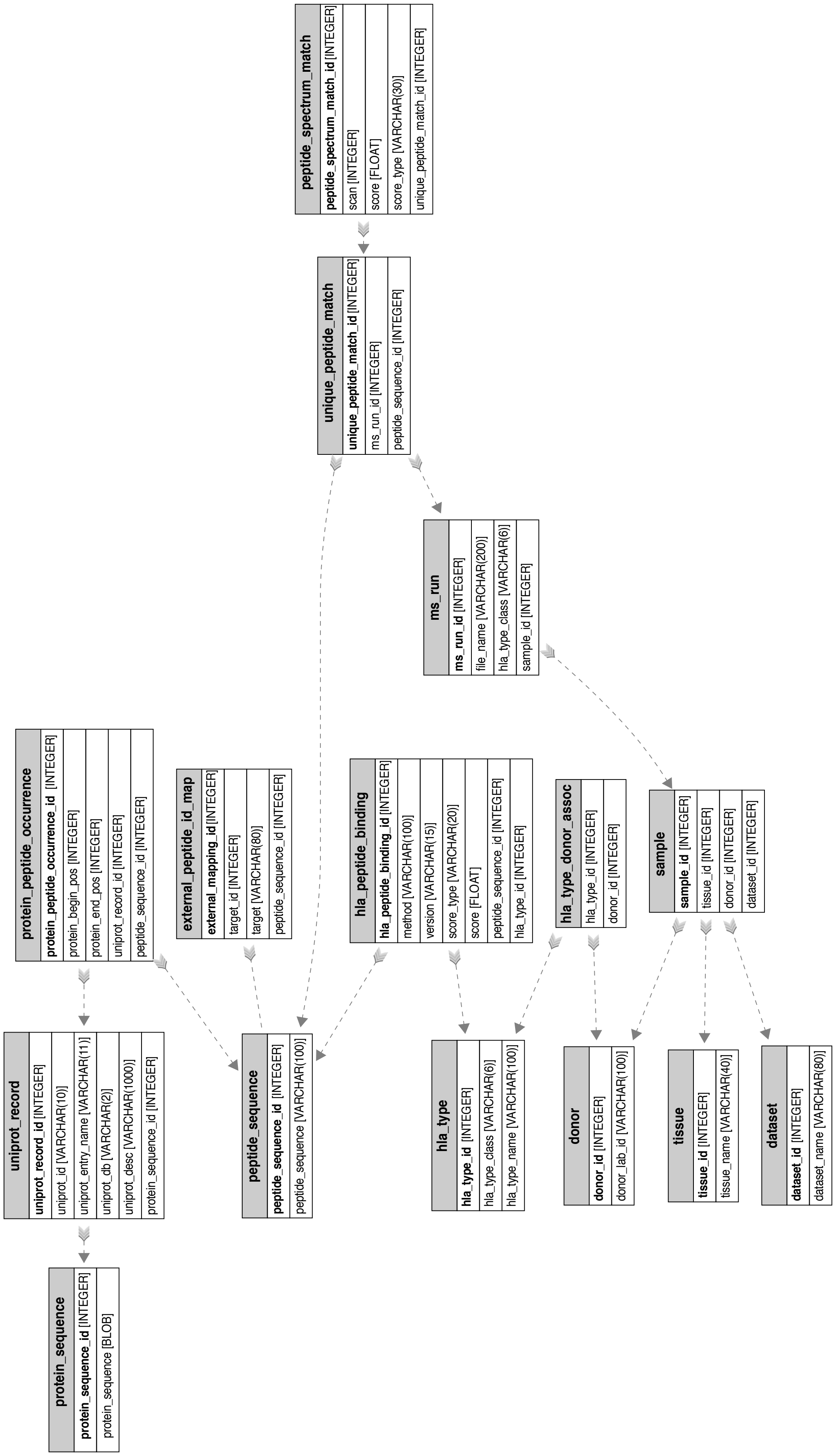

protein_sequence
protein_sequence_id [INTEGER]
protein_sequence [BLOB]
uniprot_record
uniprot_record_id [INTEGER]
uniprot_id [VARCHAR(10)]
uniprot_entry_name [VARCHAR(11)]
uniprot_db [VARCHAR(2)]
uniprot_desc [VARCHAR(1000)]
protein_sequence_id [INTEGER]
protein_peptide_occurrence
protein_peptide_occurrence_id [INTEGER]
protein_begin_pos [INTEGER]
protein_end_pos [INTEGER]
uniprot_record_id [INTEGER]
peptide_sequence_id [INTEGER]
peptide_sequence
peptide_sequence_id [INTEGER]
peptide_sequence [VARCHAR(100)]
external_peptide_id_map
external_mapping_id [INTEGER]
target_id [INTEGER]
target [VARCHAR(80)]
peptide_sequence_id [INTEGER]
hla_peptide_binding
hla_peptide_binding_id [INTEGER]
method [VARCHAR(100)]
version [VARCHAR(15)]
score_type [VARCHAR(20)]
score [FLOAT]
peptide_sequence_id [INTEGER]
hla_type_id [INTEGER]
hla_type
hla_type_id [INTEGER]
hla_type_class [VARCHAR(6)]
hla_type_name [VARCHAR(100)]
hla_type_donor_assoc
hla_type_id [INTEGER]
donor_id [INTEGER]
donor
donor_id [INTEGER]
donor_lab_id [VARCHAR(100)]
sample
sample_id [INTEGER]
tissue_id [INTEGER]
donor_id [INTEGER]
dataset_id [INTEGER]
tissue
tissue_id [INTEGER]
tissue_name [VARCHAR(40)]
dataset
dataset_id [INTEGER]
dataset_name [VARCHAR(80)]
ms_run
ms_run_id [INTEGER]
file_name [VARCHAR(200)]
hla_type_class [VARCHAR(6)]
sample_id [INTEGER]
unique_peptide_match
unique_peptide_match_id [INTEGER]
ms_run_id [INTEGER]
peptide_sequence_id [INTEGER]
peptide_spectrum_match
peptide_spectrum_match_id [INTEGER]
scan [INTEGER]
score [FLOAT]
score_type [VARCHAR(30)]
unique_peptide_match_id [INTEGER]

Figure 5.3: Entity relationship diagram of the HLA Ligand Atlas database. The diagram is sorted from left to right by 1..N relationships, where 1 are illustrated as arrow heads and N as arrow tails. Table keys are highlighted in bold.

framework Bootstrap 4. It is hosted by an Apache webserver. Memory caching is enabled using the library memcached. The table plugin DataTables was used to provide rapid browsing and filtering for tabular data. Interactive plots were designed using Bokeh and ApexCharts. The design of the SQL database and the development of the web application was carried out by Leon Kuchenbecker as contribution to the project.

Raw data availability

The LC-MS/MS immunopeptidomics data comprised in the HLA Ligand Atlas has been deposited to the ProteomeXchange Consortium via the PRIDE partner repository [119] with the dataset identifier PXD019643. LC-MS/MS runs and sample not adhering to the implemented quality control thresholds are deposited as well.

5.3 *Results*

Overview of natural HLA Ligands contained in the HLA Ligand Atlas data resource

Throughout this study we have measured the natural HLA immunopeptidome from 29 distinct organs obtained from 21 research autopsy body donors, surmounting to 1,283 LC-MS/MS runs from 227 mostly paired HLA-I (198) and -II (221) samples using a thorough experimental and computational strategy (Figure 5.4 A). All results are provided publicly in a user-friendly web interface at https://HLA-Ligand-Atlas.org. The majority of tissue samples was obtained from only 14 subjects, while five thymus samples (removed during heart surgery) and two ovary samples (removed preventively) where contributed by additional individuals. Overall, we identified 89,471 HLA-I and 145,190 HLA-II peptides at a local peptide-level FDR of 1% and estimated upper bound of the global peptide-level FDRs of 4.5% and 3.9% for HLA-I and -II peptides, respectively. In a thorough comparison with the main known publically available databases: SysteMHCAtlas [120] and the IEDB [201], this data set boosts the total number of registered HLA ligands from 356,477 to 388,436 for HLA-I and from 74,016 to 192,889 for HLA-II (Figure 5.4 B). The resulting peptide identifications could be attributed to a total of 51 HLA-I and 81 HLA-II allotypes. (Figure 5.5 A)

While limited in the number of different individuals represented in the data resource, the subjects do cover among the most frequently occuring allotypes representing a large part of the human population. Hence, we sought to approximate the worldwide population coverage of HLA allele frequencies comprised in the HLA Ligand Atlas. For this purpose, we computed population coverages using the IEDB Analysis Resources (`http://tools.iedb.org/population/`) [201]. As a result, we observe a population representation frequency of 95.1%, 73.6%, 93%, for HLA-A (n=16), -B (n=21), and -C (n=14) alleles, when considering at least one HLA allele match per individual respectively. Within the same constraints 78.8%, 99.5%, 98.2%, 92.3% for HLA-DPB1 (n=9), -DQA1 (n=11), -DQB1 (n=12), and DRB1 (n=19) alleles are represented respectively (Table C.3).

Allotype characteristics of identified natural HLA ligands

The identified peptides in the immunopeptidomics experiments fit to the well-known length distributions encountered for HLA class I (mostly 8-12mers) and class II ligands (mostly 8-25mers). The mode of the overall peptide length distribution indicates the highest abundance of 9mers (60%) for HLA-I and 15mers (18%) for HLA-II. (Figure 5.5 A) In addition, we observe a strong difference in the source proteins covered by HLA-I and -II ligands. Indeed, gene ontology enrichment of proteins that are exclusively covered by HLA-I and -II ligands confirmed the well-established fact that the cellular compartments associated with the HLA-I antigen presentation pathway are primarily intracellular where as for HLA-II extracellular proteins [41] (Figure 5.5 B,C).

While 85% of the HLA-I ligands are predicted to bind to the respective subject's HLA allotype, this holds true for only 49% of the HLA-II ligands. A major shortcoming of HLA-II binding prediction models is their strong bias towards 15mers, while most HLA-II peptides

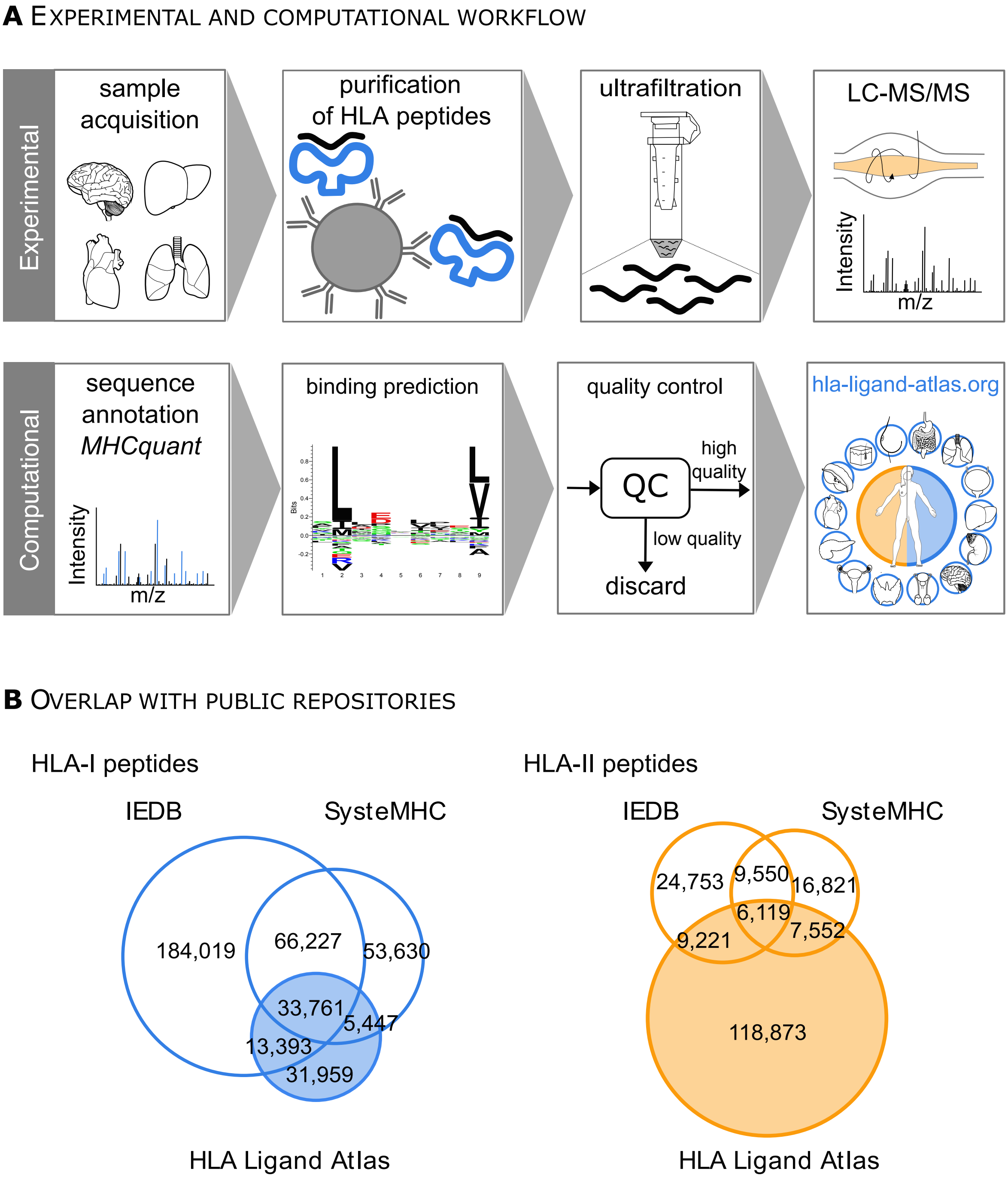


Figure 5.4: A) Experimental and computational high-throughput analysis workflow: Diverse samples of different donors and tissues were collected, purified and filtrated for MHC peptides to undergo LC-MS/MS acquisition. Raw MS data evaluation was achieved using the MHCquant bioinformatics pipeline, binding affinity prediction and a stringent quality control scheme to be released on the web page hla-ligand-atlas.org. (Figure adapted from the original version by Ana Marcu) B) Peptide overlap of the HLA Ligand Atlas with two large existing public databases containing MHC peptides: IEDB and SysteMHC.

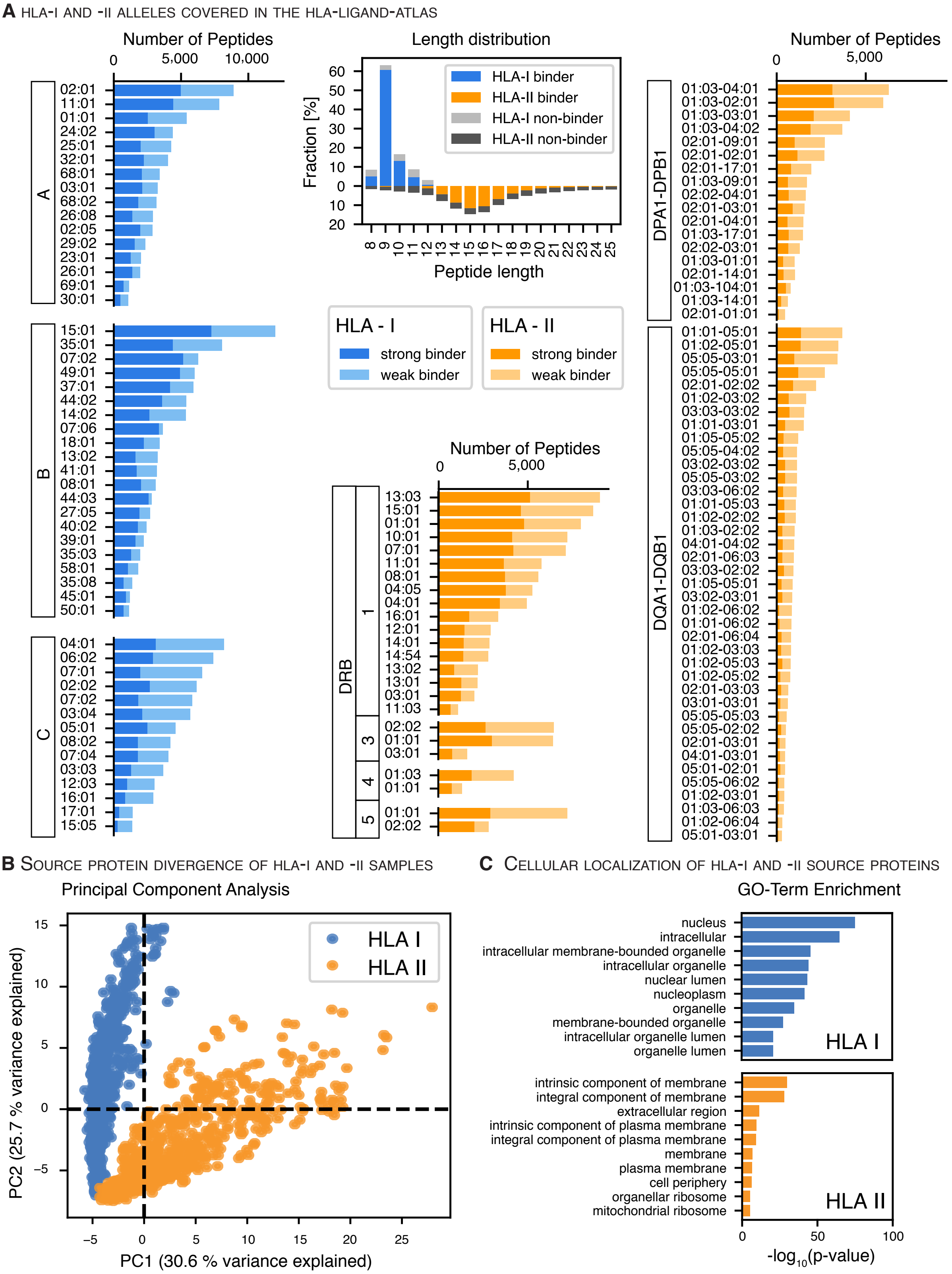


Figure 5.5: A) The diversity of HLA-I (blue) and -II alleles (orange) covered in the HLA Ligand Atlas and the number of peptides matching to each allele by binding predictions. In addition, the length distribution of identified peptides for HLA-I and -II as well as predicted non-binders is shown in the center. (Figure adapted from the original version by Leon Kuchenbecker) B) The divergence of source proteins of HLA-I and -II peptides is shown based on an unsupervised principal component analysis of HLA-I (blue dots) and -II samples (orange dots). C) Gene ontology (GO) term enrichment analysis results of HLA-I and -II exclusive source proteins with respect to their cellular localization.

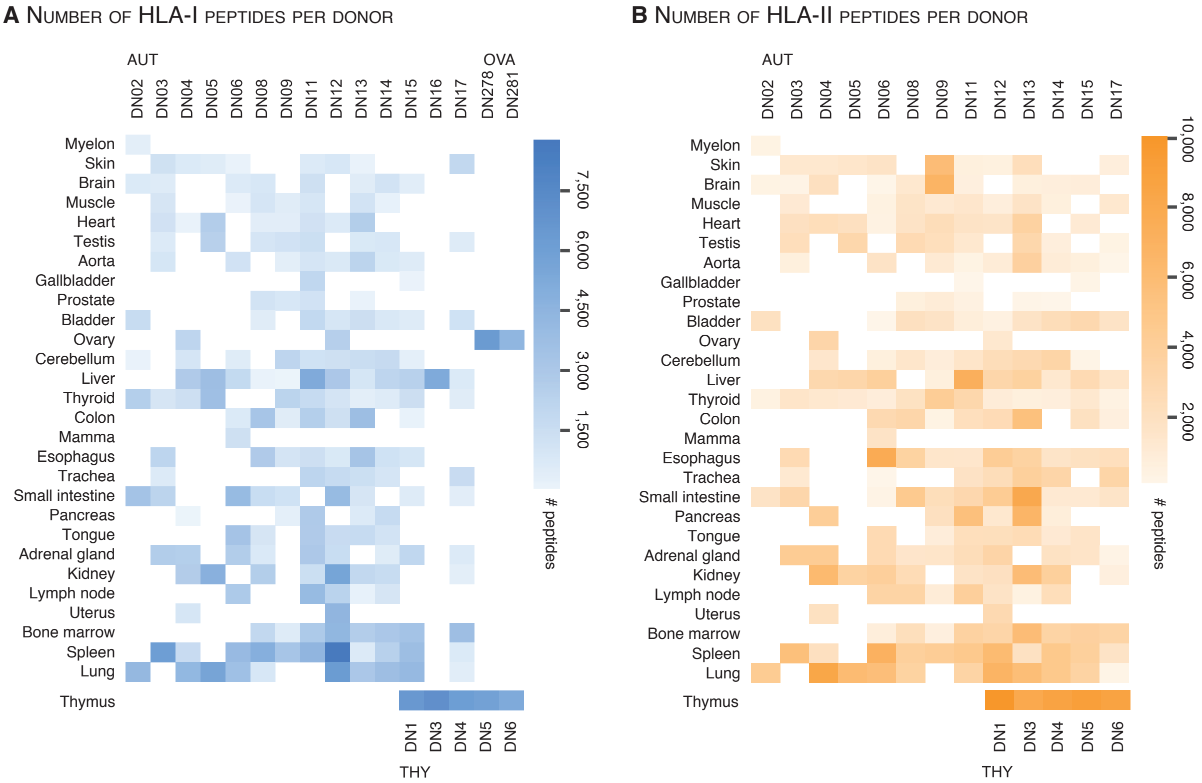


Figure 5.6: A, B) HLA-I peptide yields per tissue and subject are illustrated in a heatmap for HLA-I and –II. The color range is in accordance with the number of peptides identified in each sample as indicated in the legend on the right (HLA-I – blue, HLA-II – orange).

across all other length variants are at a disadvantage. Certain HLA allotypes, such as HLA-A*02:01, -B*15:01, and -C*04:01 are predicted with the highest number of unique strong and weak HLA-I binders. Similarly, HLA-DRB1 alleles were predicted to bind most of the HLA-II ligands characterized. The increased number of peptides presented on a subset of HLA alleles can be attributed to their frequency among the analyzed individuals, to their potentially high copy number on cells, or perhaps to positive binding prediction biases towards frequent HLA alleles. (Figure 5.5 A)

Different modes of immunopeptidome variation across samples

When investigating the immunopeptidome diversity across all samples for both HLA-I and -II alleles, we observed a strong variance in the number of presented peptides identified across donors and tissues. (Figure 5.6) Consequently, we assessed the extent of similarity and heterogenity of the immunopeptidome on both source protein and HLA ligand level between individuals and tissues. For this purpose, we computed pairwise overlaps between all samples by means of the Jaccard similarity index either on the level of identified HLA ligands or on the level of the respective corresponding source proteins.

As a result, we made several observations: (1) In most comparisons samples of the same individuals are more similar than samples from the same tissue or random comparisons (2) generally, source protein level in contrast to the HLA ligand level comparisons result in higher overlaps between samples due to the less direct influence of a individual HLA allotypes

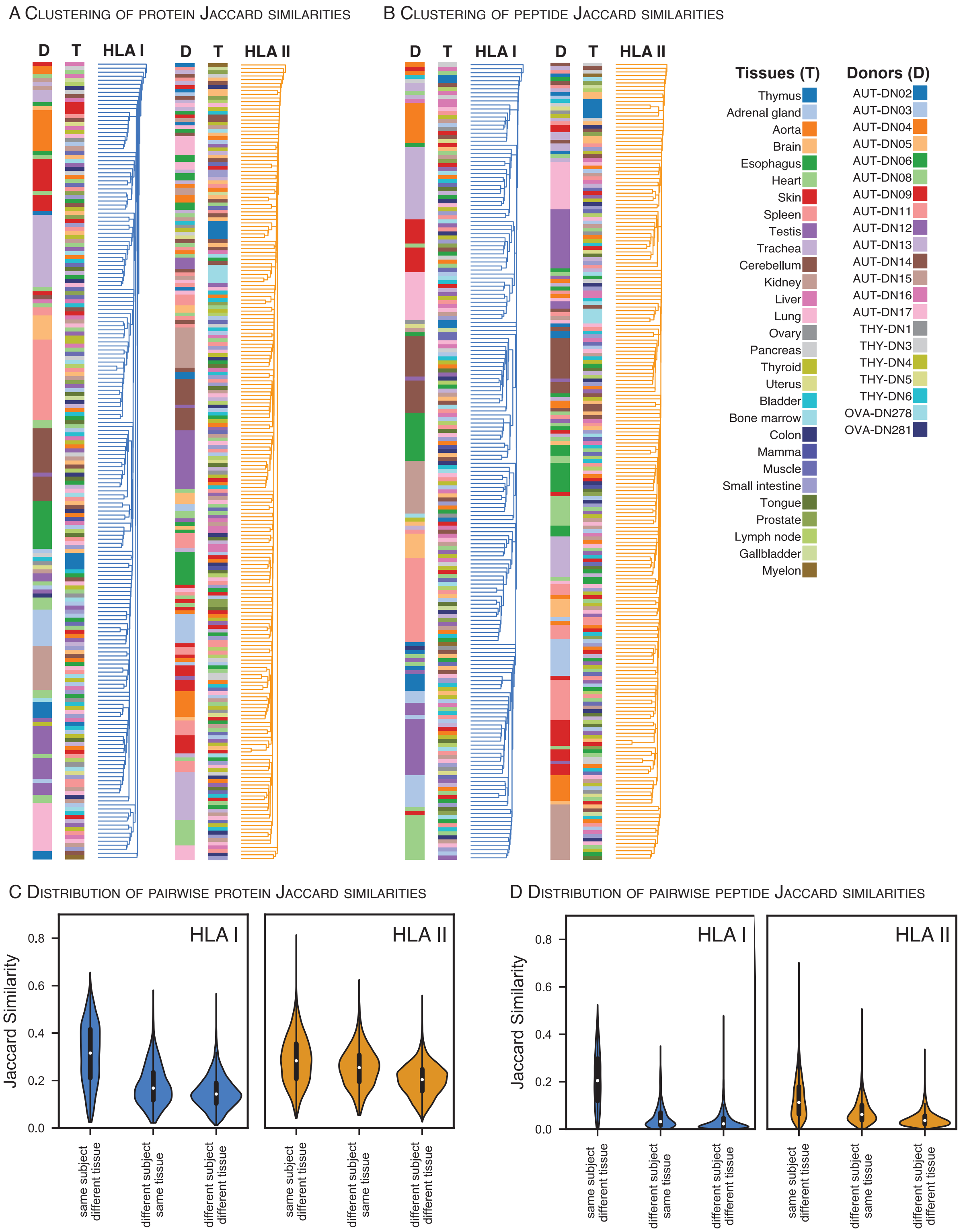


Figure 5.7: A, B) Pairwise hierarchical clustering of samples based on the Jaccard similarity between HLA-I (blue) and HLA-II (orange) on source protein and peptide level, respectively. The dendrogram illustrates the nearest neighbours based on the similarity between tissues and subjects. C, D) Violin plots illustrate the distribution of the Jaccard similarity index for each pairwise comparison between the same subject - different tissues; different subjects - the same tissue, and different subject - different tissues on source protein and peptide level, respectively. (Figure adapted from the original version by Leon Kuchenbecker)

and (3) inter-tissue and random overlaps are slightly higher for HLA-II rather than HLA-I molecules. The high heterogenity of the HLA immunopeptidome between subjects is additionally demonstrated by the results of a hierarchical clustering analysis based on the Jaccard similarity matrix. (Figure 5.7) While the result might seem partially biased due to the intrinsic HLA allotype differences across the individuals, we consistently reproduced the effect also when focusing on allele matched samples and peptides only.

Interestingly, the analysed thymus samples stand out from the other samples regarding heterogeneity and antigen coverage. In fact, a lower degree of subject individuality and an increased share of covered source proteins is revealed, resulting in multiple of the thymus samples being the nearest neighbours in the hierarchical clustering of HLA-I and -II source protein and HLA-II peptide Jaccard similarities.

HLA ligand identifications vary consistently across tissues

Despite the inter-individual (i.e. inter-allotype) variance, there exists a trend of peptide identifications across tissues. Evidently, tissues gradually separate into higher and lower HLA peptide yielding for both HLA classes (Figure 5.8). Tissues that represent the lower end of the spectrum of both HLA-I and -II identifications across all subjects include the skin, aorta, brain, and the gallbladder. In contrast, tissues such as the thymus, lung, spleen, bone marrow, and kidney (Figure 5.8 A, B) are at the higher end of the spectrum of HLA-I and -II peptide identifications. In accordance with these results, most of these high yielding organs have well-characterized immune-related functions or are central components of the lymphatic system [26].

In order to systematically evaluate this effect, we employed a linear model correlating the median number of identified HLA-I/-II peptides with RNA expression values (RPKM) of immune-related genes. The corresponding RNA sequencing data for this comparison was retrieved from a recent study targeting the respective genes across a wide number of individuals and tissues [182] (Figure 5.8 C, D; Figure 5.9). In accordance, we observe a significant correlation between expression values of immune-related genes and HLA-I and -II peptide yields. The highest correlation was detected between genes of the immunoproteasome and the number of HLA-I ligand identifications per tissue (R^2=0.506, rho=0.775, p=0.0001). As for HLA-II, ligand identifications correlate best with the expression of the HLA-DRB1 gene (R^2=0.221, rho=0.416, p=0.0236) among all comparisons.

Small subsets of HLA ligand source proteins are tissue-exclusive

To evaluate the degree of tissue-specificity within the immunopeptidome, we grouped all samples by their tissue of origin and extracted the sets of HLA ligands and their source proteins exclusively found in one tissue. Among those, we observe a particularly small number of HLA-I (ranging from five in mamma to 681 in thymus, on average 1.3% of the proteins found in a given tissue), and HLA-II (ranging from seven in ovary to 541 in thymus, on average 4.1% of the proteins found in a given tissue) ligand source proteins identified exclusively in one organ (Figure 5.10).

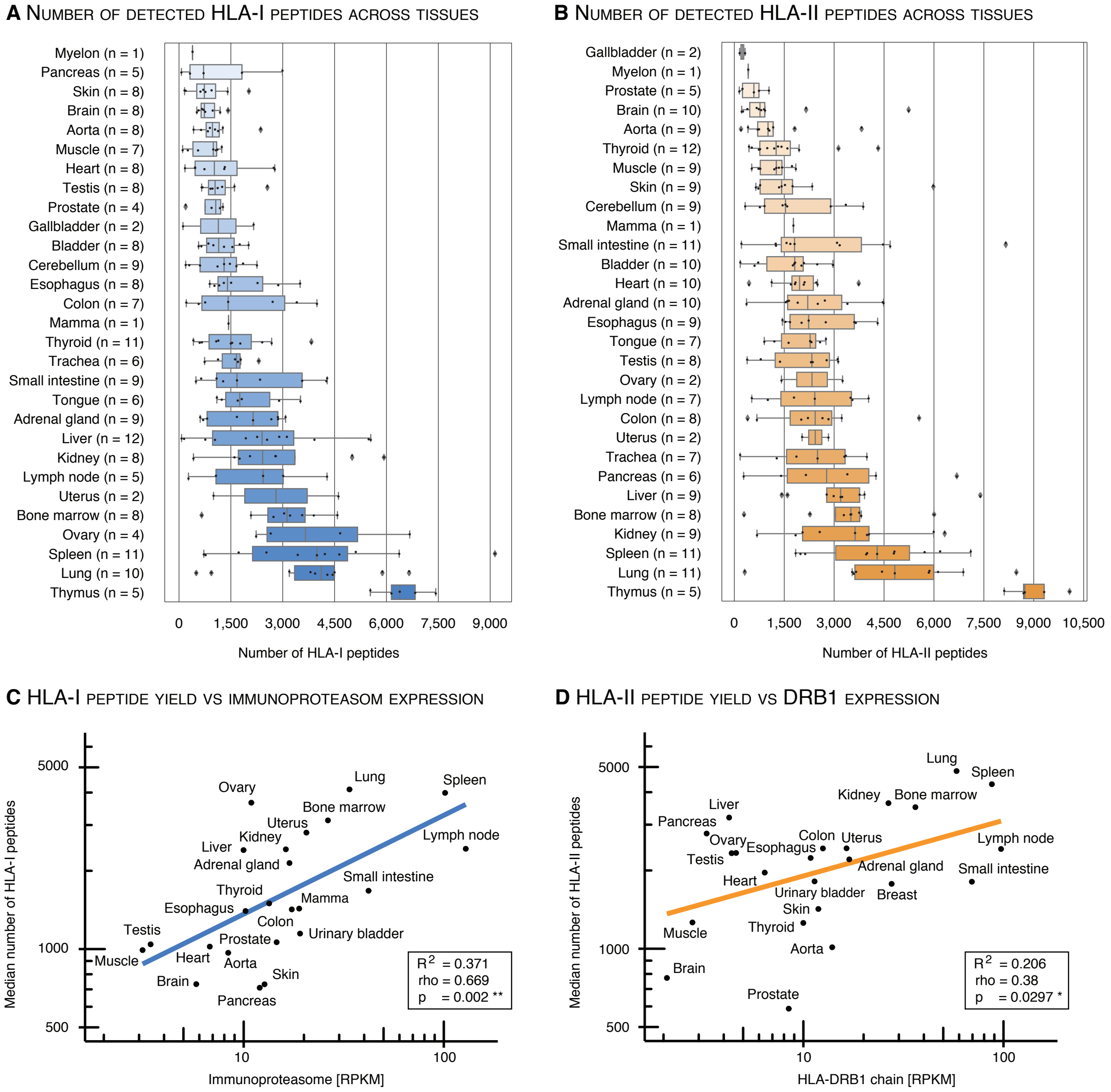


Figure 5.8: A, B) Tissues exhibit a gradual separation based on the HLA-I (blue) and -II (orange) immunopeptidome yield: The number of identified peptides per sample (subject and tissue combinations) was sorted by median immunopeptidome yield per tissue. Boxes span the inner two quantiles of the distribution and whiskers extend by the same length outside the box. Remaining outlier samples are indicated as black diamonds. The number of subjects contributing to each tissue is illustrated on the y-axis in parenthesis. C, D) A linear model was used to correlate the $\log_{10}$ transformed HLA-I (blue) and -II (orange) median peptide yields with $\log_{10}$ transformed median gene expression counts (RPKM) of the immunoproteasome and HLA-DRB1 per tissue.

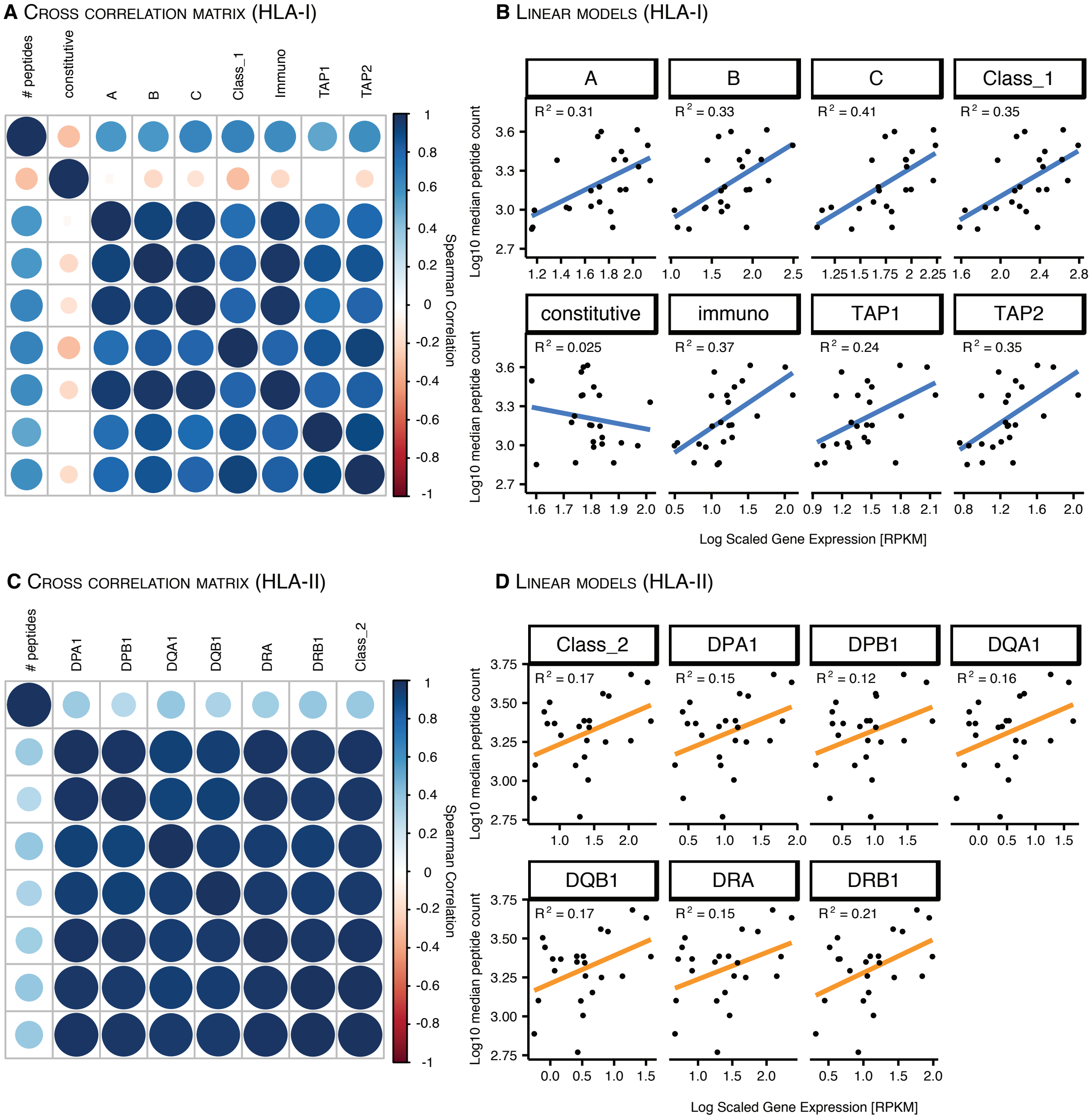
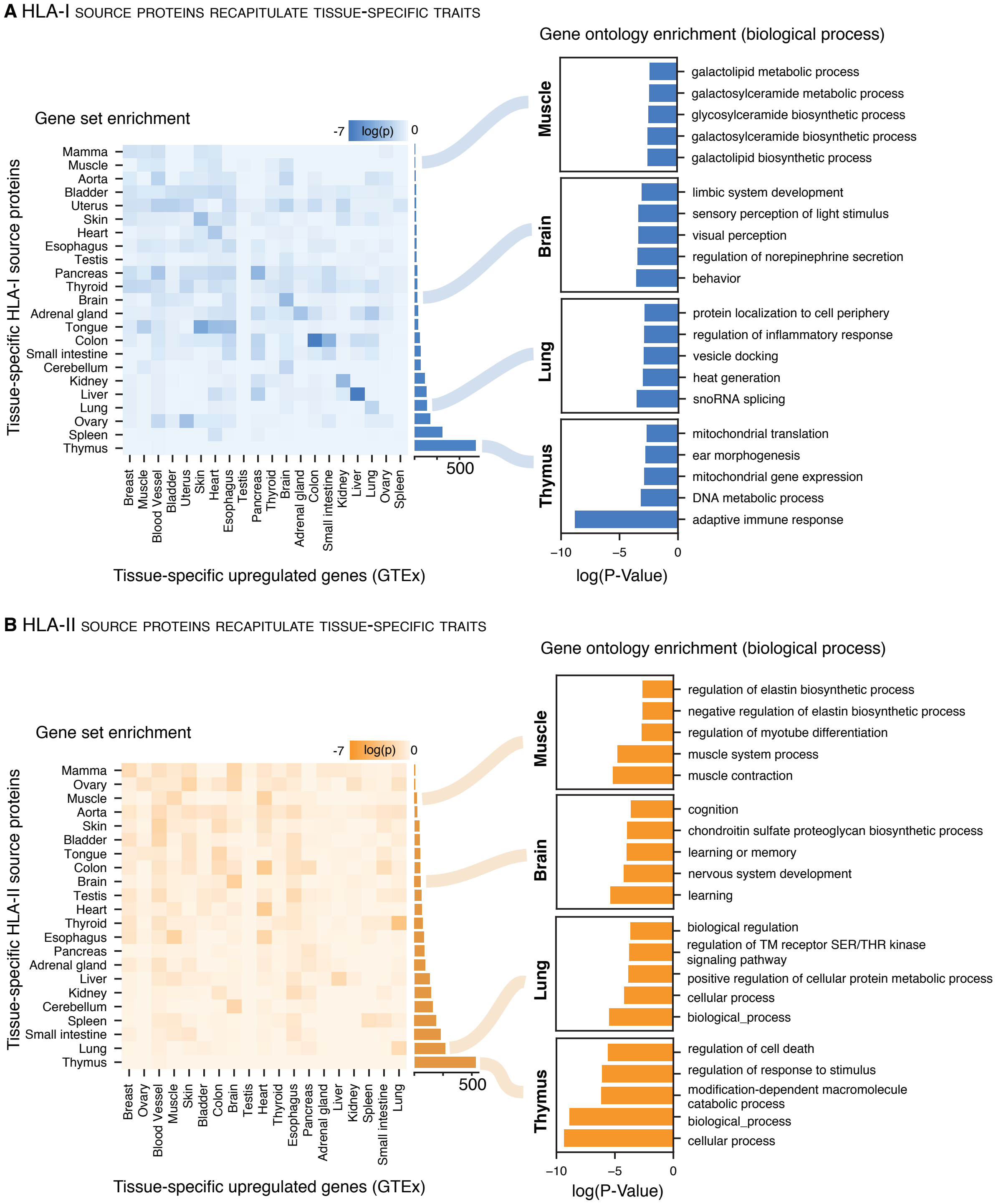


Figure 5.9: A, C) Symmetrical cross-correlation matrices were generated, illustrating the spearman correlation coefficients (rho) of the number of identified HLA-I and -II ligands with expression levels (RPKM) of relevant genes in the HLA-I and -II antigen presentation pathways and among each other. The color-coded dots and their size represent the degree (Spearman rho) of positive (blue) or negative correlation (red). B, D) In addition, linear models illustrate HLA-I and -II immunopeptidome yields correlated to log_{10} scaled gene expression values (RPKM) of genes involved in the HLA-I and -II antigen processing pathway, respectively. The R^2 correlation coefficient is depicted in the top left corner for each comparison.



A HLA-I source proteins recapitulate tissue-specific traits

B HLA-II source proteins recapitulate tissue-specific traits

Figure 5.10: A, B) Gene set enrichment (left) was tested for each tissue by correlating unique HLA-I and –II source proteins per tissue with upregulated genes as annotated in GTEx. Heatmaps depict $\log_{10}$ p-values (Fisher's exact test) for each pairwise comparison. The number of tissue-specific HLA-I and –II source proteins is depicted by the bar plot for each tissue at the right-hand side of the heatmaps. In addition, GO term enrichment (right) of biological processes was performed using the panther DB webservice for selected tissues with the same set of HLA-I and -II tissue-specific source proteins. Top 5 enriched terms with respect to their $\log_{10}$ p-value (Fisher's exact test) were selected.

Despite the low number of detected tissue-exclusive HLA ligands, we sought to determine whether their source proteins reflect organ-specific biology and whether this effect is conserved between the transcriptome and immunopeptidome. For this purpose, known transcriptome upregulated genes per tissue (retrieved from the GTEx repository) were compared with our detected sets organ-exclusive HLA-I and -II source proteins (Figure 5.10 A, B, left hand side). Indeed, we observe that organ-specific biology is conserved between transcriptome and immunopeptidome, since HLA-I and -II source proteins exhibit an enrichment for upregulated genes from the respective tissue. In addition, functional proximity between organs such as the tongue, heart and muscle or brain and cerebellum is reflected in this analysis. At last, gene ontologies (GO) of those HLA ligand source proteins found exclusively on one tissue are indicative of respective organ-specific functions, too for example "adaptive immune response" for the thymus or "nervous system development" for the brain. (Figure 5.10 A, B, right hand side)

However, clear associations between enriched gene sets and gene ontologies of HLA-I/-II source proteins are not evident in all examined organs. For example the spleen or the testis do not show clear enrichment for transcriptome-upregulated genes even though possessing a number of tissue-exclusive protein identifications. Moreover, in general it appears that organ-specific traits are more evident for HLA-I- as opposed to HLA-II source proteins, as supported by lower p-values with GTEx enriched transcripts and function-specific GO terms.

Thus, in summary we observe a slight tissue-specific influence on the immunopeptidome, in particular HLA-I, however this effect ranges from low to not evident in all organs investigated in our study.

Cryptic peptides are prevalent in benign immunopeptidomes

Since recently cryptic peptides (also known as non-canonical peptides) have gained much interest as new sources of HLA-presented peptides [112, 148, 198] and have been claimed to be mainly tumor-associated [197], we did an additional assessment of our data with respect to these targets. As our inital database search was restricted to the set of known reviewed protein sequences, here we researched the HLA-I-restricted LC-MS/MS data with Peptide-PRISM [23] (Figure 5.11 A) a recently developed database independent *de novo* search strategy.

Results from this search identified 1,466 cryptic peptides in the benign immunopeptidome, including one that was previously thought to be tumor-associated only [197] (Figure 5.11 F). Of note, different HLA allotypes appear to have unequally frequent presentation propensities for cryptic peptides, as for example nearly 15% of all identified cryptic peptides were attributed to A*11:01, followed by B*07:02 and A*03:01 (Figure 5.11 B). The finding is supported by 41.13% of these cryptic peptides being identified in more than one individual of the HLA-Ligand-Atlas cohort (Table S3). Moreover, when validating their chromatographic retention time behaviour with a predictive model, both cryptic and conventional peptides share similar properties and most cryptic peptide identifications elute in close proximity to the predicted value (Figure 5.11 D).

The genomic origin of the identified cryptic HLA-I ligands can be classified into the following categories with decreasing frequency: 5′-UTR (48%), followed by Off-Frame (34%), ncRNAs (12%), 3′-UTR (2%), intergenic (2%), and finally intronic regions (1%) (Figure 5.11 C). In accor-

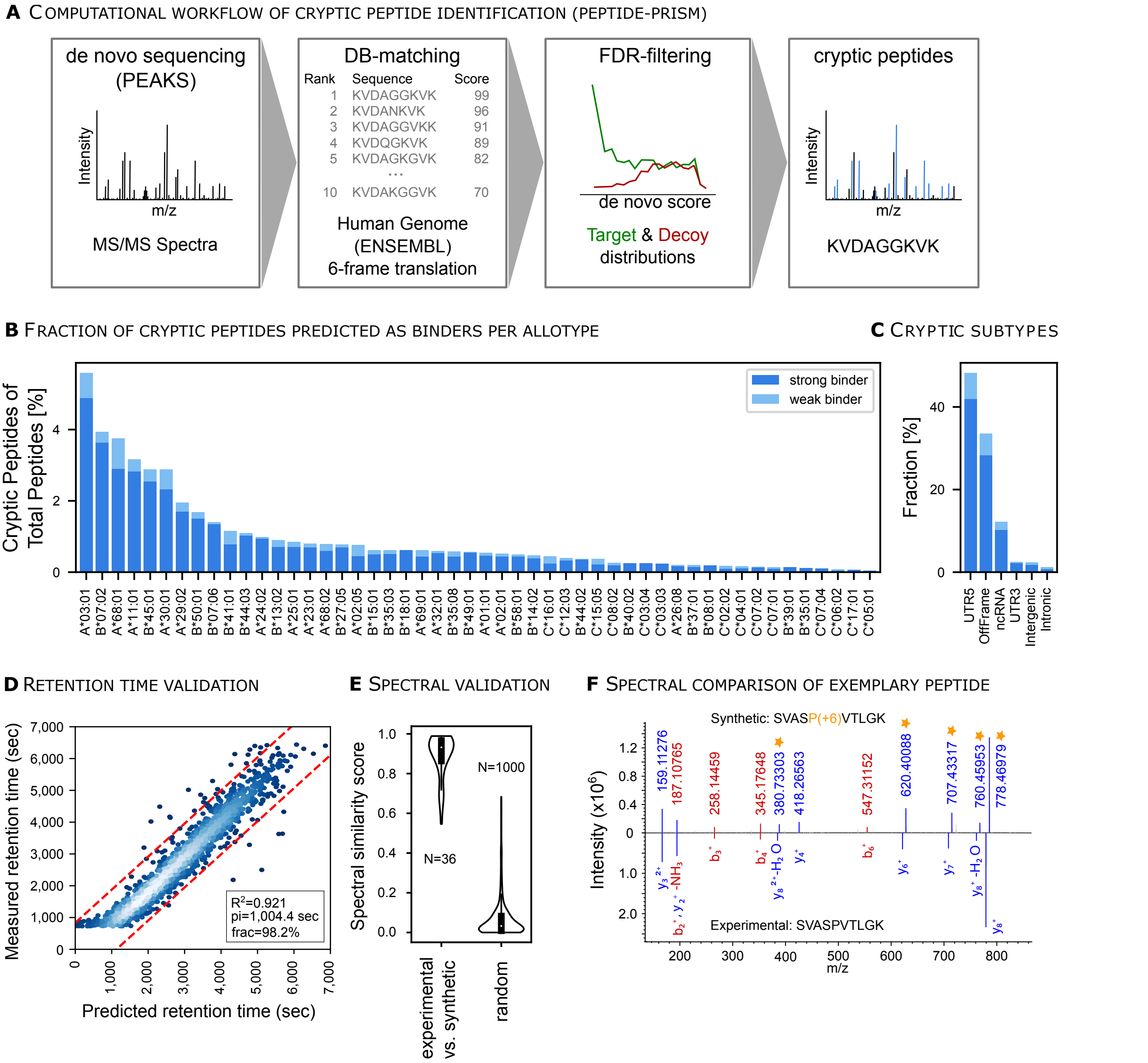


Figure 5.11: Cryptic peptides are prevalent in benign immunopeptidomes: A) Spectra were searched with Peptide-PRISM to identify peptides of cryptic origin. Briefly, de novo sequencing was performed, and top 10 sequences per spectra were queried against a database consisting of the 3-frame translated transcriptome (Ensemble 90). Target-Decoy search was performed per database stratum, separately for canonical and cryptic peptides. B) The HLA-allotype distribution of cryptic peptides was plotted in relation to cryptic and canonical peptides predicted to bind to the respective HLA allotype across all subjects and tissues. C) Distribution of identified cryptic peptides categorized into multiple non-coding genomic regions. D) Linear model correlating measured retention times (RT) of cryptic peptides with their predicted RTs trained on canonical peptide RTs. Corresponding R^2, pi (width of the prediction interval – red dashed lines), and frac (the number of peptides falling into the prediction interval) are indicated in the bottom right. E) 36 cryptic peptides were selected for spectral validation with synthetic peptides. The similarity between the synthetic and experimental spectrum was computed by correlation scores in contrast to 1000 random comparisons. F) Exemplary spectral comparison of the cryptic peptide SVASPVTLGK and its synthesized heavy isotope-labelled counterpart (P+6). Matching b (red) and y ions (blue) are indicated as well as the isotope mass shifted ions (orange stars) of the synthesized peptide. (Figure adapted from the original version containing parts by Andreas Schlosser, Leon Kuchenbecker and Ana Marcu)

dance with previous studies [23, 198] the predominant origin of cryptic peptides is from the 5′-UTR.

Finally, we chose the 36 cryptic peptides with the most evidence (low q-value and shared among subjects) for spectral validation by comparison with a synthetic counterpart. Next to manually verifying the cryptic peptide identifications, we computed a similarity score between the spectra obtained from the experimental vs. synthetic peptides (Figure 5.11 E). In contrast to randomly selected comparisons, similarity scores were highly elevated among the experimental vs. synthetic peptides, reassuring the quality of these peptide identifications. Thus these results suggest that cryptic peptides are not per-se tumor-specific, yet their frequency might be reduced in benign tissues [23].

Quantitative time series of immunopeptidome degradation

In order to exclude a time dependent decay of the immunopeptidome that might have influenced the results of our analysis, we additionally carried out a time series experiment of two benign ovaries and one benign liver. As a result we did not observe a profound qualitative or quantitative degradation of the immunopeptidome for up to 72 h after tissue removal, as neither the ratio of predicted HLA binders to non-binders (Figure 5.12 left hand side) nor a large fraction of the peptide intensities (Figure 5.12 right hand side) changed significantly throughout the time course. Only, This supports the feasibility of employing autopsy tissue as material for immunopeptidomics assays and reassures the quality of the acquired data.

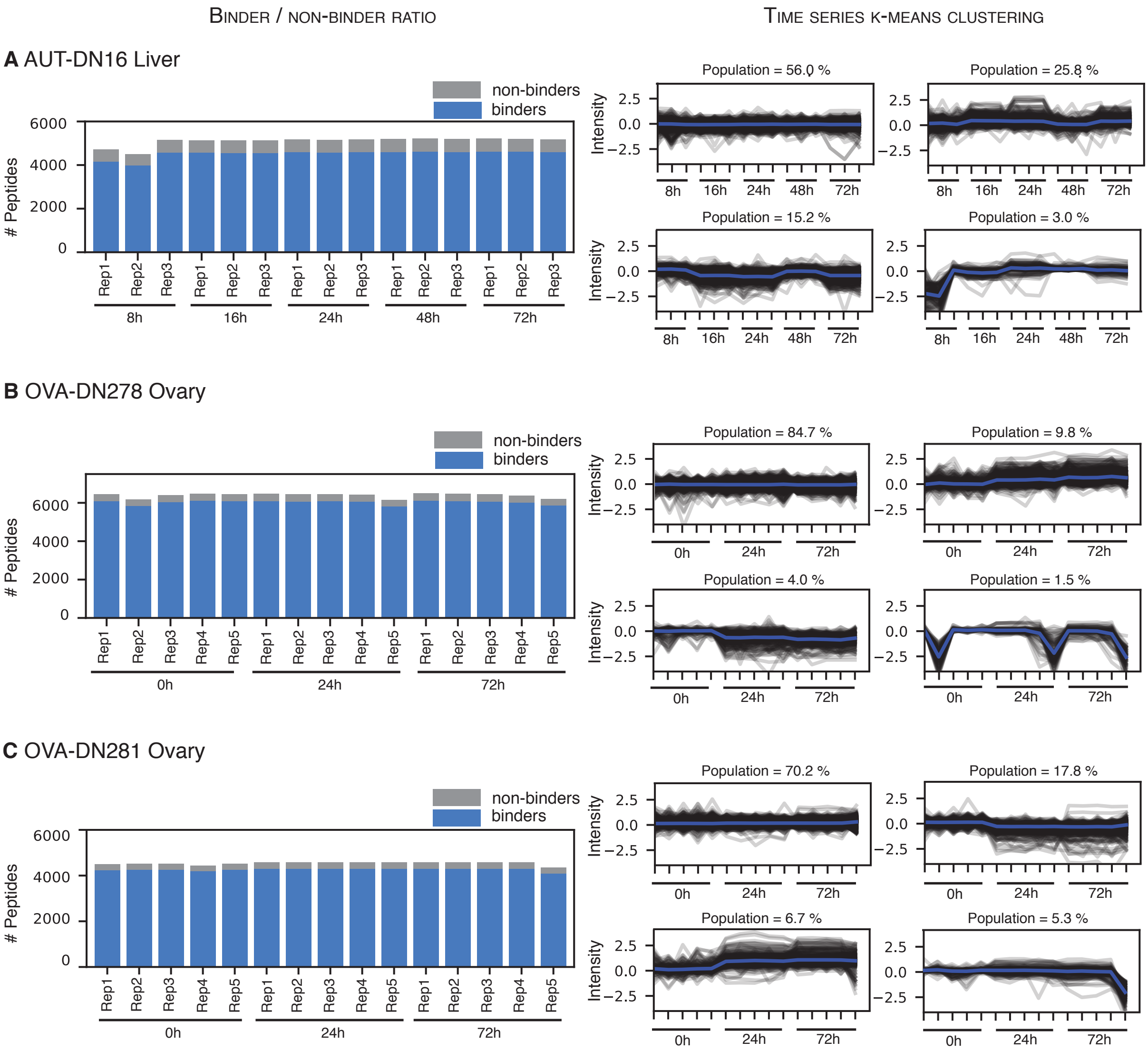


Figure 5.12: The time series experiment was carried out on three biological samples from three subjects (A: AUT-DN16 Liver, B: OVA-DN278, C: OVA-DN281). Bar plots (left hand side) indicate the number of identified HLA-I predicted binders (blue) and predicted non-binders (grey) across technical replicates and time points for each time series. Time series (right hand side) indicate individual clusters (K-Means using four seeds) of trajectories of quantified MS1 intensity across technical replicates and time points. Intermediate trend lines for each cluster are indicated in blue and the percentage of trajectories populating a given cluster is annotated at the top edge of each plot. The trajectories of all time points were set in relation to the initial time point. The analysis reveals that the number of identified peptides and their percentage of predicted HLA-binders is constant and that most trajectories do not vary much across time points.

5.4 *Discussion*

In this work we measured and analysed the currently largest high-resolution MS study of the immunopeptidome of non-malignant human tissues. We cover a multitude of different HLA-I and -II alleles, which are among the most frequent in the human population and thus make this dataset interesting to worldwide research in immunology. In addition, we sampled most organs and tissue types of the human body including central components of the immune system such as the thymus, that is often not accessible to common research biobanks. Moreover, for the first time this enabled us to do a comprehensive comparison of the variation of the immunopeptidome across these tissues, since all samples were handled and measured under the same conditions.

Throughout the analysis of the data described, we were able to confirm several well-established facts and some new discoveries about HLA peptide presentation. For example, we were able to reproduce the general length distribution of HLA-I and -II peptide ligands and their respective intracellular and extracellular origin, respectively. In addition, we confirm previous observations of hotspots of peptide presentation within a given protein sequence [200]. In particular, we also observe that many MS-based discovered HLA-II ligands are not predicted to bind to any of the subjects HLA alleles, highlighting the ongoing need for better prediction methods of HLA-II peptides.

When comparing variability across samples, tissues and individuals we encountered a very large general heterogeneity and an overweight of the inter-individual factor. Clearly, the immunopeptidomes of subjects rather than tissues are more similar, regardless of the level on which the data analysis is based on in contrast to observations of human proteome or transcriptome studies [175, 178, 189]. However, this and the fact that the similarity between samples is in general very sparse, even on source protein-level, can be attributed to the underlying HLA alleles leading to different peptides and corresponding proteins presented by in each subject. This observation is in accordance with a previous study, showing that melanoma metastases from the same patient show substantial differences in their immunopeptidomes [210]. Nonetheless, it should be kept in mind that identification of both source proteins and HLA ligands is intrinsically biased, due to its affectedness by inherent inadequate sampling and detection in MS.

The investigation of the immunopeptidomes across tissues lead to the discovery that there are rather quantitative than qualitative differences evident. Despite our finding that there are only few tissue specific HLA ligands and corresponding source proteins, there is a clear difference between the amount peptides that are isolated and identified from the various tissues. This is in accordance with tissue-wide proteomics studies finding only small numbers of organ exclusive protein identifications [178]. Moreover, in two even more systematic, quantitative follow up analyses of the human proteome and transcriptome across multiple tissues the conclusion was drawn that differences between tissues are more likely quantitative than defined by the presence or absence of certain proteins [211, 212].

The fact that the immunoproteasome might play a role with respect to the different amounts of peptides found across healthy tissues, including tissues for which no primary immunological function would be expected, is supported by two independent proteomic studies identifiying the immunoproteasome in the entire healthy human proteome as well [176, 178]. In addition, immunoproteasom expression has recently been claimed to be take strong influences on the presented HLA peptidome repertoire and is thus associated with response to checkpoint inhibitor therapy in Melanoma [213].

Less strong correlates found for HLA-II on the other hand, is a comprehensible result, since the immunoproteasom is not involved in the HLA-II presentation pathway. Only the HLA-DR protein chain is a strong and variable contributor to the dimeric HLA-II molecules, in contrast to the invariant α chain, which might support the weak correlation found in this case. Moreover, higher expression values for HLA-DRB1 compared to other HLA-II allotypes have been described in early studies on gastric epithelium for example [214] as well. However, during the immunopurification procedure we applied the Tü39L243 antibody, having high specificity for HLA-DR but possibly lower specificity for different other HLA-II allotypes. Hence, we cannot exclude a skewed identification ratio towards HLA-DRB allotypes. Finally, it should be noted that the high paired yield of HLA-I and -II ligands in some of the tissues could be indicative of an increased infiltration by different immune cells. However, since bulk tissue was analyzed and an attempt of sorting factions of cells was not in the scope of the study, a definite statement whether the peptide presentation occurred in tissue cells itself or rather on tissue-infiltrating immune cells cannot be made.

At last, we validated and confirmed the presence of non-canonical or cryptic peptides in the benign immunopeptidome. This fact highlights that these peptide species are to some extend a natural phenomenon and raises the importance for comparing findings in cancer tissues with healthy human tissues instead of NATs [186]. Nevertheless, our findings are still preliminary and will need to be analysed more quantitatively in the future.

With this work we hope to contribute to various branches of research in immunology ranging from basic understanding of immunological processes to infectious disease or cancer therapy in the future. Finally, the raw MS measurements are provided to the public and we encourage reanalysis of the data with future methods and search databases as it might lead to new insights such as the identification of additional peptides not included in our evaluation at this stage.

CHAPTER 6

HepaVac

"Multi-omics discovery of neoantigens in hepatocellular carcinoma"

This chapter includes partially identical or adapted content with permission from:

Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma

M. W. Löffler*, C. Mohr*, L. Bichmann, L. K. Freudenmann, M. Walzer, C. M. Schroeder, N. Trautwein, F. J. Hilke, R. S. Zinser, L. Mühlenbruch, D. J. Kowalewski, H. Schuster, M. Sturm, J. Matthes, O. Riess, S. Czemmel, S. Nahnsen, I. Königsrainer, K. Thiel, S. Nadalin, S. Beckert, H. Bösmüller, F. Fend, A. Velic, B. Maček, S. P. Haen, L. Buonaguro, O. Kohlbacher, S. Stevanović, A. Königsrainer, H.-G. Rammensee, HEPAVAC Consortium.

Genome Medicine. 11, 28 (2019)

Authors marked with * contributed equally to research work and manuscript

A detailed description of the contributions to the project by coauthors is provided in the Appendix D

6.1 *Introduction*

Hepatocellular carcinoma (HCC) is among the cancer malignancies with the highest death toll on a global scale [24] and with very limited therapeutic options. Long-term survival in advanced stages in disease progression is rare [25]. Even though the microenvironment of the liver is tolerogenic and supresses immune responses [215], antigen-specific T-cell responses do occur [216]. Immune infiltration of HCCs by T-cells [217] and spontaneous immune responses have been shown to correlate with longer survival [218], however their overall impact is rather weak and insufficient on their own in the liver. Thus, immunotherapies that could unleash the full potency of the immune system hold great promise.

Immune checkpoint (ICP) inhibitors have recently tracted much attention in cancer immunotherapy, and might allow to achieve an increase of effectiveness to fight malignancies [219]. In contrast to established cytostatic pharmaceutical treatments, this new class of drugs has enabled long-term survival in advanced and metastatic disease previously considered incurable [220]. Induced T-cell activity against mutated neoepitopes arising from somatic tumor mutations has been proposed as a likely mode of action for ICP inhibitors [221].

Accordingly, elevated mutational load and respectively raised neoantigen qualities [222] were determined to be strong correlates for survival after ICP therapy of several tumors [223] including in melanoma (Mel) [224], lung cancer [225] and colorectal carcinoma [226]. In addition, evidence suggests that T-cell responses to neoantigens can generate tremendous clinical effects, such as demonstrated in case reports on advanced Mel [227] and metastatic cholangiocarcinoma [228].

Ultimately, success and feasibility of neoantigen-targeted cancer immunotherapy is likely to depend on the tumor entity, since their mutational loads can vary strongly [229, 230]. As revealed by NGS, in HCC only a small proportion of about 10 % of patients showed mutations potentially accessible for tailored drug therapy [231]. However, HCCs fall into the mid range of mutational burden among the different tumor entities and preliminary data for ICP inhibitors showed objective response rates in 15–20 % of patients combined with a manageable safety profile [232]. Thus, neoantigens are in principle an interesting case for therapeutic intervention in HCC.

This chapter showcases an in-depth multi-omics analysis encompassing whole exome and transcriptome sequencing, combined with proteome and HLA ligandome profiling in selected HCC patients aiming to obtain evidence for the natural presentation of exome-derived mutated HLA ligands. In contrast to the expectations on this project, neoantigens remained elusive to the approach used, hence the biological difficulty as well as technological limits for this therapeutic strategy in HCC in contrast to Mel is discussed. In order to distinguish between the different types of evidence ranging from variants, predictions and experimentally validated neo-epitopes ($\mathrm{NE}^{\mathrm{lig}}$) on the various layers of biological complexity, we use the following glossary of definitions:

Abbreviation	Description
Var	Somatic variant (SNVs, InDels, frameshift variant)
Var^{ns}	Non-synonymous somatic variant
Var^{exp}	Expressed non-synonymous somatic variant
PNE	Predicted neo-epitope
PNE^{exp}	Predicted neo-epitope with evidence on transcript level
PNE^{prot}	Predicted neo-epitope with evidence on proteome level
NE^{lig}	Neo-epitope with evidence on HLA ligandome level
WT^{lig}	Wild-type peptide corresponding to PNE with evidence on HLA ligandome level

6.2 *Materials and Methods*

Ethics approval and informed consent

This study was conducted in accordance with the Declaration of Helsinki and approved by the local institutional review board at the University Hospital of Tübingen, Germany. All participants provided written informed consent before study inclusion.

Clinical specimens

Clinical specimens from patients (n=16) undergoing liver resection for hepatocellular carcinomas (HCC) encompassing both non-malignant and malignant liver tissue as well as peripheral blood were obtained directly after surgery and cryo-preserved (for patient characteristics see Table C.10). HCC diagnosis and predominant tumor fraction within samples were histologically confirmed by an expert pathologist. All included patients were negative for chronic viral hepatitis (hepatitis B and/or C) and without systemic pretreatment for their malignancy.

Next-Generation Sequencing

Extraction of DNA/RNA from fresh frozen tissue and PBMCs was performed using the AllPrep DNA/RNA Kit (Qiagen) from fresh frozen tissue and PBMCs, respectively.

For whole exome sequencing (WES) samples (HCC023-027, HCC034, and HCC036) were prepared using the SureSelectXT Human All Exon v5 or v6 Kit (Agilent, Waldbronn, Germany). For whole transcriptome sequencing (WTS) samples were prepared with the TruSeq Stranded mRNA Kit (Illumina, Eindhoven, Netherlands). Paired-end sequencing was performed with the HiSeq 2500 or NextSeq500 System (Illumina).

For WES DNA libraries were prepared for samples (HCC028, HCC030, HCC035, and HCC038-045) with SureSelect XT Human All Exon v6 Kit (Agilent, Waldbronn, Germany) and sequenced in paired-end mode on a HiSeq 4000 System (Illumina, Eindhoven, Netherlands). RNA library preparation was performed using the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Clontech, Saint-Germain-en-Laye, France) and sequenced on a HiSeq 4000 System (Illumina, Eindhoven, Netherlands).

Protein in-gel digestion for shotgun proteome analysis

Eluted protein samples were purified by SDS-PAGE. Coomassie-stained gel pieces were digested using trypsin. Extracted peptides were desalted using C18 Stage tips and subjected to LC-MS/MS analysis.

Mass spectrometric analysis of shotgun proteome analysis

Liquid chromatography tandem mass spectrometry (LC-MS/MS) analyses were performed on an EasyLC nano-HPLC (Proxeon Biosystems, Roskilde, Denmark) coupled to an LTQ Orbitrap Elite (Thermo Fisher). Peptide mixtures were separated on a 15 cm fused silica emitter of 75 µm inner diameter (Proxeon), in-house packed with reversed-phase ReproSil-Pur C18-AQ 3 µm resin (Dr. Maisch GmbH, Ammerbuch, Germany). Peptides were injected with solvent A (0.5 % acetic acid) at a flow rate of 500 nl/min and separated at 200 nl/min. Separation was performed using a linear 130 min gradient of 5-33 % solvent B (80 % ACN in 0.5 % acetic acid). Each of four samples was run as one technical replicate. LTQ Orbitrap Elite was operated in the positive ion mode. Precursor ions were acquired in the mass range from 300 to 2,000 m/z followed by MS/MS spectra acquisition of the 20 most intense precursor ions. Higher-energy CID (HCD) MS/MS spectra were acquired with a resolution of 15,000 and a target value of 40,000. The normalized collision energy was set to 35, activation time to 0.1 ms and the first mass to 120 Th. Fragmented masses were excluded for 60 s after MS/MS. The target values were 1E6 charges for the MS scans in the Orbitrap and 5,000 charges for the MS/MS scans with a maximum fill time of 100 ms and 150 ms, respectively.

Isolation of naturally presented HLA ligands from tissues for HLA ligandomics

HLA-I peptide complexes were isolated from HCC and corresponding (non-malignant) liver tissue samples by immunoaffinity purification as described previously [233] , using the pan-HLA-I specific monoclonal antibody W6/32 [126] (produced in-house at the Department of Immunology) and eluted using 0.2 % trifluoroacetic acid.

Mass spectrometric analysis for HLA ligandomics

Peptide extracts were separated by UHPLC (UltiMate 3000 RSLCnano System, Dionex) at a flow rate of 175 nl/min using a 50 µ25 cm C18 column (PepMap RSLC, 2 µm particle size, Thermo Fisher) and a linear gradient ranging from 3 to 40 % solvent B over the course of 90 min (Solvent A: 0.15 % formic acid; Solvent B: 80 % ACN) in several technical replicates, as described previously [233].

Eluting peptides were analyzed in an online-coupled LTQ Orbitrap XL mass spectrometer (Thermo Fisher) operated in automated data-dependent acquisition (DDA) mode. In the Orbitrap, survey scans of peptides with 400-650 m/z as well as 2+ and 3+ as permitted charge states were recorded at a resolution of 60,000 with subsequent selection of the five most abundant precursor ions for collision-induced dissociation (CID). The normalized collision energy was set to 35, activation time to 30 ms and the isolation width to 2.0 m/z. MS/MS spectra were acquired in the linear ion trap (LTQ) and corresponding precursor ions were dynamically excluded for 3 s after fragmentation. Each sample was acquired in up to five technical replicates.

To enhance sensitivity of neo-antigenic peptide identification, selected ion monitoring (SIM; LTQ Orbitrap XL) and parallel reaction monitoring (PRM) (Orbitrap Fusion Lumos, Thermo Fisher) for selected samples was additionally performed. A table listing the exact peptide sequences and their heavy isotope labels is contained in the original publication of Löffler and Mohr et al. [234] in the corresponding supplementary material tables S7 and S8.

Heavy isotope-labelled synthetic peptides for the SIM approach were purchased from Thermo Fisher. Synthetic peptide retention times (RT) were assessed in an HLA-I peptide matrix eluted from JY cells. Subsequently, these were used to create a scheduled SIM method triggering fragmentation of PNEs (2+ precursor ions) independent of their relative abundance in patient peptide extracts. Due to the number of PNEs, three SIM measurements were scheduled and thus several measurements per tumor sample were necessary. SIM scans of HCC025 and HCC026 were performed with the same UHPLC settings as top 5 CID measurements, whereas HCC027 SIM acquisition was performed using a 50 µm × 50 cm C18 column (PepMap RSLC, 2 µm particle size, Thermo Fisher) and a linear gradient ranging from 3 to 40 % solvent B over the course of 140 min. The normalized collision energy was set to 35, activation time to 30 ms and the isolation width to 1.5-2.0 m/z.

Heavy isotope-labelled synthetic peptides for PRM targeted tandem MS (PRM tMS2) were manufactured in-house by solid-phase peptide synthesis at a purity >60 %. PRM tMS2 methods targeting the heavy isotope-labelled synthetic peptide or its natural counterpart (2+ and 3+ precursor ions) were created using Skyline v4.1 [235, 236]. The retention time (RT) as well as the amount of each synthetic peptide necessary for reliable detection was assessed by titration in an HLA-I peptide matrix eluted from JY cells. Synthetic peptide purity as determined by high performance liquid chromatography (HPLC) and 75 % peptide content (reference values of nitrogen determination: 20 % TFA and 5-10 % H 2 O) were considered for weighed-in amounts. Peptides dissolved in 10 % DMSO were spiked at 4 – 20 fmol/µl and 5 µl were injected for PRM tMS2 measurements on an Orbitrap Fusion Lumos.

Raw files of these titration measurements were processed with Proteome Discoverer v1.4 (Thermo Fisher) using the SEQUEST HT search engine [82]. Based on synthetic peptide RTs ±12 min, PRM tMS2 data acquisition of tumor and autologous non-malignant liver samples was scheduled. Peptide extracts were separated by UHPLC (UltiMate 3000 RSLCnano System, Dionex) at a flow rate of 300 nl/min using a 50 µm × 25 cm C18 column (PepMap RSLC, 2 µm particle size, Thermo Fisher) and a linear gradient ranging from 3 to 40 % solvent B over the course of 90 min (Solvent A: 0.15 % formic acid; Solvent B: 80 % ACN). Eluting peptides were analyzed in an online coupled LTQ Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher). In the Orbitrap, survey scans of precursor ions (HCC025: 320-670 m/z, HCC026: 300-650 m/z; 2+ and 3+ as permitted charge states) were recorded at a resolution of 120,000 with subsequent selection for collision-induced dissociation (CID). The normalized collision energy was set to 35 and the isolation width to 1.4 m/z. At a resolution of 60,000, MS/MS spectra were acquired in the Orbitrap.

In addition to PRM tMS2 measurements, one top five run in DDA mode (HCC025: 320-670 m/z, HCC026: 300-650 m/z) per sample was performed. In the Orbitrap, survey scans of precursor ions (HCC025: 320-670 m/z, HCC026: 300-650 m/z; 2+ and 3+ as permitted charge states) were recorded at a resolution of 120,000 with subsequent selection for collision-induced dissociation (CID). The normalized collision energy was set to 35 and the isolation width to 1.4 m/z. At a resolution of 30,000, MS/MS spectra were acquired in the Orbitrap.

Synthetic peptides in an HLA-I peptide matrix were back-to-back eluted and acquired in scheduled PRM tMS2 (4 fmol/µl) or in top 5 DDA mode (10 fmol/µl), respectively. DDA measurements of synthetic peptides (10 fmol/µl), were processed with Proteome Discoverer v1.4 (Thermo Fisher) using the SEQUEST HT search engine [82] and served as spectral library for analysis of scheduled PRM tMS2 data in Skyline [235, 236].

PatientID	**Whole exome sequencing**	**Transcriptomics**	**Proteomics**	**HLA ligandomics**
	T+N	**T+N**	**T+N**	**T+N**
HCC023	✓	✓	✓	✓
HCC024	✓	✓	✓	✓
HCC025	✓	✓	✓	✓
HCC026	✓	✓	✓	✓
HCC027	✓	✓	✓	✓
HCC028	✓	✓	✗	✓
HCC030	✓	✓	✗	✓
HCC034	✓	✓	✓	✓
HCC035	✓	✓	✗	✓
HCC036	✓	T only	✓	✓
HCC038	✓	✓	✗	✓
HCC040	✓	✓	✗	✓
HCC041	✓	✓	✗	✓
HCC042	✓	✓	✗	✓
HCC043	✓	✓	✗	✓
HCC045	✓	✓	✗	✓

Table 6.1: Overview of biological omics layers that were assessed per patient in the study. Tissue samples per patients are labeled as tumor (T) and non-malignant (N) and if not available (✗).

Bioinformatics analysis

Variant calling from whole exome sequencing data

Generated reads were processed using the megSAP pipeline (`https://github.com/imgag/megSAP`) and the ngs-bits package (`https://github.com/imgag/ngs-bits`) by the Department of Medical Genetics and Applied Genomics (Tübingen, Germany) [HCC023-HCC027, HCC034, HCC036].

Adapter trimming was performed with SeqPurge [237] . Reads were mapped against the Genome Reference Consortium Human Build 37 (GRCh37) using BWA-mem [238] and Samblaster [239] was used for duplicate annotation. Local realignment of reads in target regions was done with ABRA [240]. Overlapping reads were trimmed with an in-house tool for reduction of false-positive variants with very low allele frequencies.

Somatic variant calling was performed using Strelka and Strelka2 [241, 242] . Derived variants were annotated with SnpEff/SnpSift [243, 244] , vcflib (`https://github.com/ekg/vcflib`), and dbNFSP [245]. High-confidence variants were obtained using custom filter criteria and further annotated with in-house variant frequencies, tumor RNA depth and allele frequencies. RNA reads were preprocessed to remove adaptor sequences in the same way and then mapped with STAR [246] to the same reference genome.

Otherwise, sequenced reads were demultiplexed with Illumina bcl2fastq 2.19. and Skewer 0.2.2 [247] was used for adapter trimming, followed by read mapping with an in-house version of BWA-mem [238] v0.72 against an in-house version of hg19. Local realignment of reads in target regions was done with ABRA [240]and duplicate reads were discarded using SAMtools v0.1.18 [248] . Somatic variants, called with a proprietary software (CeGaT GmbH, Tübingen, Germany), were filtered for a minimal coverage of 30x in tumor and non-malignant tissue and an allele frequency greater than 0.05 in tumor tissue and three-fold less in non-malignant tissue. In case of HCC028, HCC030, HCC035, and HCC038-HCC045, somatic mutations were annotated using SnpEff 4.1k [243]

HLA typing from whole exome sequencing data

Typing at four-digit resolution using WES data was performed by OptiType [249] for HLA-I alleles and confirmed in selected cases by molecular HLA typing (using clinically validated LUMINEX and sequence-based typing) during clinical routines (see Supplementary Table 3).

Gene expression analysis from transcriptome sequencing data

Gene expression values were calculated as fragments per kilobase of exon per million reads mapped (FPKM) of the corresponding transcripts and RNA tumor sequencing depth at the corresponding variant position. Mapping of RNA reads was done using TopHat 2 (v2.0.12) [250] . Adapters were removed beforehand with CutAdapt (–discard-trimmed) based on FastQC results (v0.10). Counts for mapped RNA reads were calculated using HTSeq (0.6.1p2) [251] . FPKM values were calculated as follows:

$$FPKM = \frac{10^9 xC}{NxL} \tag{6.1}$$

where, L is the exon length in base pairs for the corresponding gene, C is the number of reads that mapped to a gene (number of counts from HTSeq run), and N is the total number of unique mapped reads in the sample.

Proteomic data analysis

Proteome MS data from in-gel digestions was processed for protein identification using with MaxQuant software suite v.1.5.2.8 using the Andromeda search engine [96]. The human reference database was obtained from UniProt (taxonomy ID 9606, containing 91,646 protein entries and 285 commonly occurring laboratory contaminants) and concatenated with the

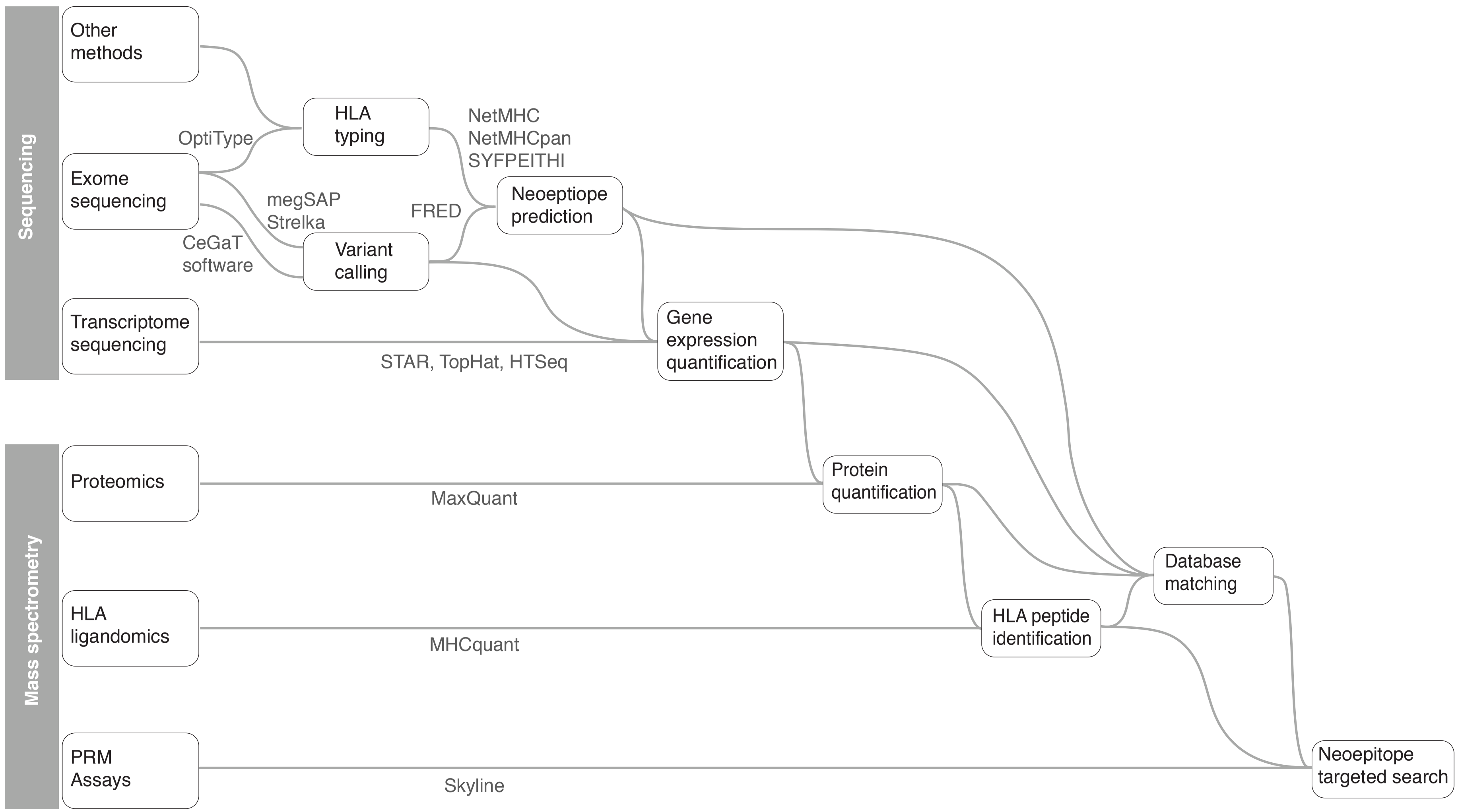


Figure 6.1: Neoepitope discovery multi-omics workflow overview: Mulitple software tools (grey) were applied to corresponding biological sequencing and mass spectrometry methods and their results combined in order to trace the presence of mutated neoepitopes in HCC and Mel patient samples. A more detailed explanation of each step is contained in the corresponding method section. The software tools were applied individually and the pipeline does not exist as one combined tool. Several analysis steps were carried out by coauthors of the project as mentioned in Appendix D.

patient-specific mutanome. Endoprotease trypsin was fixed as protease with a maximum of two missed cleavages. Oxidation of methionines and N-terminal acetylation were specified as variable modifications, whereas carbamidomethylation of cysteines was defined as a fixed modification. Initial maximum allowed mass tolerance was set to six ppm. A FDR threshold of 1 % was applied at peptide and protein level.

Label-free protein quantification was done using MaxQuant v1.5.0.0 [96] . Parameter groups were defined for non-malignant liver- and tumor-derived raw files, respectively. The multiplicity was set to one. Protein N-terminal acetylation as well as oxidation of methionine residues were selected as variable modifications, whereas carbamidomethylation of cysteine residues was set as fixed modification. TrypsinP was selected as enzyme with specific digestion mode. Further, the match type MatchFromAndTo was specified and the number of MaxMissedCleavages was set to two. Requantification and matching between runs were enabled. As a reference, the Swiss-Prot reviewed human proteome (version UP000005640, derived: 16/02/2016) was utilized.

HLA ligandomics data analysis

MS data analysis obtained from HLA-immunoprecipitates was assessed using a Python implemented development version of the MHCquant bioinformatics workflow (see Chapter 3) available in the qPortal instance [19] at the Quantitative Biology Center, Tübingen under the name "Ligandomics ID Coprocessing 2.1" [123]. It uses the same functionalities as the Nextflow and KNIME implementations of MHCquant (see Chapter 3) provided by tools of the open-source software library for LC/MS OpenMS 2.3 [17]. Identification and post-scoring were performed using the OpenMS adapter to Comet 2016.01 rev. 3 [83] and Percolator (3.1.1) [88] . HLA ligand identification was performed against a personalized version of the human reference proteome (Swiss-Prot, reviewed UP000005640), including the patient-specific mutanome.

Database search was carried out without enzymatic restriction and oxidation of methionine residues as the only dynamic modification (maximal number of modifications per peptide set to 3). The digest mass range was set to 800-2,500. Precursor charge was fixed to 2-3 and the precursor mass tolerance was set to 5 ppm. In addition, a fragment bin tolerance of 1.0 Da and a fragment bin offset of 0.4 Da was set and neutral losses were included for each peptide spectrum match (PSM). A 5 % PSM FDR threshold was calculated using Percolator, based on a competitive target-decoy approach using reversed decoy sequences and merged identifications of all replicate runs if available. Peptide quantification was achieved using MapAlignerIdentification and FeatureFinderIdentification [131] with default settings. IDs of replicates were treated as internal IDs and the median intensity of consensus features was used as final quantification value. Only quantified identifications were considered to be valid hits. HLA-I annotation was performed using an in-house version of SYFPEITHI [32] , netMHC 4.0 [138] , and netMHCpan 3.0 [37, 78].

HLA affinity prediction for PNE^{lig}

Peptides of 8-11 amino acids length were constructed by sliding a shifting window of the peptide length over the affected mutated positions. Resulting peptides were filtered against the human proteome (UniProt UP000005640, 02/29/16) and the Ensembl proteome reference (release 84, 04/27/2016) to avoid the selection of identical peptides, contained within wild-type proteins. In case of frameshift mutations, the reading frame offset was monitored in order to determine sequences of alternative reading frames, resulting in altered amino acid sequences and therefore yielding neo-epitopes. Transcript information was retrieved via BioMart, based on the stable database version of GRCh37 (http://feb2014.archive.ensembl.org). HLA binding prediction was performed with SYFPEITHI [32] , netMHC 4.0 [138], and netMHCpan 3.0 [37, 78].

The workflow was implemented using FRED2 [128] . All reported predictions include variant details, mutated peptide sequence, HLA allele, prediction method, corresponding binding score, half maximal score, and a qualitative distinction between binding and non-binding peptides, which is based on the score of the corresponding method. SYFPEITHI-predicted peptides were considered binders, when prediction scores exceeded half of the maximal score of the corresponding HLA allotype matrix. According to netMHC and netMHCpan, predicted peptides with affinities (IC 50 values in nM) $\leq$ 500 nM were selected. Results were further annotated with gene expression values, protein quantification values, and the results of HLA ligandome analysis.

Database matching for prioritization of WT^{lig}

HLA ligandome database queries refer to the in-house database maintained at the Department of Immunology encompassing > 300,000 unique HLA-I peptides identified through MS/MS in diverse tissues (non-malignant as well as with pathologies including malignancies). Database matching was carried out using rSQL, querying for an exact string match of the respective wild-type ligand (W_{lig}) matching to the respective predicted neo-epitope (PNE). All HLA-I allotypes of our HCC and Mel cohort were covered by respective samples in the database. Each sample containing the respective ligand was counted as a separate match.

333,431 different HLA-I peptides have been identified on benign (n=631), malignant (n=780) or (n=115) human samples with pathologies including both primary tissues and (established) cell lines. In total, the database comprises 2,646,952 HLA-I peptides corresponding to 49 different HLA alleles, encompassing 18 HLA-A, 27 HLA-B, and 14 HLA-C alleles, including duplicates.

Queries against the Immune Epitope Database (IEDB; `http://www.iedb.org/`) [201] were performed after filtering for HLA-I ligands, annotated as positive, positive-high, positive-intermediate, and positive-low.

Data storage and management

All data was stored, managed, as well as partially analysed via the qPortal instance [19] at the Quantitative Biology Center, Tübingen, if not stated otherwise (Project IDs: IVAC_HEPA_VAC and MNF_RAMMENSEE). In addition, the LC-MS/MS proteomics and immunopeptidomics data has been deposited to the ProteomeXchange Consortium via the PRIDE partner repository [119] with the dataset identifiers PXD013057 and PXD004894.

6.3 *Results*

Traces of mutation-derived HLA ligands on different omics levels

With the aim to identify individual somatic tumor mutations resulting in natural HLA ligands presented to T-cells, hence neo-epitopes, analyses of malignant and non-malignant liver tissue resected during surgery for HCC from 16 patients was performed. This multi-omics approach encompassed the search for evidence of neo-epitopes on different biological levels including exome (n=16), transcriptome (n=16), shotgun proteome (n=7) as well as the HLA ligandome (HLA-presented peptides; n=16). (Figure 6.2, Table 6.1)

EXOME On average 151±40 somatic variants (Var), including single nucleotide variants, insertions/deletions, and frameshift variants, were detected per HCC, when referenced against DNA from blood. 44% of these (66±19) caused changes in the amino acid sequence of the encoded protein (i.e., Var^{ns} - non-synonymous variants). When assessing the number of predicted HLA-binding neo-epitopes (PNE) per patient suitable to bind to their individual set of HLA alleles, an average number of 244±77 peptides from Var^{ns} per patient was predicted, exceeding the binding threshold.

TRANSCRIPTOME Additional orthogonal evidence for PNEs was gained by annotating them with RNA level transcriptome data. These expression annotated predicted peptides (PNE^{exp}) decreased the total amount of PNE by half (49±8 % of PNE) to an average of 118±40 predicted peptides.

PROTEOME In order to gain combined additional evidence for PNEs on transcriptome and proteome level, PNE were annotated with log2-intensities of shotgun proteome data of HCCs when available (n=7). For a total of 159 PNE, corresponding source proteins (n=33) were detectable (on average for 22.7±21.1 PNE^{prot} per patient). Only in one patient, no such evidence for PNE-associated proteins was found (HCC034), whereas only a fraction of 9.8±8.6% of PNE had additional evidence on shotgun proteome level.

LIGANDOME To assess the existence of mutated HLA ligands with confidence, uHPLC-coupled MS/MS was employed to identify naturally presented HLA ligands from HCC and adjacent-benign liver tissue. (Figure 6.2 C) These analyses yielded on average 1403±621 HLA-I presented peptides from HCC (FDR 5%, length 8-11 AA) and 1159±525 peptides from the non-malignant liver samples (FDR 5%, length 8-11 AA). After prediction of binding affinities for the respective patient HLA allotypes, an average of 1026±451 peptides per tumor and 867±450 peptides per non-malignant liver sample remained. On average, 51±11% of these HLA ligands were shared between matched malignant and benign liver samples. 73±10% of all MS-detected peptides identified on tumor tissue were predicted binders. Similar numbers were observed for non-malignant tissue (72%±11%). The amount of shared predicted binders of MS-detected peptides between matched malignant and benign liver samples averaged at 58%±12%. Importantly, we did not find evidence for naturally presented mutated HLA ligands in HCC, independent of filtering criteria. However, we were able to identify one wild-type HLA ligand (WT^{lig}) corresponding to PNE in two patients each.

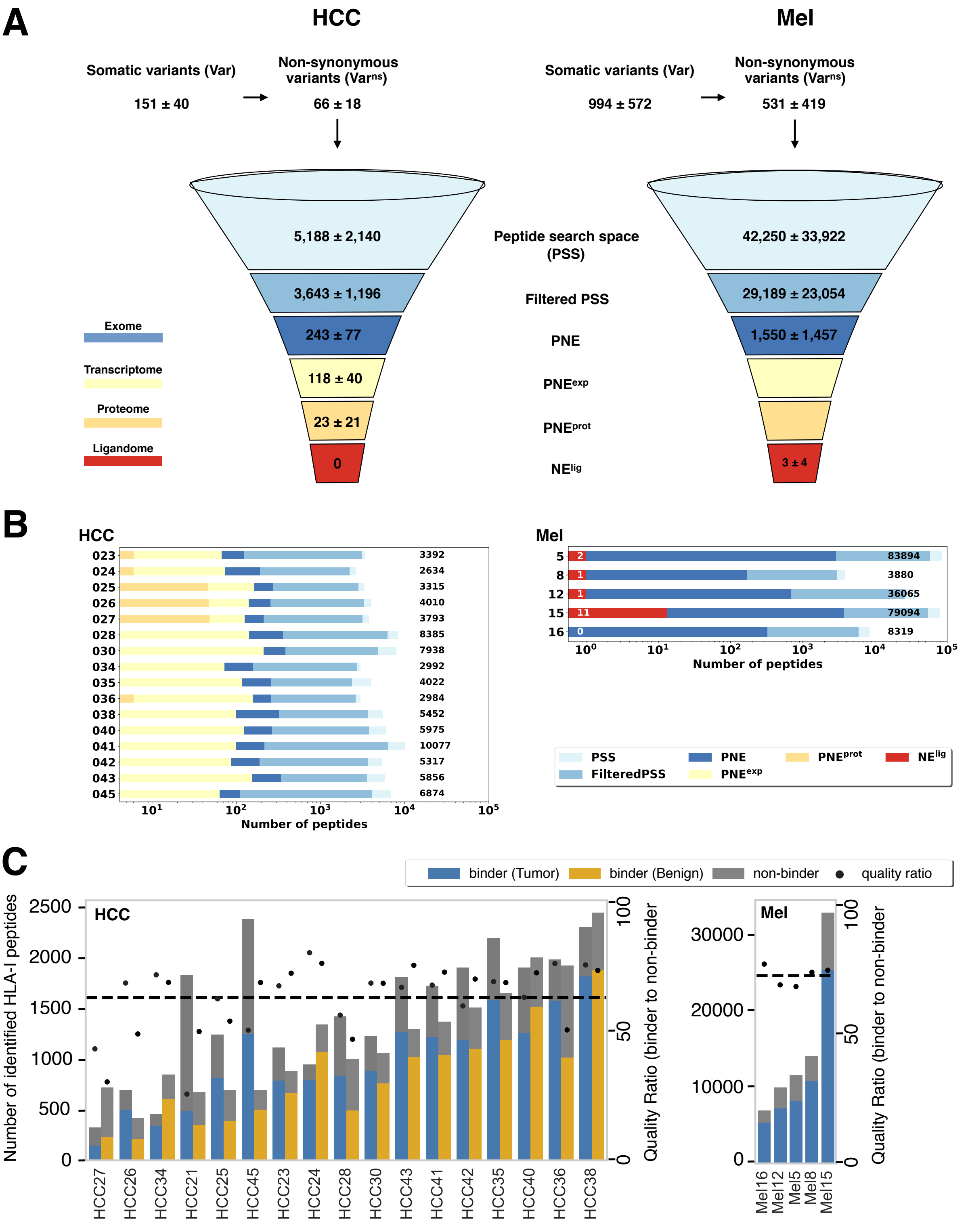


Figure 6.2: Evidence for predicted neoepitopes on different Omics levels. A) Somatic (Var) and non-somatic variants (Var^{ns}), peptide search space (PSS), resulting predicted neoepitopes (PNE) and their evidence on the various Omics levels transcriptome (PNE^{exp}), proteome (PNE^{prot}) and HLA ligandome (NE^{lig}) are compared for the HCC and the Mel dataset. The numbers represent the mean ± standard deviation. B) Numbers of peptides per patient on all included omics leves resulting from processing with the here used insilico neoepitope identification pipeline. Total counts based on the peptide search space are annotated in black, and NE^{lig} if available are shown in red. (Figure A and B adapted from the original version by Christopher Mohr) C) HLA ligandomics peptide yields of the HCC cohort (left) in contrast to Mel (right) for tumor (blue) and benign (yellow) tissue samples. Predicted HLA binder ratio is indicated as second axis on the right and non-binding peptides are added to the bar plot in grey.

Benchmarking HCC and Mel HLA ligandomics datasets

To demonstrate the high sensitivity of our neoepitope identification pipeline, we additionally processed a publicly available data set of five Mel patients as a reference. The amounts of Var^{ns} and PNE show remarkable differences between HCC and Mel. (Figure 6.2) Whereas in two cases, Mel samples showed comparable properties to HCCs with respect to the numbers of Var^{ns} and PNE (Mel8, Mel16), numbers were substantially increased in the majority of Mel samples (Mel5, Mel12, Mel15). This corresponds to an average of 531 Var^{ns} in Mel in comparison to 66 in HCC, resulting in a eight-fold larger peptide search space (PSS) in Mel. Derived PNE amounted to an average of 243 in HCC in contrast to 1.550 in the Mel data set.

Assessing respective data on a per patient basis (Figure 6.2 B), it is obvious that the HCC data set is more homogenous (PSS: 2.500 to 10.000, PNE: 111 to 382), whereas in Mel the PSS ranges from 4.000 to 84.000 (PNE: 169 to 3717) resulting from a substantially different TMB.

Using our ligandomics identification pipeline, we were able to reprocess the melanoma raw MS data that had been described by Bassani-Sternberg et al. [46]. The ligandomics peptide yields are significanlty lower for our HCC samples than those encountered for Mel (Figure 6.2 C). While we were able to reconfirm all of the neo-epitopes (NE^{lig}) in their MS data set [Mel5 (n=2); Mel8 (n=1); Mel15 (n=8)], we further discovered one additional NE^{lig} for Mel12 and three additional NE^{lig} for Mel15. However, only one NE^{lig} was discovered in a sample (Mel8) with mutational burden properties comparable to our HCC cohort.

Therefore, our comparatively homogenous HCC cohort and their analyzed tumor tissue samples, for which no NE^{lig} were discovered, differ substantially (by at least one order of magnitude) from that of Mel, with regard to the mutational load and the amount of identifiable HLA-presented peptides.

Evidence for mutated proteins on shotgun proteome level

Despite the lack of identification of NE^{lig} in our HCC cohort on HLA ligandome level, we aimed at gathering the best available evidence for the presence of mutated proteins on shotgun proteome level. A tryptic digest of cell lysates was used, knowing that detection of respective variants is difficult, the sensitivity of the technique limited [252] and governed by a variety of influencing factors.

As a result, we detected two somatic mutations on proteome level in HCC025 and HCC026. (Figure 6.3) In HCC025, the variant ALB (K375E) was identified, which was corroborated on both exome (Var) and transcriptome level (Var^{exp}). However, in addtion we detected this Var in non-malignant liver tissue and peripheral blood from the respective patient, possibly explained by the fact that the tumor synthesized a mutated ALB protein secreted into circulation. In contrast, for HCC026, a Var^{exp} in the helicase RECQL (H19R) was verified based on an additional tryptic cleavage site introduced through the gained arginin. This resulted in a truncated protein undetectable in the corresponding non-malignant liver tissue (Figure 6.3 B).

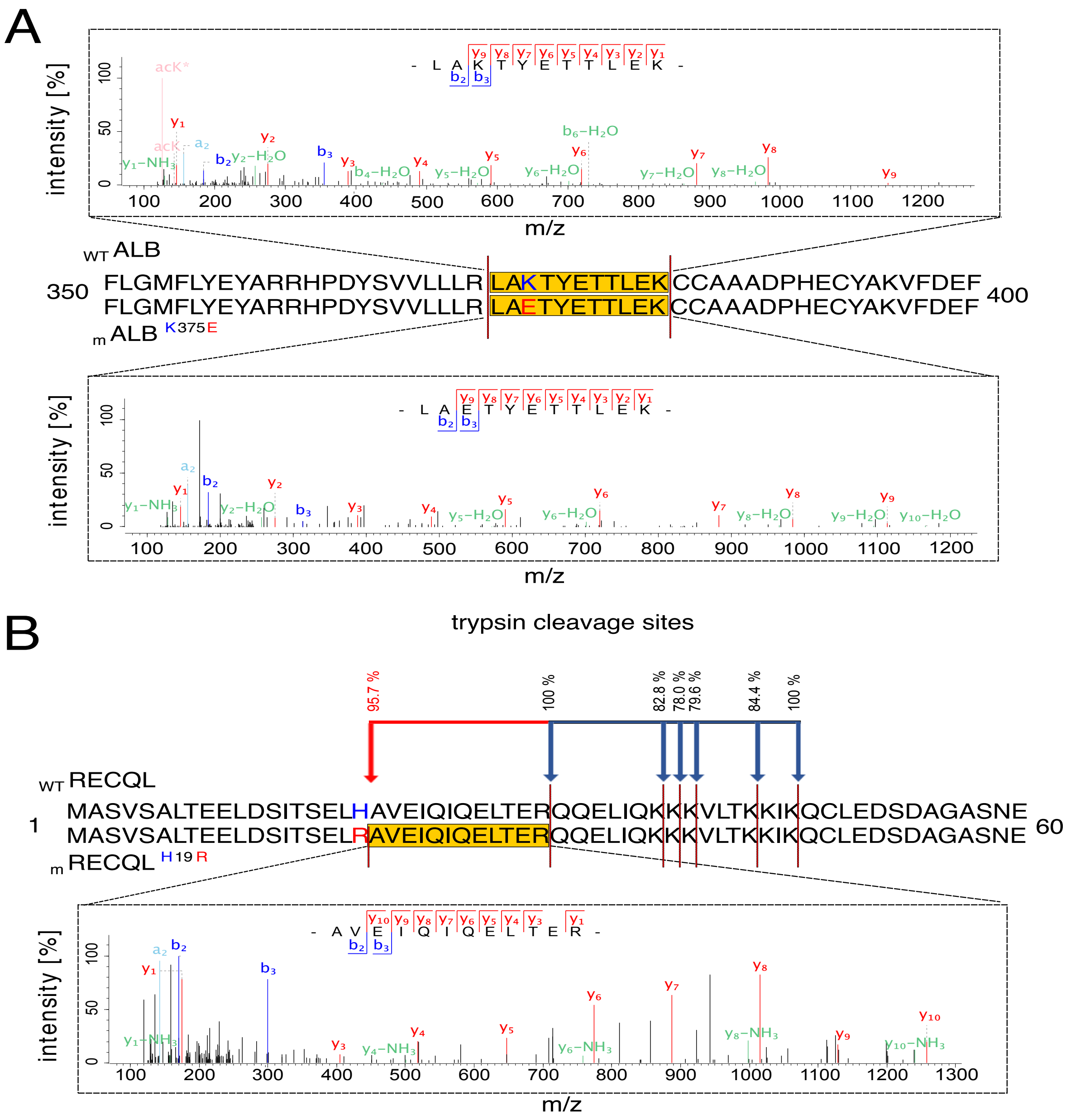


Figure 6.3: Evidence for mutated proteins in the shotgun proteome and database matching. A) Annotated spectra of albumin (ALB) showing sequences of wild-type (LAKTYETTLEK; top) and mutated (LAETYETTLEK; bottom) protein measured by LC-MS/MS. B) Annotated spectra of RecQ like helicase (RECQL) showing sequences of the peptide AVEIQIQELTER resulting from an additional tryptic cleavage side added directly in front of this sequence through a mutation from histidine (H) to arginine (R), evidenced in HCC tissue only.

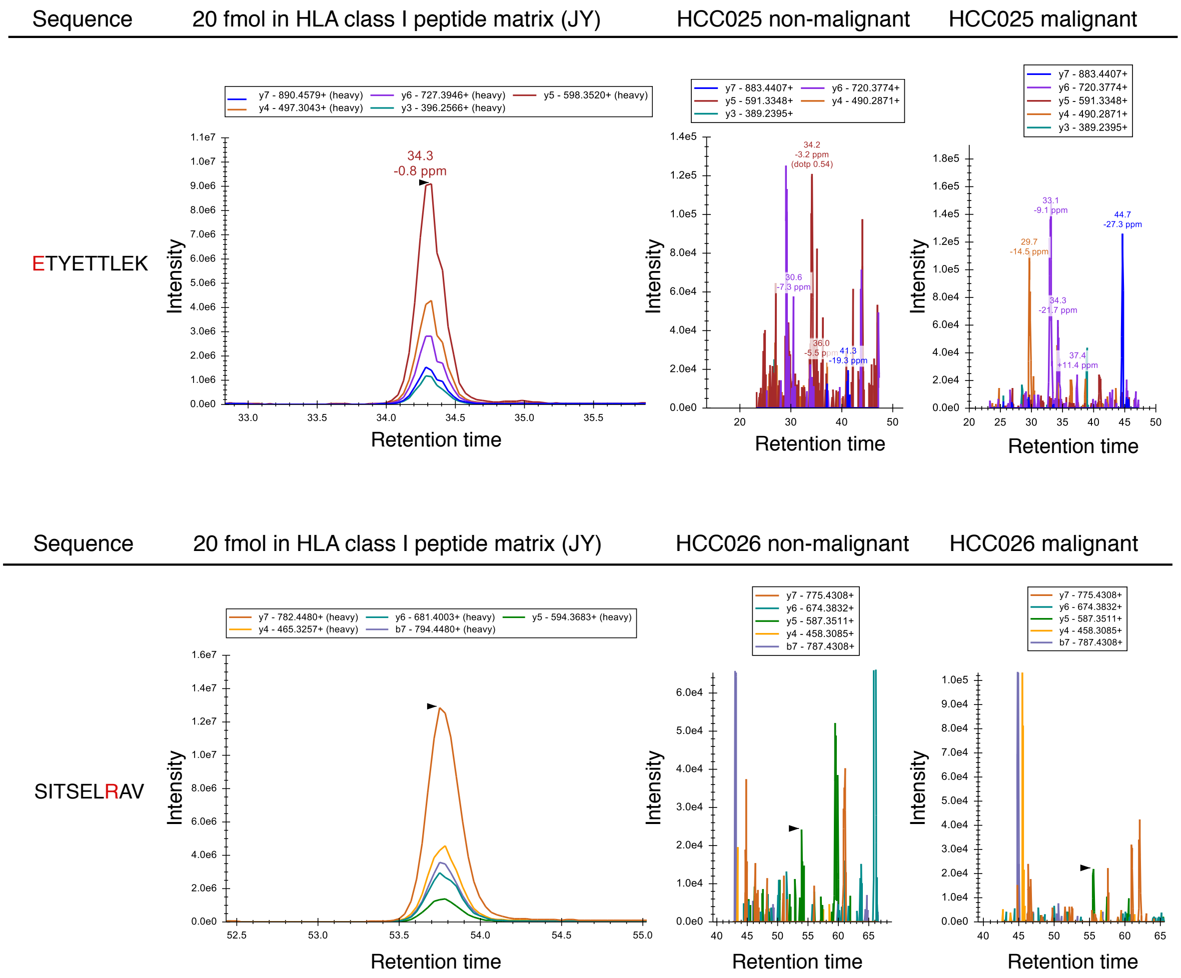


Figure 6.4: Targeted PRM measurement results of the putative PNEs ETYETTLEK and SITSELHAV arising from the mutations ALB K375E and RECQL H19R. Extracted ion chromatograms (XICs) of the corresponding peptide transitions in a matrix of 20 fmol spiked in synthetic peptide into JY cell extract in contrast to untreated non-malignant and malignant tissue samples. No evidence for the presence of the same transition peak group is apparent in any of the tissue samples. (Figure adapted from the original version by Lena Freudenmann)

Targeted mass spectrometry for discovery of mutated HLA ligands

In order to gain addtional confidence in the lack of identifiable NE^{lig} in HCC and exclude technical limitations of our DDA based MS search, targeted MS/MS measurements were employed for multiple promissing candidate NE^{lig}. 17-20 PNE from three HCCs (HCC025-27) were selected for a single-ion monitoring (SIM) approach using heavy isotope-labelled peptides as a reference. Of particular interest in this regard were peptides harbouring the mutations confirmed by proteomics (PNE^{prot}), ALB (HCC025/ ALB K375E) as well as RECQL (HCC026/ RECQL H19R). For HCC025-26, therefore additional PRM measurements targeting the best ranking predicted mutated HLA ligands as well as corresponding wild-type HLA ligands were performed, covering the positions of interest. However, none of these attempts could not confirm any of these PNE^{prot} as a naturally presented HLA ligand (Figure 6.4).

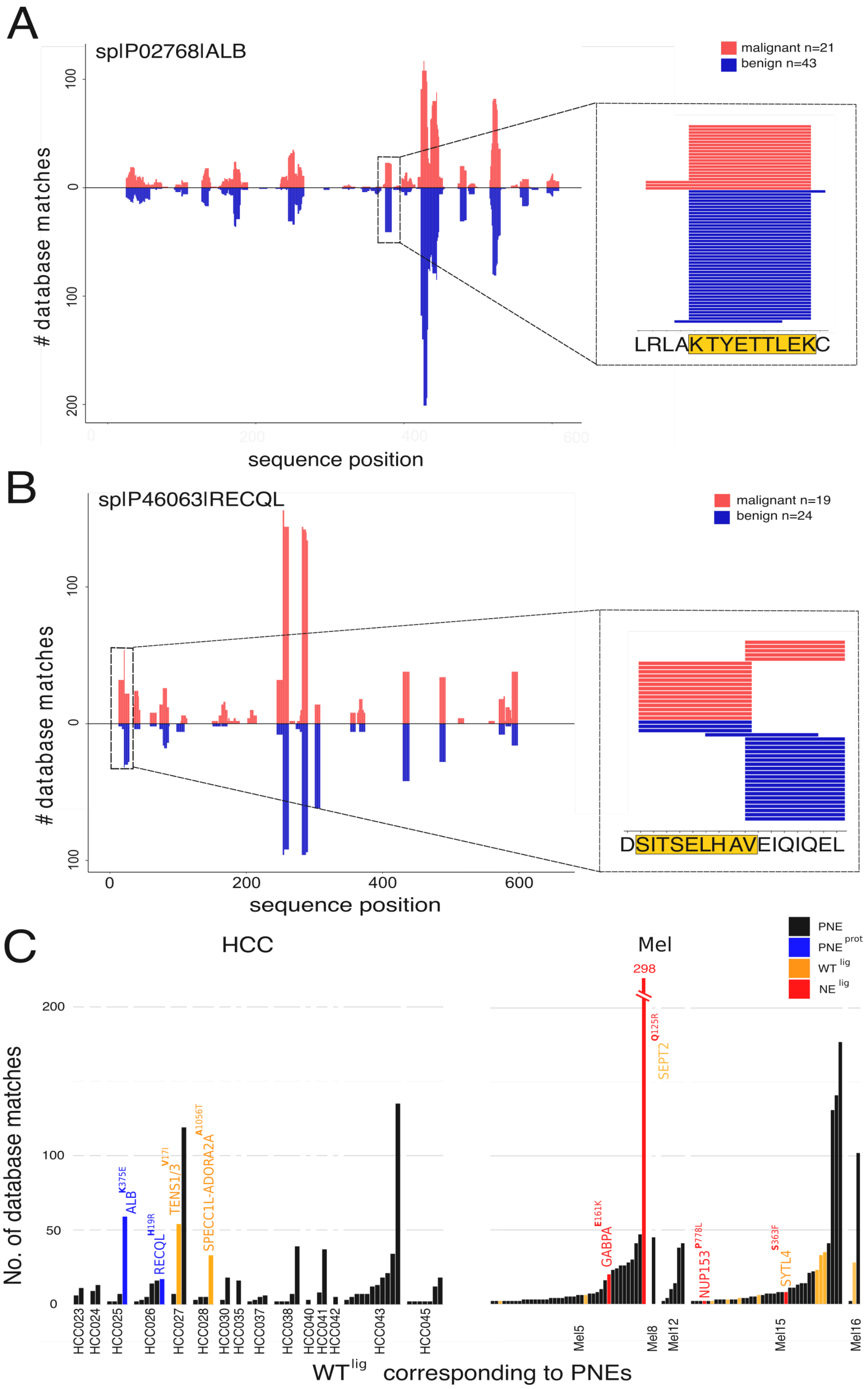


Figure 6.5: A) Database matching of natural HLA ligands with wild-type peptide sequences (with diverse HLA restrictions) covering the exact position evidenced as mutated in ALB. B) Database matching of natural HLA ligands with wild-type peptide sequences (with diverse HLA restrictions) covering the exact position evidenced as mutated in RECQL. C) Number of database matches of wild-type ligands (WT^{lig}) corresponding to predicted mutated neoepitopes (PNE). PNE with additional evidence in HCC and Mel are highlighted - (1) black: wild-type sequence of PNE contained in database; (2) yellow: wild-type sequence peptide corresponding to PNE confirmed in autologous tissue as natural HLA ligand by MS; (3) blue: mutated protein confirmed by shotgun proteomics - PNE^{prot}; (4) red: PNE confirmed as natural HLA ligand by MS - NE^{lig}.

Prioritizing predicted mutated HLA ligands in absence of HLA ligandome evidence

Ultimatly, lacking detection of a mutated HLA ligand does not necessarily equal absence due to several reasons: inter alia 1) the detection limits of the MS instrumentation, 2) ionizability of respective peptides and 3) unkown temporal dynamics of the HLA ligandome landscape. Here we therefore propose a knowledge based approach using previously measured wild-type HLA ligands corresponding to a PNE ($\mathrm{WT}^{\mathrm{lig}}$) as one way for prioritize PNE. Hence, we assume the more frequently a $\mathrm{WT}^{\mathrm{lig}}$ was detected in previous measurements the more likely it is the corresponding $\mathrm{PNE}^{\mathrm{lig}}$ counterpart will exist if the mutation does not negatively impact its HLA binding affinity. To this end, we compared the number of database matches of all $\mathrm{WT}^{\mathrm{lig}}$ in HCC and Mel to our in-house database of HLA ligands measured over the last decades. (Figure 6.5)

The two pinpointed $\mathrm{PNE}^{\mathrm{prot}}$ in ALB (59 matches) and RECQL (17 matches) give rise to the two most frequently found $\mathrm{WT}^{\mathrm{lig}}$ in our database for those particular patients. In addition, nearly all patients carry at least one mutation that could potentially give rise to a PNE whose $\mathrm{WT}^{\mathrm{lig}}$ were measured multiple times in the past. In accordance, some of the directly observed $\mathrm{NE}^{\mathrm{lig}}$ on Mel and their corresponding $\mathrm{WT}^{\mathrm{lig}}$ produced multiple hits in our database, such as GABPA (20 matches), SYTL4 (8 matches), NUP153 (2 matches) and outstandingly SEPT2 (298 matches). In addition, $\mathrm{WT}^{\mathrm{lig}}$ TENS1/3 (54 matches) of patient HCC27 and SPECC1L-ADORA (33 matches) of patient HCC28 were detected in their own ligandome extract, highly favouring the presence of its $\mathrm{NE}^{\mathrm{lig}}$ counterpart as well, although it could not be detected. These 4 ligands and $\mathrm{WT}^{\mathrm{lig}}$ SEPT2 have been documented in the IEDB previously as well. Ultimately, these results question ligandome level detection depth and may prove utility of large available knowledge bases for HLA immunopeptidomics [116].

Alternative targets for immunotherapy of HCC

Nontheless, as mutated HLA ligands are less frequent and more difficult to be identified in HCC than Mel in our approach, other targets with potential therapeutic relevance might be revealed such as HLA ligands from known cancer testis antigens. Hence, we screened our HCC dataset for proteins previously described as cancer-testis antigens (CTAs) [253] and found eight different HLA-I ligands mapping to six CTA.

These few CTA encompass ARMC3 (Q5W041), ATAD2 (Q6PL18), MAEL (Q96JY0), PRAME (P78395), proteins of the SSX family, and TFDP3 (Q5H9I0) (Table 6.2). However, comparing these identified peptides to the recently published resource of naturally presented HLA ligands on benign tissues (HLA Ligand Atlas) [163] revealed that two of these (AYAIIKEEL-ARMC3 and SLLQHLIGL - PRAME) are presented in the healthy state as well. Hence, to avoid autoimmune reactions these two should rather not be considered for targets in HCC immunotherapy.

CTA (UniprotID)	HLA peptide	HLA restriction	Patient ID
ARMC3 (Q5W041)	EQIEDLAKY	A*26:01	HCC045
ATAD2 (Q6PL18)	AYAIIKEEL	A*24:02	HCC023
ATAD2 (Q6PL18)	AEFRTNKTL	B*44:03	HCC045
MAEL (Q96JY0)	MVVLDAGRY	A*26:01	HCC045
PRAME (P78395)	SLLQHLIGL	B*08:01	HCC041
SSX1 (Q16384)	AFDDIATYF	C*04:01	HCC035
SSX *	RLRERKQLV	B*08:01	HCC041
TFDP3 (Q5H9I0)	EVVGELVAKF	A*26:01	HCC045

Table 6.2: CTAs as alternative targets for immunotherapy of HCC: Identified peptides that match to known CTAs such as SSX1 (Q16384); SSX2 (Q16385); SSX3 (Q99909); SSX4 (O60224); SSX6 (Q7RTT6); SSX7 (Q7RTT5); SSX9 (Q7RTT3)

6.4 *Discussion*

In this work we have aimed to discover genome mutations in a small cohort of HCC patients that would give rise to HLA ligands presenting these mutations to their immunesystem and that would be amendable for personalized immunotherapy. By choosing a mutli-omics approach to search the various levels of biological complexity involving the genome, transcriptome, proteome and HLA ligandome we were seeking to trace mutations of prioritized therapeutic relevance. While we were not able to get direct experimental evidence for the presentation of particular mutated HLA ligands in any of the HCC patients, despite in depth orthogonal search including targeted MS approaches, we managed to pinpoint mutations present in several omics layers and found a number of promissing PNE^{lig} candidates for immunotherapy. In fact, almost all HCC patients revealed to express RNA sequences (PNE^{exp}) and translate proteins (PNE^{prot}) of multiple of those genes affected through cancer somatic mutations and that were predicted to be HLA-presented. In addition, some of the mutations are in regions that were previously directly observed to be presented on HLA peptides in other samples (WT^{lig}). Finally, when searching for non-mutated immunological targets we observe a few peptides matching to known cancer testis antigens that might provide an alternative to neoepitope based immunotherapy to some patients.

When comparing our approach and computational analysis pipeline to the recently published study of Mel [46], we were able to reproduce the results and indeed identify mutated HLA ligands. Hence, we derive several explanations for the discovery of mutated neoepitopes in Mel but not in HCC. Firstly, Mel is among the tumor entities with high mutational load in contrast to HCC [229], resulting in several magnitudes higher amounts of possible neoepitope candidates (PNE) in Mel than HCC. Thus tumor biology and mutational load are fundamental underlying factors when considering suitability of immunotherapeutic approaches. In addition, the HLA immunopeptidome yields identified in the Mel data set in contrast to the HCC cohort differ by several magnitudes as well, allowing to identify a much greater amount of HLA-presented peptides per patient and mapping them to the mutated search space. Finally, a much larger amount of the predicted neoepitopes in the Mel cohort have non-mutated counter parts (WT^{lig}) that we have previously identified in other measurements, increasing the odds of finding this ligand in an MS experiment.

In conclusion, we figure that immunotherapy is a less promissing option in tumors of lower TMB such as HCC, because NE^{lig} are a probabilstically rare encounter rather than in entities of high TMB such as Mel. However, since we found evidence on various omics layers for tumor somatic mutations and were able to identify peptides from non-mutated cancer testis antigens this study should not hinder the enthusiasm for a wider scope in the exploration of immunological targets for HCC. The therapeutic window for late stage HCC patients is small also for other approaches [25] and small amounts of mutations might be suitable enough to induce personalized T-cell therapy [216]. Ultimately, other targets for example from alternatively spliced [48] or cryptic translation events [197] may be found in the future giving rise to a greater selection of suitable targets.

CHAPTER 7

Conclusion

Conclusion

This thesis describes the computational analysis of HLA immunopeptidomics mass spectrometry data and discusses its relevance for several topics ranging from basic research in immunology to clinical applications. While half of the research work is devoted to the computational development and comparative analysis of mass spectrometry methodology to measure the HLA bound immunopeptidome, the other half focuses on the biological discovery and its significance for general human and tumor immunology.

methodological advancements

In Chapter 3, MHCquant [123] - a novel automated computational pipeline to process large sets of mass spectrometry measurements was described. The pipeline is containerized and can be efficiently executed on HPC systems, which allowed to analyse more than 1400 MS runs in about 26 hours. This facilitates immunopeptidome research enormously, since at the time of this thesis this was an endeavour that would have taken weeks to months if carried out manually using existing commercial or other software solutions. Moreover, the automated analysis also provides a way to work more reproducibly and enables easy and coherent reprocessing of previous MS experiments with improved methods or newer versions at a later stage. In addition we found that the employed open-source search engine Comet [82] in combination with the Percolator algorithm [88] for FDR scoring vastly outperformed other search engines with respect to the number of peptides discovered in each MS experiment. Confidence in the correctness of the additionally identified peptides was gained by assessing their properties with respect to their chromatographic retention time behaviour and sequence motifs that are predicted to fit well to corresponding HLA allotypes. Finally, when employing the workflow to a data set of Melanoma cancer samples [46], we retrieved three novel mutated neoepitope peptide identifications that were previously not found. As evidence of neoepitopes is of tremendous interest to cancer immunotherapy, this clearly demonstrates the potential that new MS data processing methods have for the field of immunopeptidomics, which has been pointed out by other researchers in the field as well [15, 22, 113].

In the following Chapter 4, the application of data-independent acquisition mass spectrometry (DIA-SWATH MS) - a recently developed technological improvement that had shown great advantages for clinical applications [155] - was investigated for application to immunopeptidome analysis. As part of this work DIAproteomics - an additional automated and containerized computational processing pipeline for DIA data was constructed. In contrast to DDA - the commonly applied method for measuring MS samples, we confirmed findings for proteomics studies [69, 153, 154] that DIA achieves a much greater reproducibility among replicate measurements and a higher share of peptide identifications across samples. As DIA is able to trigger MS2 fragmentation of nearly all precursors present in a sample, it can overcome the shortcoming of DDA's inadequate sampling. As a result, this enables the additional identification of lower abundant precursor ions in particular. However, even when using DIA the total number of shared peptide identifications across samples of different tissues and patients is still quite low. While this results is influenced by the incomplete immunpurification procedure, it indicates that there is a great biological variance present across samples and this should be taken into account when designing general immunotherapeutic strategies for T-cell therapy or peptide vaccination [44] of multiple patients. Nevertheless, same tissues of different patients share commonly presented peptides, when focusing on a particular HLA allele, which are found with higher sensitivity when applying the DIA methodology.

IMMUNOLOGICAL DISCOVERIES

Having the bioinformatic pipelines at hand to coherently process large amounts of immunopeptidomics data, in a highly collaborative project we set out to acquire and analyse samples of human tissues. As a result, Chapter 5 highlights the findings of the currently largest set of immunopeptidomics samples from non-malignant human research autopsy body donors. Similar to breakthrough discoveries such as the sequencing of the human genome [173, 174] or the mapping of the human proteome [176–178], we anticipate that with this research we have contributed to the decoding of the human immunopeptidome as another orthogonal biological layer as human tissue-dependent antigen processing has never been explored as extensive before. The acquired set covers 29 distinct tissues including the thymus, 51 HLA-I and 81 HLA-II individual HLA allotypes and all data was released into a publically available webservice free to download and explore for the scientific community [163]. In total, we observed a great variance in the amount of HLA peptides identified, their sequences and source proteins. Sources of variability were coming from the different tissues analysed, the characteristics of donors, their respective HLA allotypes and the technical variability ot the immunpurification and DDA measurements. While the donor characteristics exceeded the inter-tissue variability, we were able to decipher a small extend of tissue specific effects present in the immunopeptidome. Most significantly, we observed different amounts of peptides presented by the various tissues, gradually separating them into high and low yielding tissues. This could be linked to immune related gene expression as well as central functions of these tissues in the immune system or high degrees of immune infiltrating cells. In addition, the small sets of source proteins of tissue specific HLA ligands do reflect organ specific biology when querying their gene ontologies. Ultimately, we researched the acquired samples with a *de novo* peptide identification method [23] and were able to validate the presence of cryptic peptides stemming from non-canonical regions of the genome. This highlights the importance of investigating the healthy state immunopeptidome, as these peptides had previously been claimed to occur in tumor tissues only. [254] In conclusion, the assembled dataset is therefore a great benefit for the research of various aspects of human immunology ranging from basic biological questions to tumor immunotherapy.

At last, in Chapter 6 the same methodology was applied in another collaborative effort to the HLA immunopeptidome of hepatocellular carcinoma with the aim to discover tumor exclusive antigens suitable for immunotherapies such as neoantigen-based vaccines. By choosing a multi-omics approach through the combination with NGS exome and transcriptom sequencing as well as proteomics we seeked to get a more holistic understanding of tumor antigen presentation in particular the presentation of mutated neoepitopes. [185] In contrary to our expectations, we were not able to find evidence for mutated HLA ligands. However, we were able to pinpoint and prioritize tumor somatic mutations in the individual patients and annotate them with evidence on different omics layers. When comparing these results with the results of a previously published melanoma study [46] we concluded that technical effects as well as the biology of tumor entities have influenced the outcome. In fact, the Melanoma samples had significantly higher immunopeptidomics yields as well as mutational loads that were several magnitudes higher in contrast to the hepatocellular carcinoma patients. As HCC is ranging in the moderatly to low mutated cancer entities [229] we thus conclude that neoepitope based immunotherapy might not be suitable for this type of solid tumors. Nevertheless, other for example non-mutated tumor antigens might be found in HCC that prove to be useful for immunotherapies in the future.

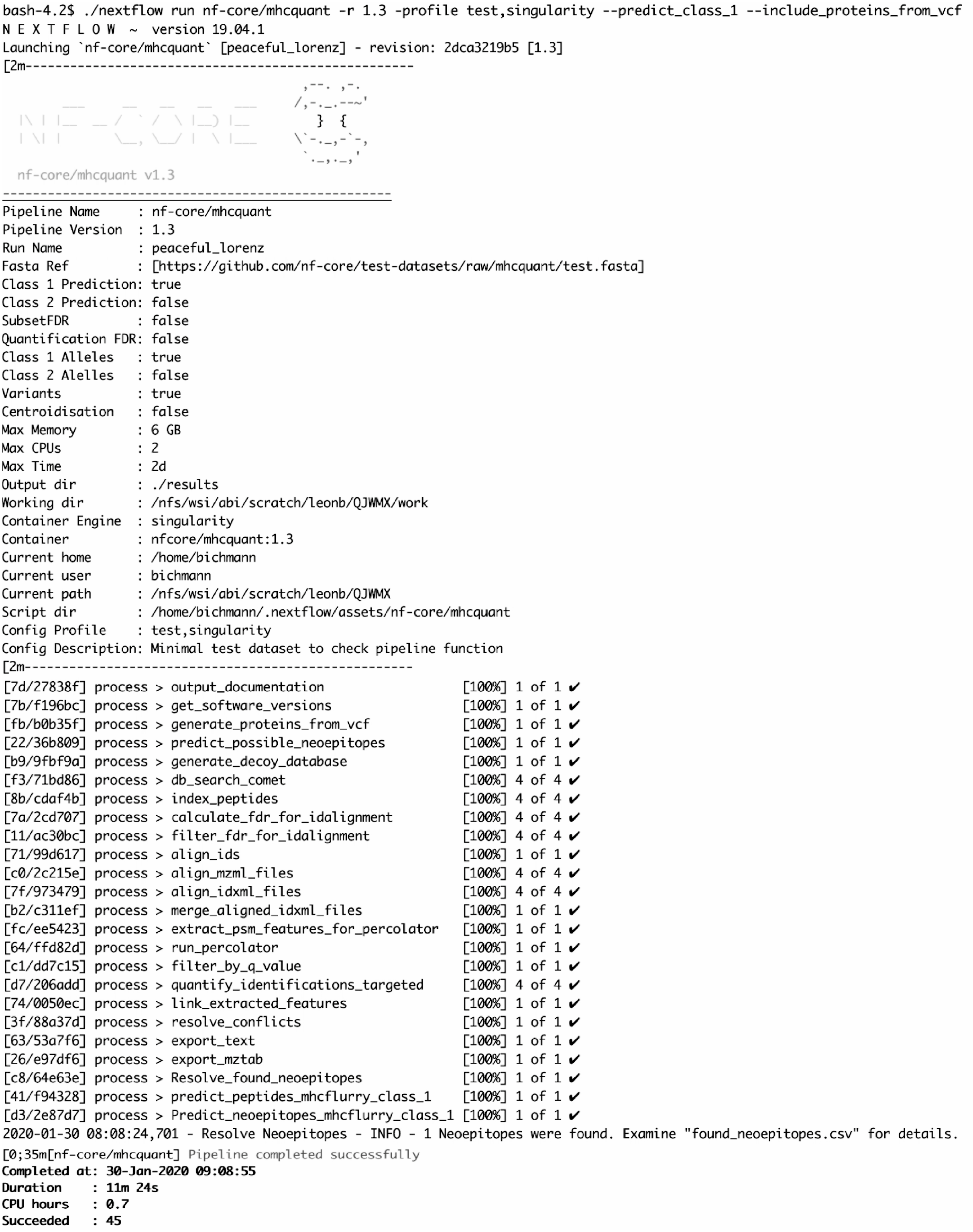
OUTLOOK

Most likely, the investigation of the immunopeptidome for vaccination and immunotherapy will be improved upon in the future with advancing technological developments in the field. Fractionation in combination with high-throughput immunopurification systems [107] as well as ion mobility based separation for mass spectrometry [151] are promissing technologies to increase the peptide yield per sample on the experimental side. On the computational side prediction of fragment intensities [80, 81] and their integration into search engines, as well as FDR considerations [22, 149] and improvements on algorithms for the deconvolution of DIA data [70, 157] are going to have an impact on the sensitivity of discoveries made from experiments. For the translation into clinical practice additional hurdles are still going to have to be considered such as the selection and prioritization of antigens from genomic alterations [255], adjuvantation and feasability logistics [256]. With the results of this thesis work, I hope to have demonstrated the capabilities of computational methodologies in order to facilitate and improve insights into human immunology and the advancement immunotherapy and I am looking forward to follow up on future developments in the field.

APPENDIX A

Additional information on bioinformatics workflows

A.1 *MHCquant*



```
bash-4.2$ ./nextflow run nf-core/mhcquant -r 1.3 -profile test,singularity --predict_class_1 --include_proteins_from_vcf
N E X T F L O W  ~  version 19.04.1
Launching `nf-core/mhcquant` [peaceful_lorenz] - revision: 2dca3219b5 [1.3]
[2m----------------------------------------------------
                                        ,--. ,-.
          ___     __   __   __   ___   /,-._.--~'
    |\ | |__  __ /  ` /  \ |__) |__       }  {
    | \| |       \__, \__/ |  \ |___   \`-._,-`-,
                                        `._,._,'
  nf-core/mhcquant v1.3
----------------------------------------------------
Pipeline Name     : nf-core/mhcquant
Pipeline Version  : 1.3
Run Name          : peaceful_lorenz
Fasta Ref         : [https://github.com/nf-core/test-datasets/raw/mhcquant/test.fasta]
Class 1 Prediction: true
Class 2 Prediction: false
SubsetFDR         : false
Quantification FDR: false
Class 1 Alleles   : true
Class 2 Alelles   : false
Variants          : true
Centroidisation   : false
Max Memory        : 6 GB
Max CPUs          : 2
Max Time          : 2d
Output dir        : ./results
Working dir       : /nfs/wsi/abi/scratch/leonb/QJWMX/work
Container Engine  : singularity
Container         : nfcore/mhcquant:1.3
Current home      : /home/bichmann
Current user      : bichmann
Current path      : /nfs/wsi/abi/scratch/leonb/QJWMX
Script dir        : /home/bichmann/.nextflow/assets/nf-core/mhcquant
Config Profile    : test,singularity
Config Description: Minimal test dataset to check pipeline function
[2m----------------------------------------------------
[7d/27838f] process > output_documentation                [100%] 1 of 1 ✔
[7b/f196bc] process > get_software_versions               [100%] 1 of 1 ✔
[fb/b0b35f] process > generate_proteins_from_vcf          [100%] 1 of 1 ✔
[22/36b809] process > predict_possible_neoepitopes        [100%] 1 of 1 ✔
[b9/9fbf9a] process > generate_decoy_database             [100%] 1 of 1 ✔
[f3/71bd86] process > db_search_comet                     [100%] 4 of 4 ✔
[8b/cdaf4b] process > index_peptides                      [100%] 4 of 4 ✔
[7a/2cd707] process > calculate_fdr_for_idalignment       [100%] 4 of 4 ✔
[11/ac30bc] process > filter_fdr_for_idalignment          [100%] 4 of 4 ✔
[71/99d617] process > align_ids                           [100%] 1 of 1 ✔
[c0/2c215e] process > align_mzml_files                    [100%] 4 of 4 ✔
[7f/973479] process > align_idxml_files                   [100%] 4 of 4 ✔
[b2/c311ef] process > merge_aligned_idxml_files           [100%] 1 of 1 ✔
[fc/ee5423] process > extract_psm_features_for_percolator [100%] 1 of 1 ✔
[64/ffd82d] process > run_percolator                      [100%] 1 of 1 ✔
[c1/dd7c15] process > filter_by_q_value                   [100%] 1 of 1 ✔
[d7/206add] process > quantify_identifications_targeted   [100%] 4 of 4 ✔
[74/0050ec] process > link_extracted_features             [100%] 1 of 1 ✔
[3f/88a37d] process > resolve_conflicts                   [100%] 1 of 1 ✔
[63/53a7f6] process > export_text                         [100%] 1 of 1 ✔
[26/e97df6] process > export_mztab                        [100%] 1 of 1 ✔
[c8/64e63e] process > Resolve_found_neoepitopes           [100%] 1 of 1 ✔
[41/f94328] process > predict_peptides_mhcflurry_class_1  [100%] 1 of 1 ✔
[d3/2e87d7] process > Predict_neoepitopes_mhcflurry_class_1 [100%] 1 of 1 ✔
2020-01-30 08:08:24,701 - Resolve Neoepitopes - INFO - 1 Neoepitopes were found. Examine "found_neoepitopes.csv" for details.
[0;35m[nf-core/mhcquant] Pipeline completed successfully
Completed at: 30-Jan-2020 09:08:55
Duration    : 11m 24s
CPU hours   : 0.7
Succeeded   : 45
```

Figure A.1: Commandline execution of the MHCquant Nextflow implementation 1.3.0 using the integrated small test data set.

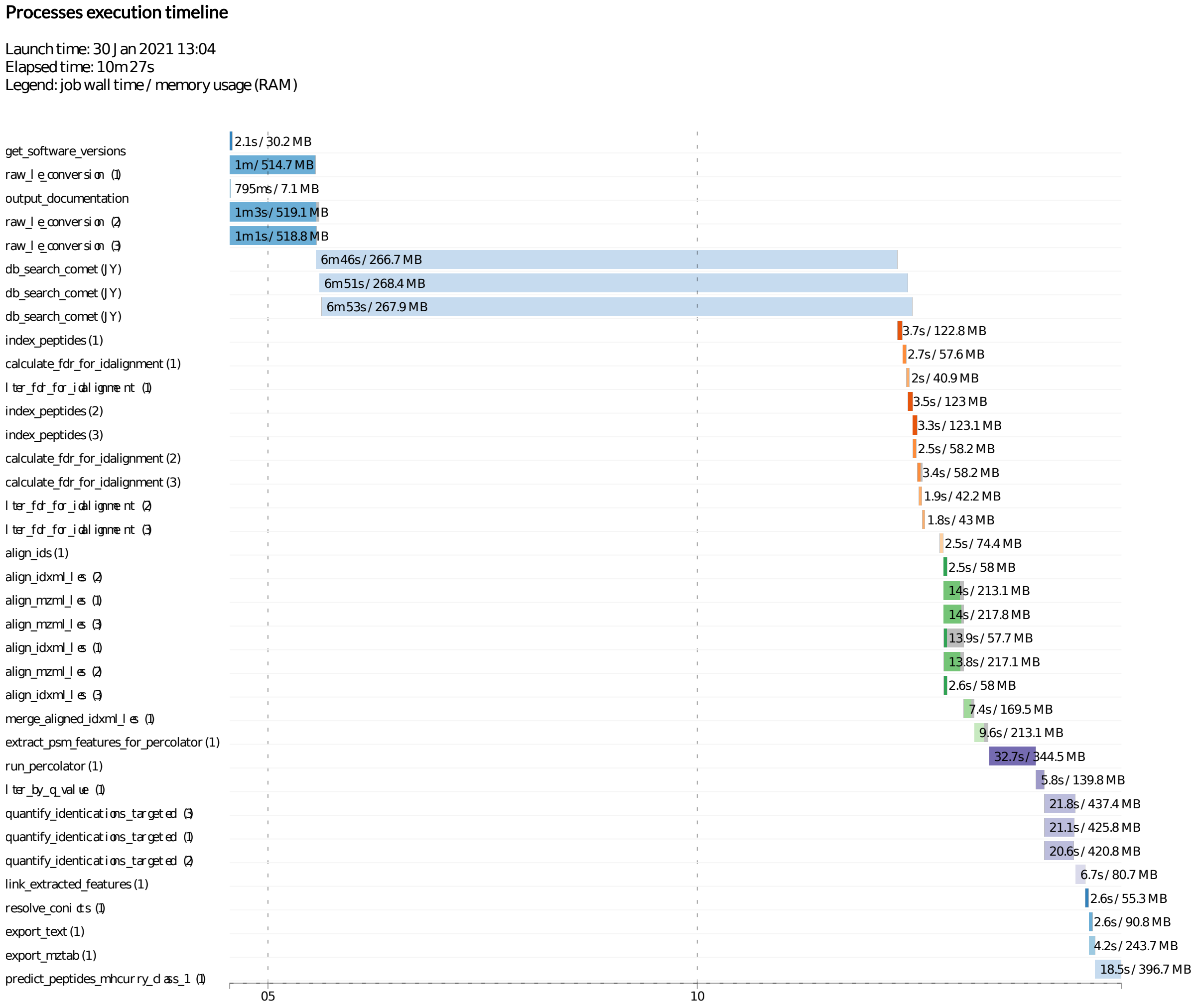


Figure A.2: The runtime of the MHCquant nextflow implementation was tested by a set of three JY standard replicate measurements on a 28 core HPC node of the de.NBI cloud infrastructure: Memory usage in Gigabytes for each process step and runtime in minutes for each process step are listed.

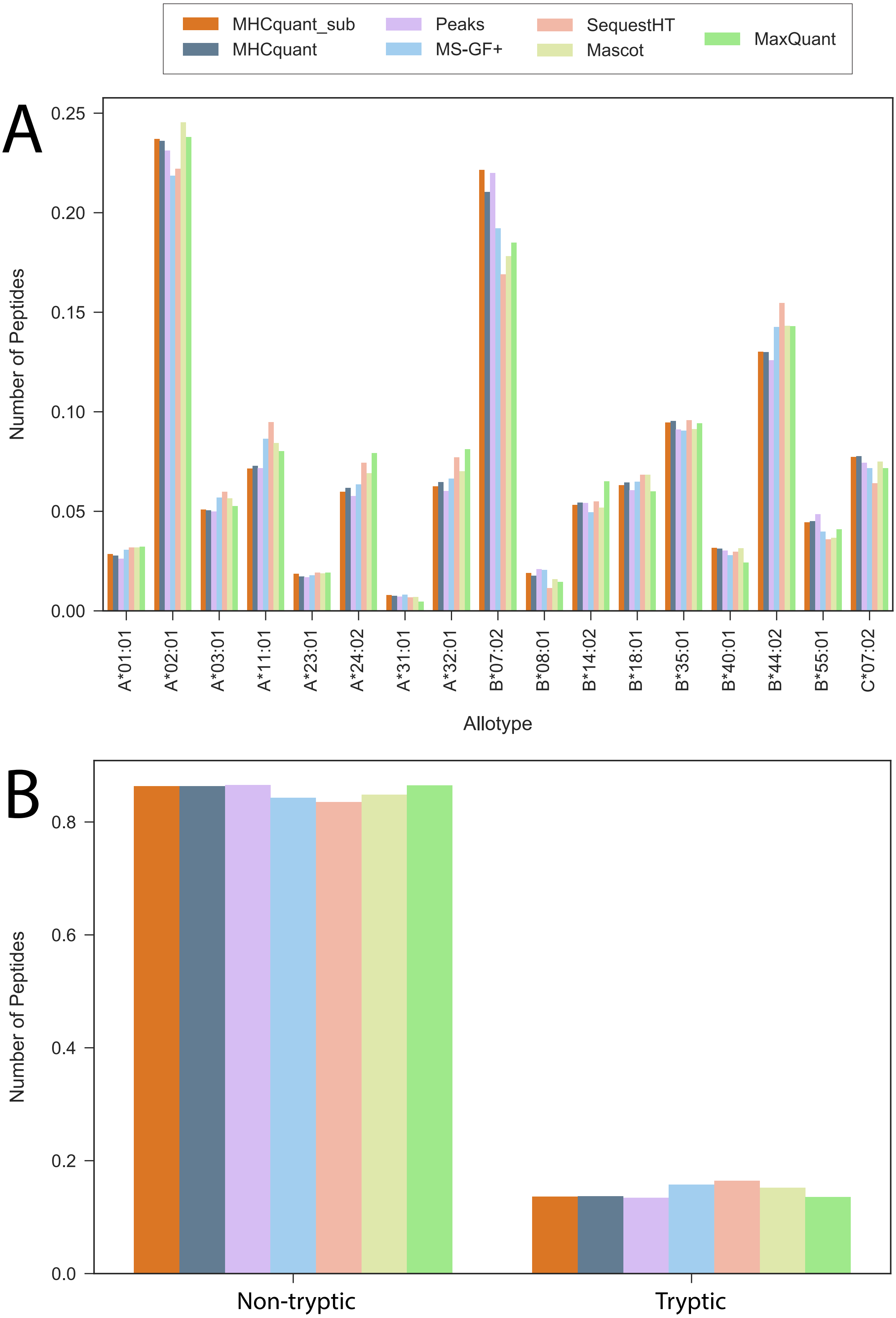


Figure A.3: Allele bias and bias to tryptic vs. non-tryptic peptide discoveries: The percentages of identified peptides matching uniquely to A) one allele, B) tryptic or non-tryptic cleavage specificity compared between search engines.

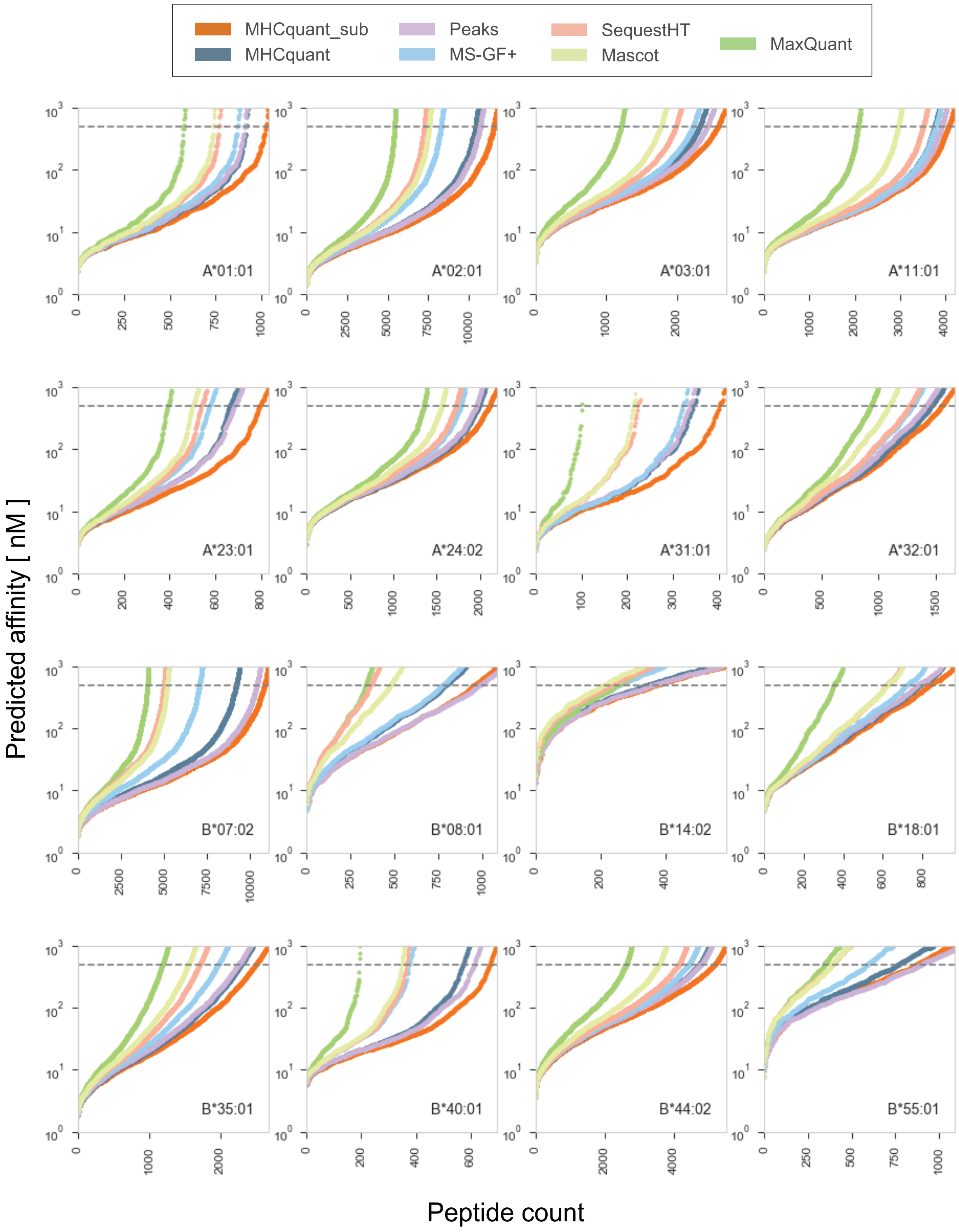
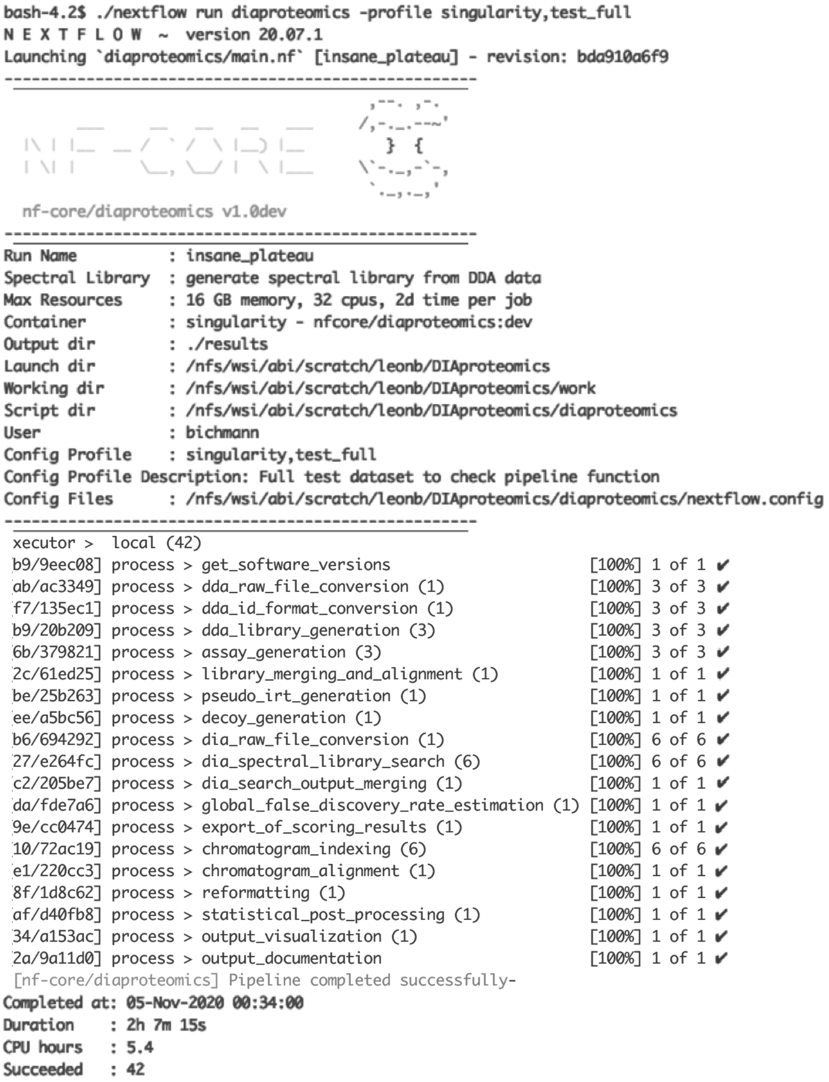


Figure A.4: Absolute numbers of identified peptides by each search engines tested in the benchmark for different alleles and their predicted affinities.

A.2 *DIAproteomics*



```
bash-4.2$ ./nextflow run diaproteomics -profile singularity,test_full
N E X T F L O W  ~  version 20.07.1
Launching `diaproteomics/main.nf` [insane_plateau] - revision: bda910a6f9
----------------------------------------------------
                                        ,--. ,-.
          ___     __   __   __   ___   /,-._.--~'
    |\ | |__  __ /  ` /  \ |__) |__       }  {
    | \| |       \__, \__/ |  \ |___   \`-._,-`-,
                                        `
                                         `._,._,'
  nf-core/diaproteomics v1.0dev
----------------------------------------------------
Run Name          : insane_plateau
Spectral Library  : generate spectral library from DDA data
Max Resources     : 16 GB memory, 32 cpus, 2d time per job
Container         : singularity - nfcore/diaproteomics:dev
Output dir        : ./results
Launch dir        : /nfs/wsi/abi/scratch/leonb/DIAproteomics
Working dir       : /nfs/wsi/abi/scratch/leonb/DIAproteomics/work
Script dir        : /nfs/wsi/abi/scratch/leonb/DIAproteomics/diaproteomics
User              : bichmann
Config Profile    : singularity,test_full
Config Profile Description: Full test dataset to check pipeline function
Config Files      : /nfs/wsi/abi/scratch/leonb/DIAproteomics/diaproteomics/nextflow.config
----------------------------------------------------
 xecutor >  local (42)
 b9/9eec08] process > get_software_versions                    [100%] 1 of 1 ✔
 ab/ac3349] process > dda_raw_file_conversion (1)              [100%] 3 of 3 ✔
 f7/135ec1] process > dda_id_format_conversion (1)             [100%] 3 of 3 ✔
 b9/20b209] process > dda_library_generation (3)               [100%] 3 of 3 ✔
 6b/379821] process > assay_generation (3)                     [100%] 3 of 3 ✔
 2c/61ed25] process > library_merging_and_alignment (1)        [100%] 1 of 1 ✔
 be/25b263] process > pseudo_irt_generation (1)                [100%] 1 of 1 ✔
 ee/a5bc56] process > decoy_generation (1)                     [100%] 1 of 1 ✔
 b6/694292] process > dia_raw_file_conversion (1)              [100%] 6 of 6 ✔
 27/e264fc] process > dia_spectral_library_search (6)          [100%] 6 of 6 ✔
 c2/205be7] process > dia_search_output_merging (1)            [100%] 1 of 1 ✔
 da/fde7a6] process > global_false_discovery_rate_estimation (1) [100%] 1 of 1 ✔
 9e/cc0474] process > export_of_scoring_results (1)            [100%] 1 of 1 ✔
 10/72ac19] process > chromatogram_indexing (6)                [100%] 6 of 6 ✔
 e1/220cc3] process > chromatogram_alignment (1)               [100%] 1 of 1 ✔
 8f/1d8c62] process > reformatting (1)                         [100%] 1 of 1 ✔
 af/d40fb8] process > statistical_post_processing (1)          [100%] 1 of 1 ✔
 34/a153ac] process > output_visualization (1)                 [100%] 1 of 1 ✔
 2a/9a11d0] process > output_documentation                     [100%] 1 of 1 ✔
 [nf-core/diaproteomics] Pipeline completed successfully-
Completed at: 05-Nov-2020 00:34:00
Duration    : 2h 7m 15s
CPU hours   : 5.4
Succeeded   : 42
```

Figure A.5: Commandline execution of the DIAproteomics Nextflow implementation v.1.1.0 using the integrated large test data set.

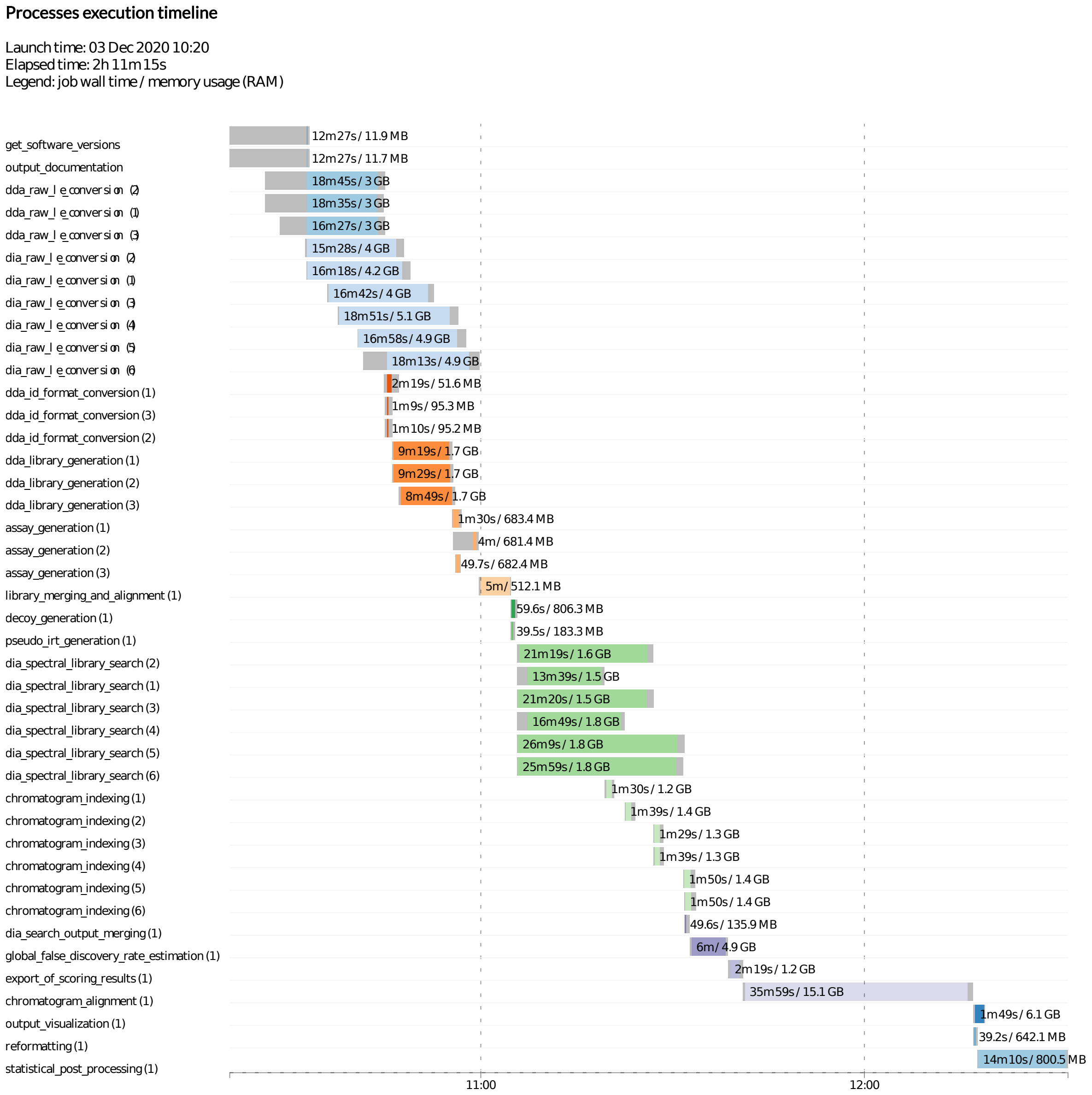


Figure A.6: Detailed overview on run times and memory usage of all integrated steps of the DIAproteomics pipeline v. 1.0 when processing the PRIDE dataset PXD003179 on the Amazon webservice cloud infrastructure.

Reproducibility

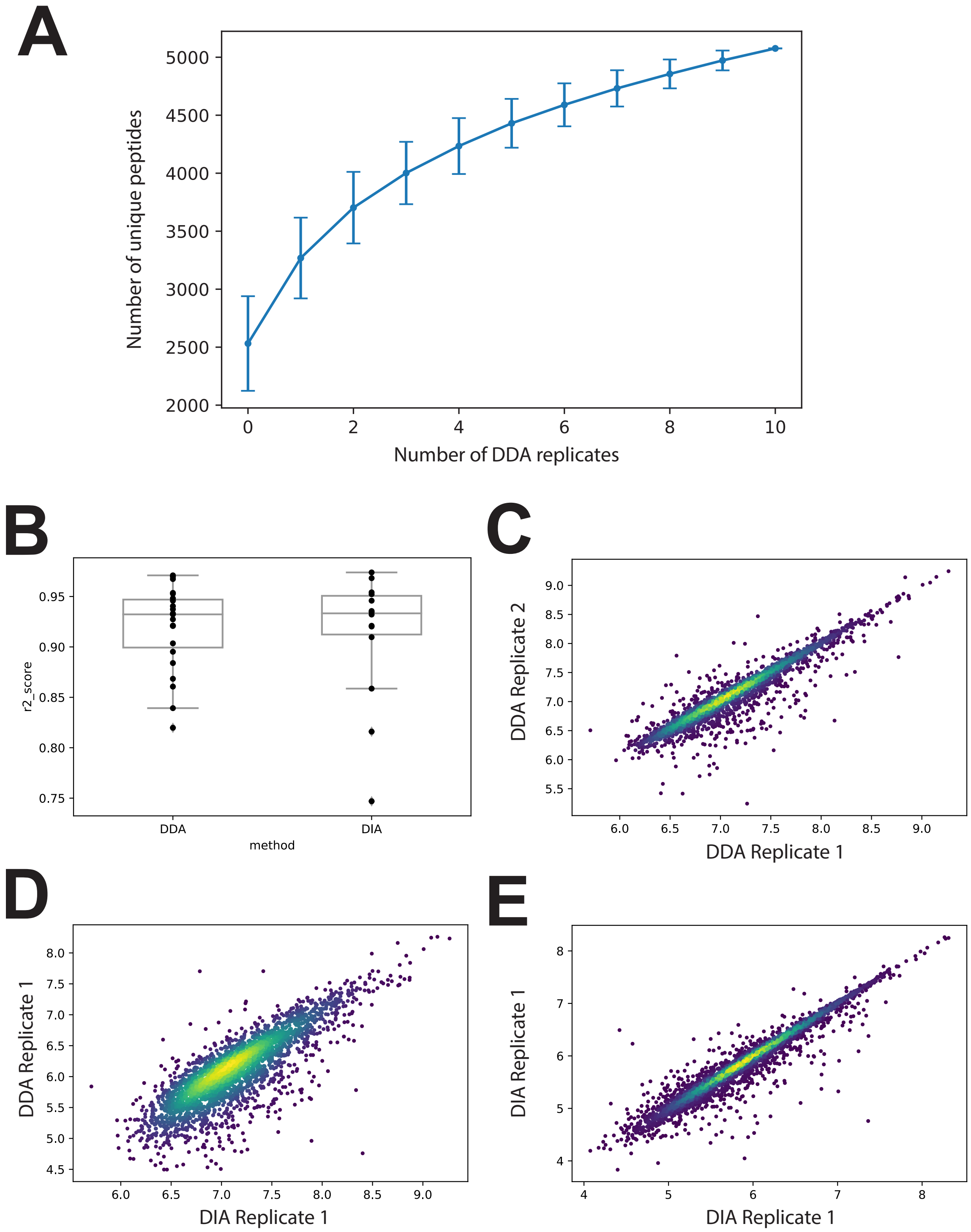


Figure A.7: Additional analysis on reproducibility was comparatively assessed for the DDA and DIA approach: A) The peptide identification (ID) increase when measuring 10 replicates of the same JY cell standard iteratively using the DDA approach. While the initial replicates vary strongly and increase the total number of IDs, the later replicates only add few new IDs indicating an accumulation of false positive IDs. B) R² coefficients from pairwise correlations of peptide quantities resulting from replicate measurements, C) Explicit correlation of the peptide quantities of two DDA replicate measurements, D) Explicit correlation of the peptide quantities of a DIA and DDA replicate measurements of the same sample, E) Explicit correlation of the peptide quantities of two DIA replicate measurements

Pairwise RT alignment using minimum spanning trees

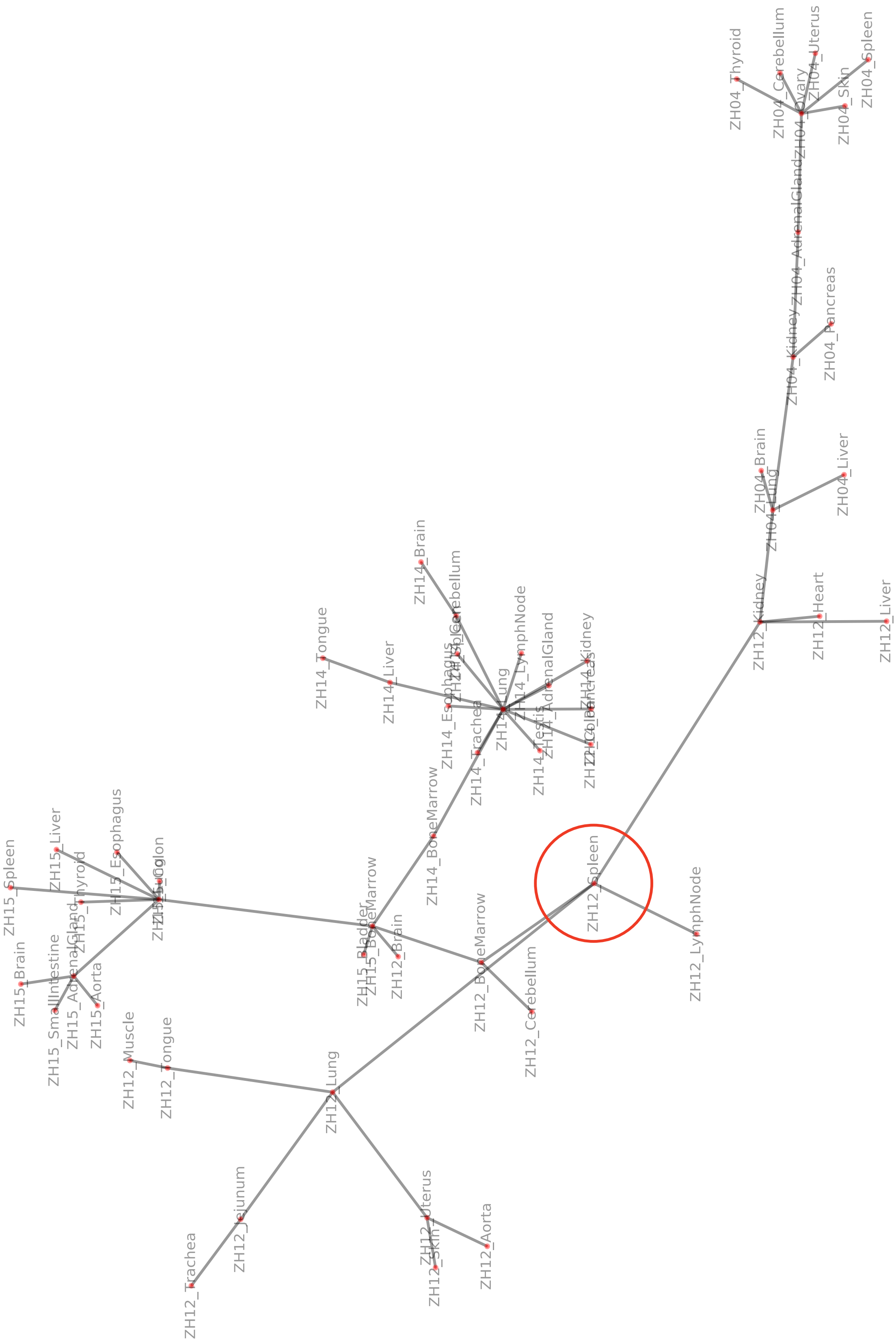


*Figure A.8: Minimum spanning tree for pairwise RT alignment between samples for the generation of a pan-allele (HLA-A*02:01) master library across multiple body donors (ZH). The tree is constructed based on the highest pairwise peptide overlap between two respective samples. All nodes undergo a pairwise linear RT alignment towards the center of the tree (marked by the red circle)*

APPENDIX B

Manual validation of mass spectra

B.1 *MHCquant*

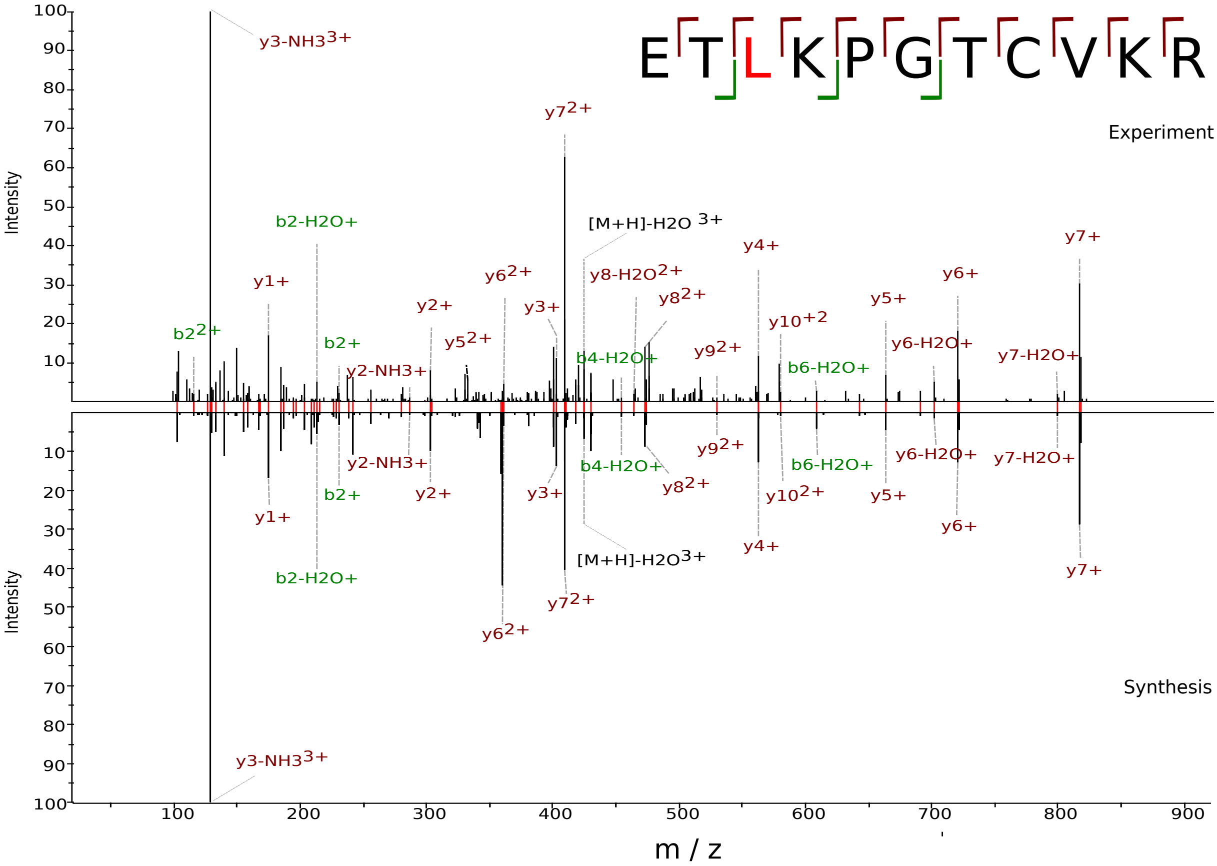
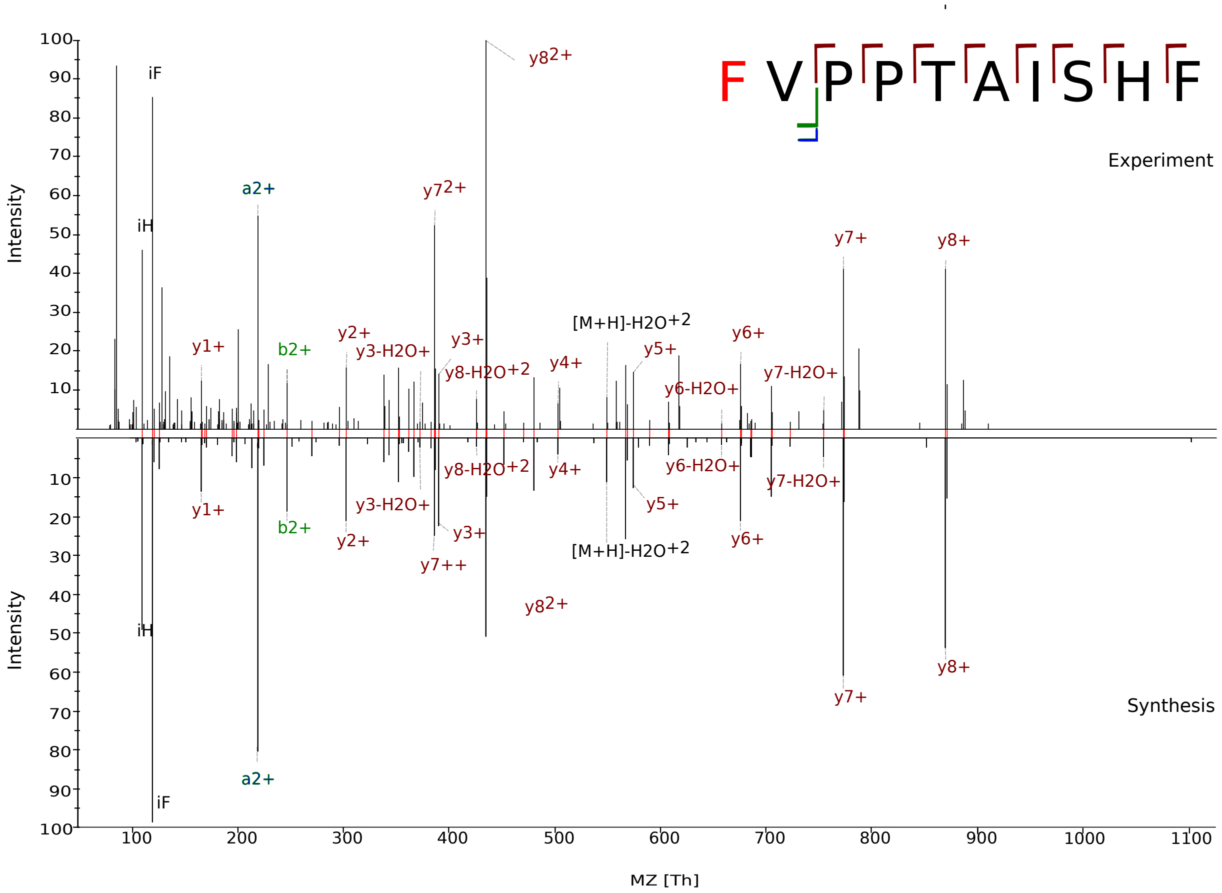
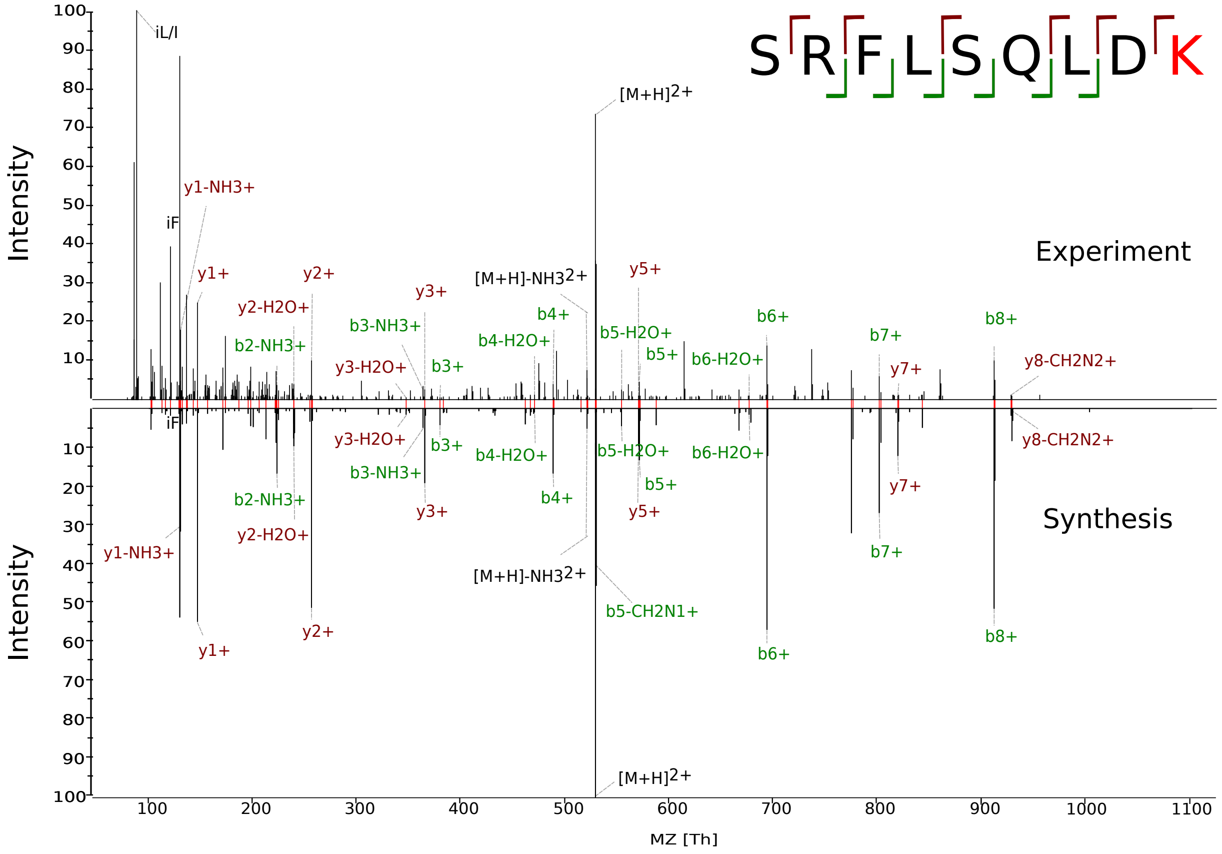


Figure B.1: Detailed peptide spectrum matches including all fragment annotations of additional potential neoepitopes - experimentally determined (upper) and synthetic (lower) peptide. The mutated amino-acid is highlighted in red, b- (green), y- (brown) and a-ions (blue) are annotated for important fragments.


F V P P T A I S H F
Experiment
Synthesis
Intensity
MZ [Th]
100
200
300
400
500
600
700
800
900
1000
1100
iF
iH
a2+
y1+
b2+
y2+
y3-H2O+
y3+
y72+
y7++
y8-H2O+2
y82+
y4+
[M+H]-H2O+2
y5+
y6-H2O+
y6+
y7-H2O+
y7+
y8+


S R F L S Q L D K
Experiment
Synthesis
Intensity
MZ [Th]
100
200
300
400
500
600
700
800
900
1000
1100
iL/I
iF
y1-NH3+
y1+
y2-H2O+
y2+
b2-NH3+
b3-NH3+
y3-H2O+
b3+
y3+
[M+H]-NH32+
[M+H]2+
b4+
b4-H2O+
b5-H2O+
b5+
b5-CH2N1+
y5+
b6+
b6-H2O+
b7+
y7+
b8+
y8-CH2N2+

B.2 *Data-independent acquisition*

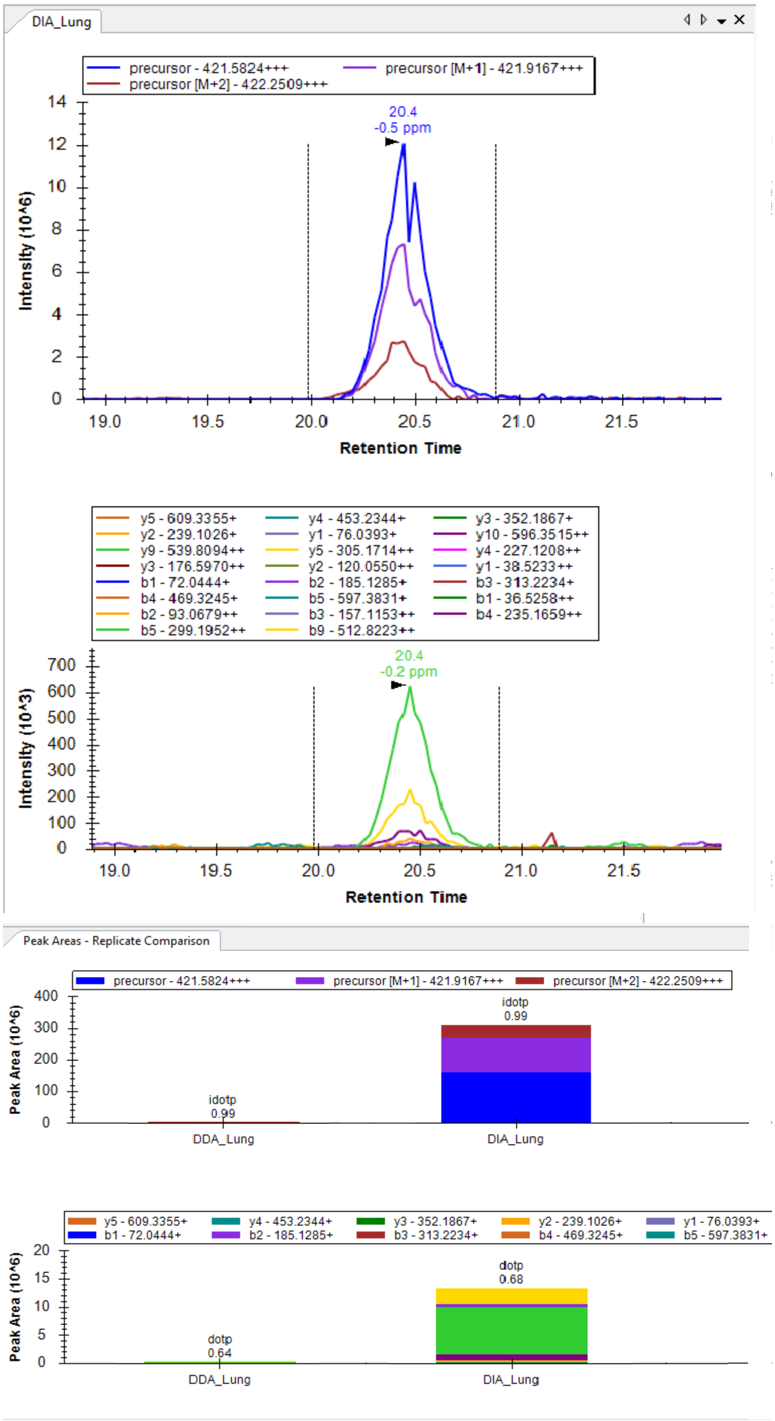


Figure B.2: Examplary XICs of the same peptide identification recovered in lung tissue using the DIA approach in contrast to the DDA approach visualized through Skyline [235]. Transition quantities are extremely low in the corresponding DDA run, which might have lead to the missing discovery in these samples. In contrast DIA recovers the peptide transition in all MS runs.

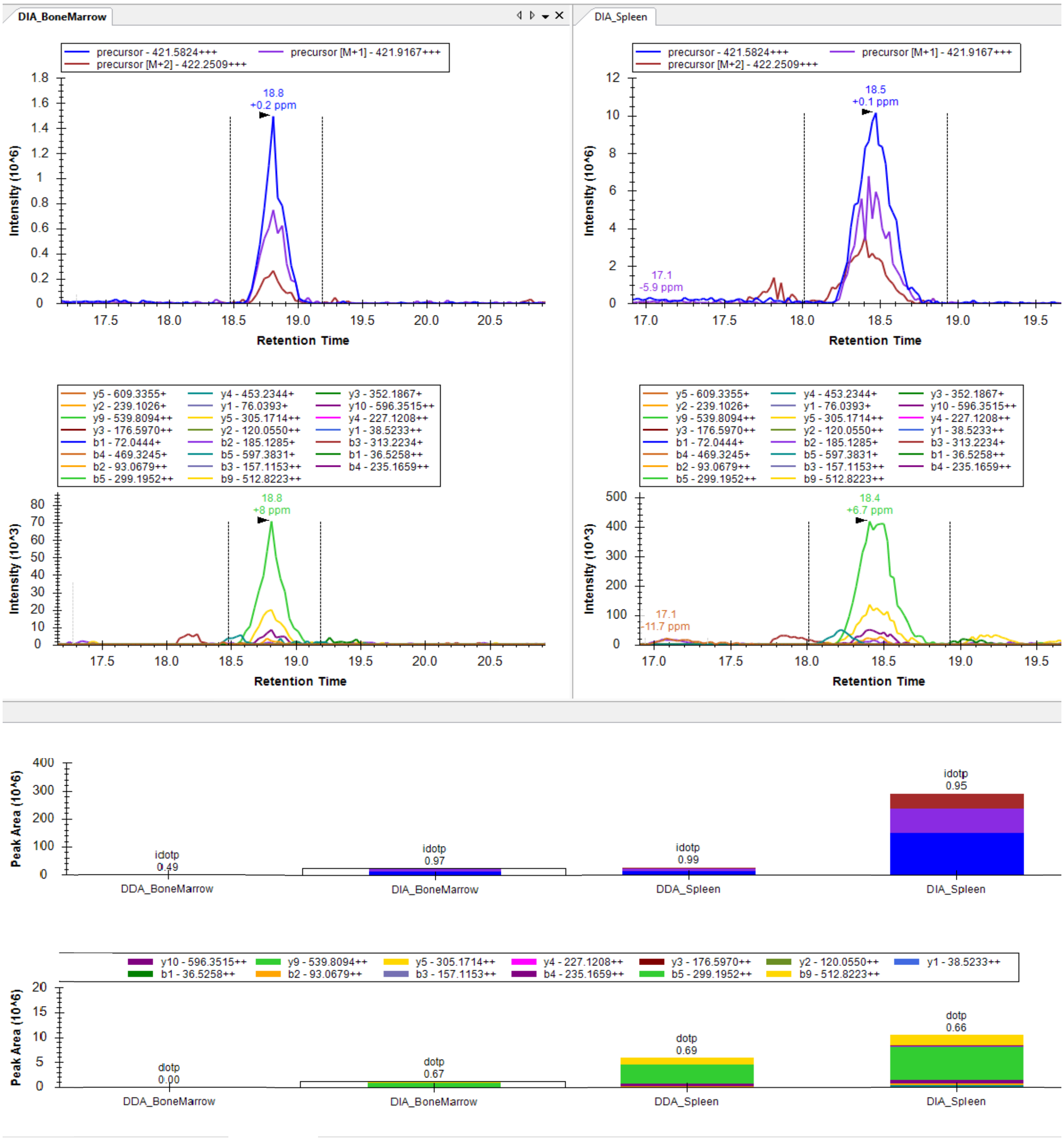


Figure B.3: Examplary XICs of the same peptide identification recovered in bone marrow and spleen tissue using the DIA approach in contrast to the DDA approach visualized through Skyline [235]. Transition quantities are low in some of the DDA runs, which might have lead to the missing discovery in these samples. In contrast DIA recovers the peptide transition in all MS runs.

B.3 *HLA Ligand Atlas*

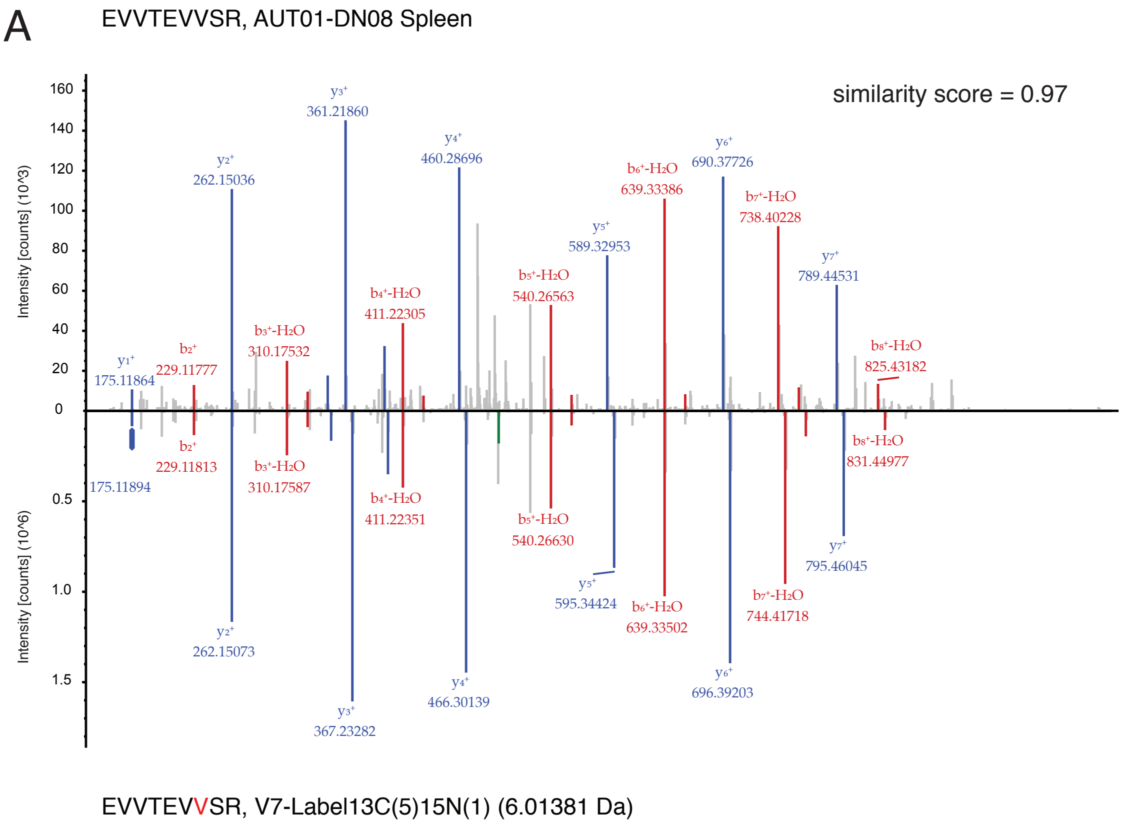
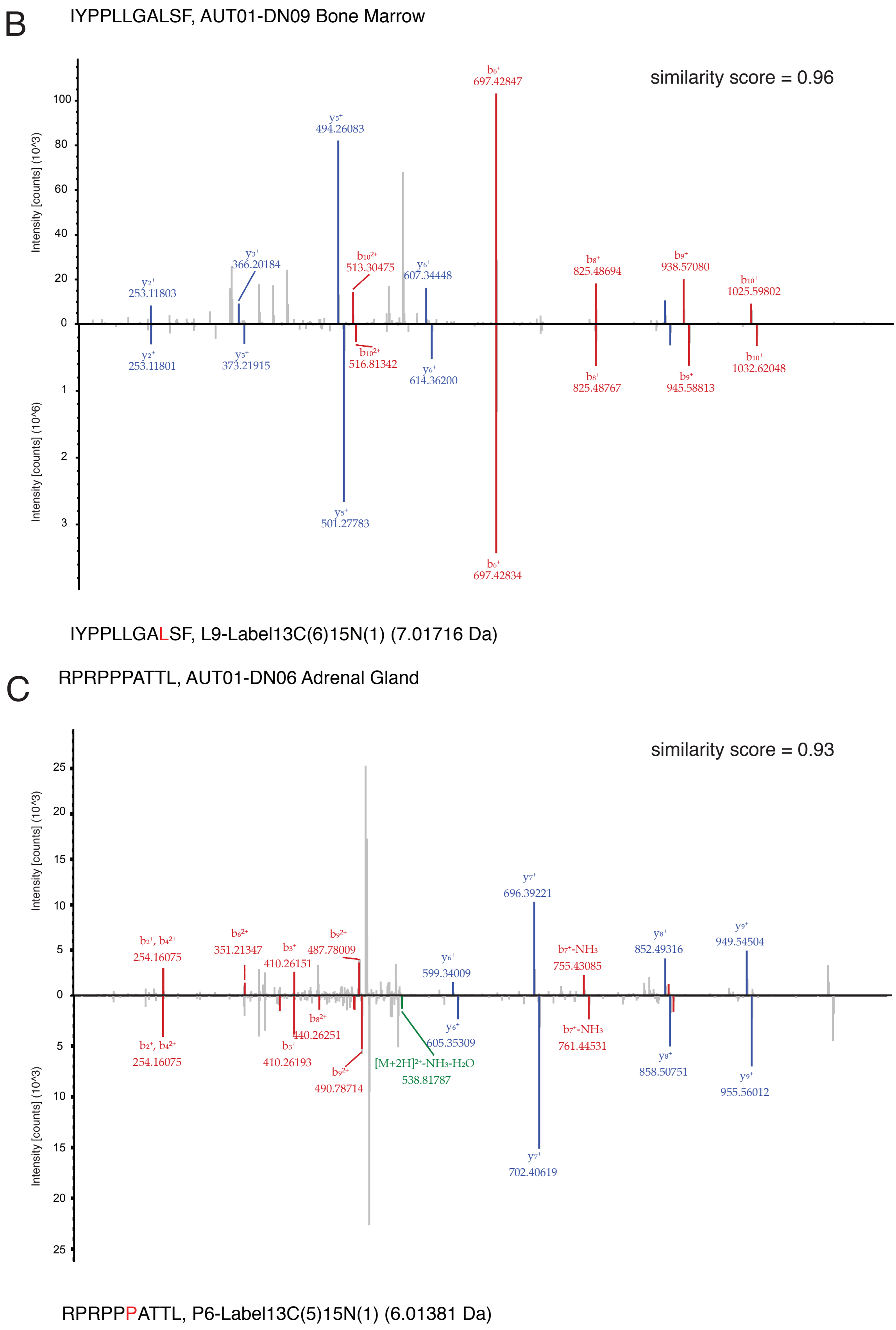
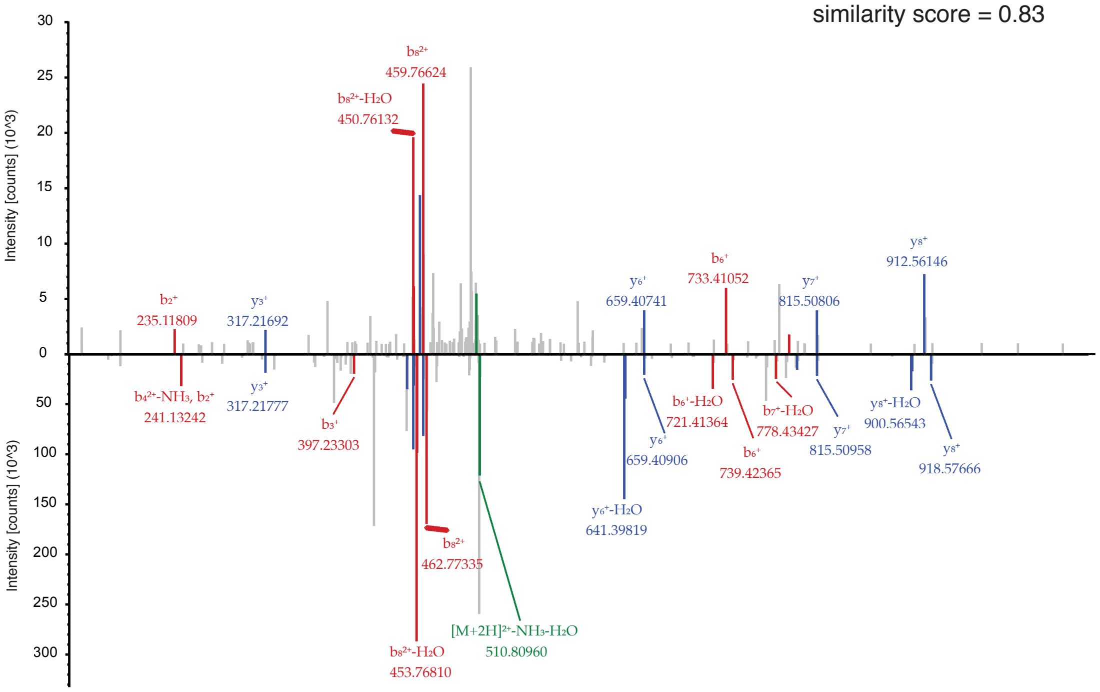


Figure B.4: A-F) Six exemplary spectral comparisons depict cryptic peptides identified in diverse tissues and subjects (upper spectrum) related to their synthetic isotope labeled counterpart (lower spectrum). The isotope labelled amino acid is highlighted in red and its corresponding label and mass are annotated. Matching b- (red), y-ions (blue) and neutral losses (green) and their corresponding fragment masses are annotated to each peak in the spectra. The spectral similarity score of the given comparison is annotated in the top right.


B
IYPPLLGALSF, AUT01-DN09 Bone Marrow
similarity score = 0.96
Intensity [counts] (10^3)
100
80
60
40
20
0
Intensity [counts] (10^6)
1
2
3
y2+ 253.11803
y3+ 366.20184
y5+ 494.26083
b10 2+ 513.30475
y6+ 607.34448
b6+ 697.42847
b8+ 825.48694
b9+ 938.57080
b10+ 1025.59802
y2+ 253.11801
y3+ 373.21915
b10 2+ 516.81342
y6+ 614.36200
y5+ 501.27783
b6+ 697.42834
b8+ 825.48767
b9+ 945.58813
b10+ 1032.62048
IYPPLLGALSF, L9-Label13C(6)15N(1) (7.01716 Da)
C
RPRPPPATTL, AUT01-DN06 Adrenal Gland
similarity score = 0.93
Intensity [counts] (10^3)
25
20
15
10
5
0
Intensity [counts] (10^3)
5
10
15
20
25
b2+, b4 2+ 254.16075
b6 2+ 351.21347
b3+ 410.26151
b9 2+ 487.78009
y6+ 599.34009
y7+ 696.39221
b7+-NH3 755.43085
y8+ 852.49316
y9+ 949.54504
b2+, b4 2+ 254.16075
b8 2+ 440.26251
b3+ 410.26193
b9 2+ 490.78714
[M+2H]2+-NH3-H2O 538.81787
y6+ 605.35309
y7+ 702.40619
b7+-NH3 761.44531
y8+ 858.50751
y9+ 955.56012
RPRPPPATTL, P6-Label13C(5)15N(1) (6.01381 Da)

D

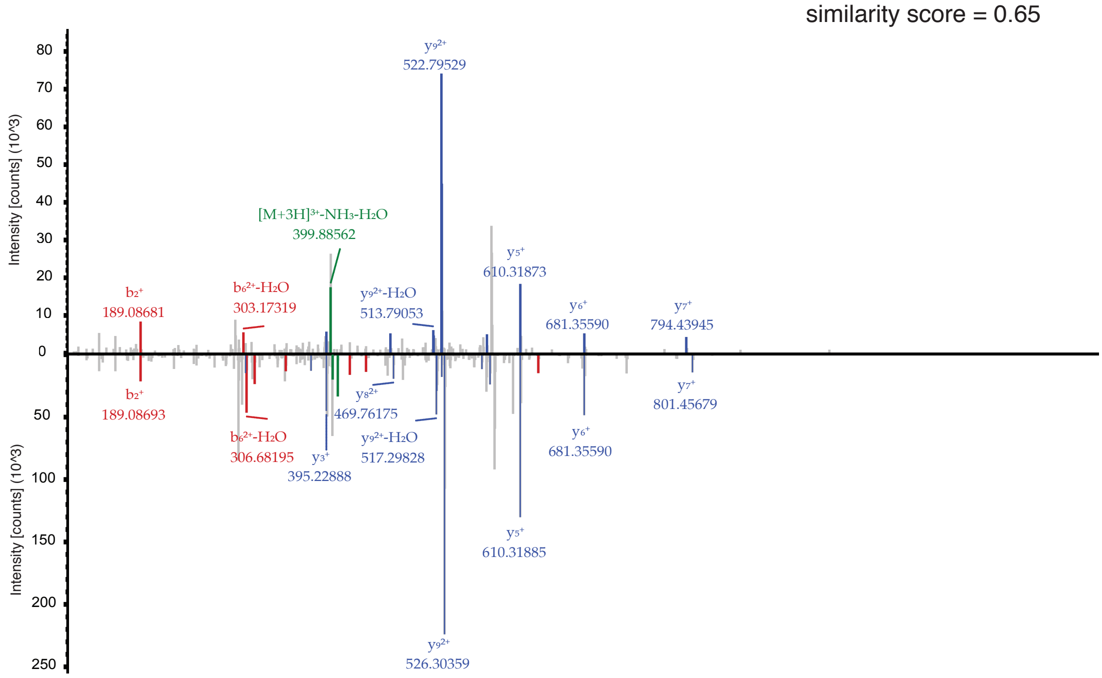
HPRTIQGKL AUT01-DN06 Lung



HPRTIQGKL, P2-Label13C(5)15N(1) (6.01381 Da)

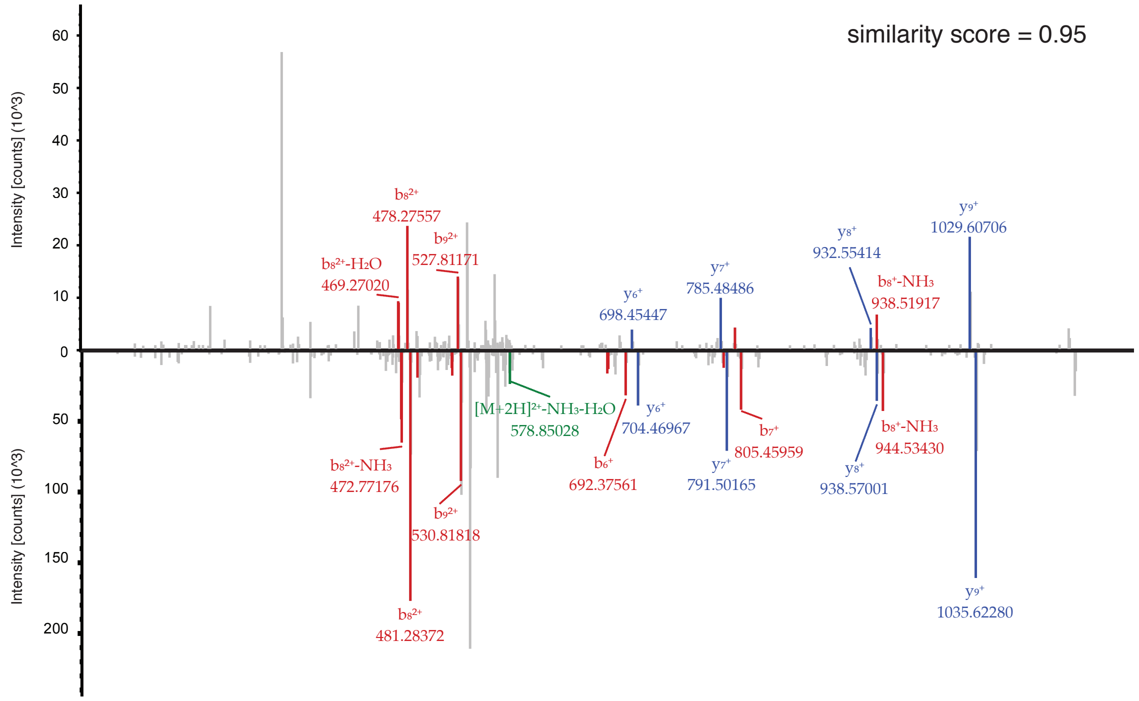
E

STIHLASQFTK, AUT01-DN08 Colon



STIHLASQFTK, L5-Label13C(6)15N(1) (7.01716 Da)

F RPFSPTLRVL, AUT01-DN06 Tongue



RPFSPTLRVL, P5-Label13C(5)15N(1) (6.01381 Da)

APPENDIX C

Characteristics of samples and patients

C.1 *MHCquant*

Subject	Cell type	HLA-A	HLA-B	HLA-C	Cell count (106̂)	Antibody used
PBMC001	PBMCs	02:01/24:02	35:01/55:01	-	1.27	3 mg
PBMC002	PBMCs	01:01/23:01	44:02/44:05	-	340	1 mg
PBMC003	PBMCs	11:01/32:01	18:01/44:02	-	170	1 mg
PBMC004	PBMCs	01:01/02:01	7:02	-	260	1 mg
PBMC005	PBMCs	03:01/07:02	35:01:00	-	370	1 mg
PBMC006	PBMCs	01:01/03:01	08:01/44:02	-	375	1 mg
PBMC007	PBMCs	02:01/31:01	18:01/40:01	-	380	1 mg
PBMC008	PBMCs	11:01/44:02	55:01:00	-	300	1 mg
PBMC009	PBMCs	24:02/32:01	07:02/14:02	-	320	1 mg
JY	EBV-BC	2:01	7:02	C*07:02	75	1 mg

Table C.1: Sample overview used for this study. Abbreviations: PBMCs, pheripheral blood mononuclear cells; EBV-BC, Epstein barr virus immortalised B cell lymphoblastoid line

C.2 *Data-independent acquisition*

Applied Swath Windows

	HLA Class I		HLA Class II	
Window	**Start m/z**	**End m/z**	**Start m/z**	**End m/z**
1	400.4537	415.4537	462.9812	487.9812
2	415.4612	430.4612	487.9937	512.9937
3	430.4688	445.4688	513.0062	538.0062
4	445.4763	460.4763	538.0187	563.0187
5	460.4838	475.4838	563.0313	588.0313
6	475.4912	490.4912	588.0438	613.0438
7	490.4987	505.4987	613.0563	638.0563
8	505.5062	520.5062	638.0687	663.0687
9	520.5137	535.5137	663.0812	688.0812
10	535.5212	550.5212	688.0938	713.0938
11	550.5335	584.5335	718.609	754.609
12	584.5505	618.5505	754.627	790.627
13	618.5675	652.5675	790.645	826.645
14			826.663	862.663
15			862.681	898.681
16			898.699	934.699
17			934.717	970.717

Table C.2: SWATH Windows for DIA MS acquisition of HLA Class I and Class II preparations

C.3 *HLA-Ligand-Atlas*

	n	coverage	average_hit	pc90
HLA-A	16	95.15%	1.47	1.12
HLA-B	21	73.66%	0.96	0.38
HLA-C	14	93.02%	1.42	1.07
HLA-DRB1	19	92.29%	1.4	1.05
HLA-DPA1	3	94.78%	1.29	1.08
HLA-DPB1	9	78.96%	1.02	0.48
HLA-DQA1	11	99.36%	1.73	1.37
HLA-DQB1	12	98.20%	1.65	1.26

Table C.3: Worldwide population coverage of HLA allele frequencies comprised in the HLA Ligand Atlas computed through the population coverage functionality of the IEDB Analysis Resources (`http://tools.iedb.org/population/`) [201]. The analysis was carried out by Lena Freudenmann as part of the publication. [208]

	THY				
Tissue	**DN1**	**DN3**	**DN4**	**DN5**	**DN6**
Thymus	1	1	1.1	1	DN2

Table C.4: Thymus sample amounts of individual donors used for the immunopurification procedure of HLA ligands prior to MS measurement for the HLA Ligand Atlas database

	TimeSeries						
Tissue	**0h**	**8h**	**16h**	**24h**	**48h**	**72h**	
Liver		1.04	1.01	1.05	1.06	1.01	**AUT-DN06**
Ovary	0.582			0.742		0.918	**OVA-DN278**
Ovary	0.737			0.675		0.579	**OVA-DN281**

Table C.5: Tissue sample amounts of all time points of the timeseries analysis used for the immunopurification procedure of HLA ligands prior to MS measurement for the HLA Ligand Atlas database

	AUT01												
Tissue	**DN02**	**DN03**	**DN04**	**DN05**	**DN06**	**DN08**	**DN09**	**DN11**	**DN12**	**DN13**	**DN14**	**DN15**	**DN17**
Adrenal Gland		1.3	1.8		1.6	0.7	0.9	1	1		1.1	0.9	1
Aorta		0.7			2.1	0.3	1.1	1.1	1	1.1	1	0.8	1.1
Bone Marrow					n.a.	1	1.4	1	0.5	1.2	0.7	1.2	0.7
Brain	n.a.	n.a.	2.9		3.2	1	1.4	1	1.2	1.1	0.9	1.1	0.9
Cerebellum	n.a.	3.7	2.5		2.5	1.1	1.1	0.7	1	1	1	1	
Colon					1.3	1.1	1.1	1.1	1.1	1		1.1	1.1
Esophagus		1.6			n.a.	0.9	1.1	1.1	1.1	1	1.1	1	1.2
Gallbladder								1			0.9	0.7	1
Heart		1	1.1	2.6	2.3	0.7	1.1	0.7	0.9	1		1	1.1
Kidney		1.2	2.9	1	2.8	0.6		1.1	1	0.9	1		1.1
Liver		3.2	3	2.4	0.9	0.7	4.7	1.2	1.1	1	1.1	1	1.1
Lung	4.5	2	2.4	2.8	2.3	1		1.3	1.2	1	1	1	1
Lymph Node					0.4	1.1	1	0.5	0.2	0.2	0.9		
Mamma					1.9								
Muscle		2.8			1.9	0.8	1.1	1	0.7	1.1	1.1		1.1
Spinal Chord	3.1												

Table C.6: Tissue sample amounts from all individual donors used for the immunopurification procedure of HLA ligands prior to MS measurement for the HLA Ligand Atlas database.

	AUT01												
Tissue	**DN02**	**DN03**	**DN04**	**DN05**	**DN06**	**DN08**	**DN09**	**DN11**	**DN12**	**DN13**	**DN14**	**DN15**	**DN17**
Ovary			2.6						0.7				
Pancreas	2.3	1.5	2.3				1.1	1.1	1.2	0.9	1		
Prostate						1	1.1	1.1		1	1		1
Skin	2.3	2.7	1.2	0.4	0.7	0.4	0.6	0.3	0.8	1			1.1
Small Intestine	4.1	3			2.4	0.8	1.2	1	0.9	0.9	1		1.1
Spleen		3.6	3.3		1.2	1.9	4.1	1.1		1.1	1.1	1.1	1.1
Stomach	2.6	2.6			1.9	1	1.8	1.2	1.2	1.1	1	0.7	
Testis		1		1		1.1	0.9	1.1		0.9	0.8		1.1
Thyroid	0.8	1.3	0.7	1	n.a.	0.8	0.8	0.8	0.9	1.1	1	1	1.3
Tongue					1.2	0.8	0.8	1	0.9	0.9	0.8		1
Trachea		0.6			n.a.			1.1	0.7	1.1	0.9		1.1
Thymus													
Urinary Bladder					0.9	0.4	1.1	1.1	1.1	1	1.1	0.8	1
Uterus			0.7						1				

Table C.7: Tissue sample amounts from all individual donors used for the immunopurification procedure of HLA ligands prior to MS measurement for the HLA Ligand Atlas database.

Subject	Sex	Age (years)	# tissues	HLA-A	HLA-B	HLA-C	Typing Method
AUT01-DN02	Female	56.7	9	11:01/68:01	15:01/35:03	03:03/04:01	SBT & Luminex + Optitype
AUT01-DN03	Male	80.6	18	01:01/11:01	15:01/35:01	03:03/04:01	SBT & Luminex + Optitype
AUT01-DN04	Female	47	13	02:01/23:01	27:05/50:01	02:02/06:02	SBT & Luminex + Optitype
AUT01-DN05	Male	89.8	7	01:01/11:01	07:02/49:01	07:01/07:02	SBT & Luminex + Optitype
AUT01-DN06	Female	60.1	22	03:01/68:02	07:02/14:02	07:02/08:02	SBT & Luminex + Optitype
AUT01-DN08	Male	59	22	32:01/68:01	15:01/44:02	03:03/07:04	SBT & Luminex
AUT01-DN09	Male	75.1	22	24:02/30:01	13:02/35:08	04:01/06:02	SBT & Luminex + Optitype
AUT01-DN11	Male	69.7	25	01:01/69:01	37:01/49:01	06:02/07:01	SBT & Luminex + Optitype
AUT01-DN12	Female	82.7	23	02:01/11:01	15:01/35:01	03:04/04:01	SBT & Luminex + Optitype
AUT01-DN13	Male	48.9	23	02:05/11:01	40:02/58:01	02:02/07:01	SBT & Luminex + Optitype
AUT01-DN14	Male	55.6	22	02:01/68:02	14:02/27:05	02:02/08:02	SBT & Luminex + Optitype
AUT01-DN15	Female	82.7	15	01:01/02:01	08:01/44:02	07:01/07:04	SBT & Luminex + Optitype
AUT01-DN16	Male	80.4	1	01:01/24:02	08:01/41:01	07:01/17:01	SBT & Luminex + Optitype
AUT01-DN17	Male	76.6	21	03:01/24:02	35:03/45:01	04:01/16:01	SBT & Luminex + Optitype
OVA01-DN278	Female	58	1	02:01:01	44:02	05:01	SBT & Luminex
OVA01-DN281	Female	62	1	11:01/26:01	08:01/35:01	07:02/04:01	SBT & Luminex
THY-DN1	Male	4 days	1	03:01:01/29:02:01	07:02:01/44:03:01	07:02:01/16:01:01	NGS (Histogenetics)
THY-DN3	Female	4 months	1	24:02:01/25:01:01	18:01:01/41:01:01	12:03:01/17:01:01	NGS (Histogenetics)
THY-DN4	Male	6 months	1	02:01:01/26:08:01	15:01:01/44:02:01	03:04:01/05:01:01	NGS (Histogenetics)
THY-DN5	Male	2 months	1	01:01:01/03:01:01	07:06:01/07:02:01	07:02:01/15:05:02	NGS (Histogenetics)
THY-DN6	Male	6 months	1	01:01:01/25:01:01	13:02:01/39:01:01	06:02:01/12:03:01	NGS (Histogenetics)

Table C.8: Characteristics of body donors used for the HLA Ligand Atlas database concerning sex, age, number of extracted tissues and HLA class I typing.

Subject	HLA-DRB1	HLA-DRB3	HLA-DRB4	HLA-DRB5	HLA-DQA1	HLA-DQB1	HLA-DPA1	HLA-DPB1
AUT01-DN02	4:01		1:03		03:02/03:01	3:01	1:03	04:02/04:01
AUT01-DN03	11:03/07:01	2:02	1:01		02:01/05:05	02:02/03:01	1:03	04:01/03:01
AUT01-DN04	13:01/07:01	1:01	1:01		02:01/01:03	02:02/06:03	02:02/01:03	04:01/03:01
AUT01-DN05	11:01/14:54	2:02			01:01/05:05	05:03/03:01	1:03	4:01
AUT01-DN06	13:03/08:01	1:01			04:01/05:05	04:02/03:01	1:03	04:02/04:01
AUT01-DN08	13:03/14:01				5:05	3:01		
AUT01-DN09	7:01		1:03		2:01	2:02	1:03	04:02/03:01
AUT01-DN11	13:02/07:01	3:01	1:03		02:01/01:02	06:04/03:03	1:03	04:01/03:01
AUT01-DN12	01:01/12:01	2:02			01:01/05:05	05:01/03:01	02:01/01:03	02:01/09:01
AUT01-DN13	15:01/10:01			1:01	01:01/01:02	05:01/06:02	1:03	02:01/04:02
AUT01-DN14	13:03/04:01	1:01	1:03		03:02/05:05	03:02/03:01	02:01/01:03	2:01
AUT01-DN15	11:01/03:01	02:02/01:01			5:01	02:01/03:01	02:01/01:03	01:01/03:01
AUT01-DN16	03:01/04:05				05:01/03:02	02:01/02:02		
AUT01-DN17	15:01/14:54	2:02		1:01	01:01/01:02	05:03/06:02	1:03	2:01
OVA01-DN278	11:01/13:01					03:01/06:03		
OVA01-DN281	04:04/11:08					03:02/03:01		
THY01-DN1	01:01/15:01			1:01	01:01/01:02	05:01/06:02	02:01/01:03	04:01/14:01
THY01-DN3	04:05/15:01		1:03	1:01	03:03/01:02	02:02/06:02	02:01/01:03	02:01/09:01
THY01-DN4	11:01/15:01	2:02		1:01	05:05/01:02	06:02/03:01	1:03	02:01/04:01
THY01-DN5	04:05/15:01		1:03	1:01	03:03/01:02	03:02/06:02	1:03	104:01/04:01
THY01-DN6	10:01/16:01			2:02	01:05/01:02	05:01/05:02	02:01/01:03	02:01/17:01

Table C.9: Characteristics of body donors used for the HLA Ligand Atlas database concerning HLA class II typing.

C.4 *HepaVac*

Patient	Age (yrs.)	Sex (m/f)	T	N	M	G	HLA-A	HLA-B	HLA-C
HCC023	68	f	pT3	pN0	cM0	2-3	24:02/29:02	37:01/44:03	06:02/16:01
HCC024	55	m	pT1	pN0	cM0	1	03:01/68:01	15:01/40:01	03:04/03:81
HCC025	70	f	pT3	pNx	cM0	1-2	02:01/11:01	37:01/44:02	06:02/07:04
HCC026	77	m	rpT	2 pN	x cM	0 2	01:01/02:01	08:01/51:01	01:02/07:01
HCC027	78	m	pT3	pN0	cM0	2-3	03:01/24:02	18:01/27:05	02:02/07:01
HCC028	75	f	pT1	pNx	cM0	2	02:01/24:02	07:02/35:03	04:01/07:02
HCC030	85	m	pT3	pNx	cM0	2	02:01/03:01	14:01/27:05	01:02/08:02
HCC034	76	m	pT3	pNx	cM0	2	11:01/23:01	18:01/44:03	04:01/07:01
HCC035	75	m	pT1	pNx	cM0	3	02:01/68:01	27:05/35:03	02:02/04:01
HCC036	70	m	pT3	pN0	cM0	3	02:01/68:01	27:05/44:02	01:02/07:04
HCC038	75	m	pT1	pNx	cM0	2	2:01	7:02	7:02
HCC040	72	m	pT1	pN0	cM0	2	03:01/68:01	44:02/51:01	07:04/15:02
HCC041	57	m	pT2	pNx	cM0	3	1:01	8:01	7:01
HCC042	63	m	pT2	pNx	cM0	3	01:01/03:01	08:01/55:01	03:03/07:01
HCC043	78	f	pT1	pNx	cM0	2	01:01/02:01	08:01/40:01	03:04/07:01
HCC045	73	m	pT1	pNx	cM0	2	01:01/26:01	44:03/47:01	04:01/06:02

Table C.10: Overview of patient characteristics contributing biological samples for the study including HLA class I allotypes, TNM tumor staging (according to Union internationale contre le cancer (UICC) and grading of HCCs.

APPENDIX D

Coauthor contributions

The following persons contributed significantly to the research work described in this thesis:

Michael Ghosh (MG), Annika Nelde (AN), Lukas Heumos (LH),Ana Marcu (AM), Sven-Leon Kuchenbecker (SLK), Christopher Mohr (CM), Lena Freudenmann (LF), Lena Mühlenbruch (LM), Daniel Kowalewski (DK), Nico Trautwein(NT), Christopher M. Schröder (CMS), Markus W. Löffler (MWL), Timo Sachsenberg (TS), Andras Szolek (AS), Oliver Kohlbacher(OK), Stefan Stefanovic (SS), Hans Georg Rammensee (HGR), Shubham Gupta (SG), George Rosenberger (GR) and Hannes Röst (HR).

supervision

Ideas, concepts and the results of this research work were discussed with my supervising principal investigators Prof. Dr. Oliver Kohlbacher (OK), Prof. Stefan Stefanovic (SS), Prof. Hans Georg Rammensee, Dr. Markus Löffler and Dr. Hannes Röst.

mhcquant

MG and AN carried out MS measurements and HLAtyping of samples used for benchmarking the MHCquant workflow. LH contributed to the integration of MHC binding predictions and generation of theoretical neoepitopes into the computational MHCquant workflow. CM tested and integrated the workflow into the Qportal webservice of the Quantitative Biology Center in Tübingen. The work was summarized in a published manuscript only with minor contributions of other authors and parts of the text of this manuscript were reused or reformulated for its incorporation into the thesis chapter.

data-independent acquisition

AM carried out DIA-SWATH MS measurements and other experimental preparations analogously to the HLA Ligand Atlas project. SG and GR provided support in the integrations of the software tools DIAlignR and EasyPQP, respectively. SLK and TS performed code review of the DIAproteomics pipeline. The work was partly summarized in a preprint manuscript only with minor contributions of other authors and parts of the text of this manuscript were reused or reformulated for its incorporation into the thesis chapter.

hla ligand atlas

AM, DK, and LF carried out MS measurements and other experimental preparations for the HLA Ligand Atlas project. In addition, LF assessed world wide HLA frequencies for the HLA Ligand Atlas project. LK implemented the HLA ligand atlas webservice and storage and management of the data in an SQL database. Moreover, LK carried out and visualized parts of the analysis of the HLA Ligand Atlas project, namely the hierarchical clustering of patient / tissue similarities, HLA allele binding affinity predictions and peptide binder distributions. AS performed the HLA typing from exome sequencing data. The work was jointly summarized in a published manuscript with the contribution of most authors and parts of the text of this manuscript were reused or reformulated for its incorporation into the thesis chapter.

HEPAVAC

NT, LF and LM carried out MS measurements and experimental preparations for the HepaVac project. In addition, LF carried out the targeted PRM measurements, analyzed and visualized the results using the Skyline software. CM and CMS analyzed the NGS data as part of the HepaVac project and computed the number of predicted neoepitopes resulting from genetic variants. In addition CM carried out the work on data integration between the various omics levels - namely genomics, transcriptomics and proteomics. Proteomics data analysis employing MaxQuant was carried out by the Tübingen Proteome Center. The work was jointly summarized in a published manuscript with the contribution of most authors and parts of the text of this manuscript were reused or reformulated for its incorporation into the thesis chapter.

OTHER COAUTHORS

Coauthors from publications concerning this dissertation that were not specifically addressed in any of the previous paragraphs were involved in the organisation, have contributed conceptual ideas, sample material or suggested manuscript editions used within the research studies.

APPENDIX E

Curriculum Vitae

Personal data

Name	Leon Clemens Bichmann
Date of Birth	December 10th, 1989
Place of Birth	Frankfurt am Main, Germany
Citizen of	Germany

Education

2017– present	Eberhard Karls Universität, Tübingen, Germany *Philosophical Doctor:* Applied Bioinformatics
Sept.– Oct., 2019	*Research visit:* University of Toronto, Canada
2013 – 2015	Technical University Dresden Dresden, Germany *Master of Science:* Molecular Bioengineering
Feb.– Sept., 2015	*Thesis:* University of California San Francisco, USA
2010 – 2013	Ludwig-Maximilians-University Munich, Germany *Bachelor of Science:* Chemistry and Biochemistry
Sept.– Oct., 2012	*Research visit:* National Yang-Ming University, Taiwan
2000 – 2009	Ziehen Gymnasium High School Frankfurt am Main, Germany *Focus:* Chemistry and Mathematics *Year abroad:* Peninsula School, Melbourne, Australia

Employment

Jan. – Nov., 2018	CureVac AG, Biomarker Research Part time data assistant Tübingen, Germany
Feb. – Dec., 2016	EAWAG, ETH Zurich Data scientist, Environmental chemistry Zurich, Switzerland
Aug. – Sept., 2011	Air Liquide, Research and Technology Student assistant Frankfurt am Main, Germany

Jun. – Jul., 2010 Siemens AG, Healthcare Sector
Technical basic internship
Erlangen, Germany

Oct. 2009 – Jun., 2010 Arbeiter Samariter Bund
Home emergency civil service
Frankfurt, Germany

Awards

Böhringer-Ingelheim travel stipend, visit UofT
European bioinformatics winter school poster prize
Otto-von-Bayer Fellowship, master thesis at UCSF
DAAD "RISE Worldwide", research visit in Taiwan
BIOMOD Competition: 2nd best team - Dresden DNAmic

APPENDIX F

Scientific publications and conference contributions

Submitted manuscripts currently under review:

1. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D. J., Freudenmann, L. K., Backert, L., Mühlenbruch, L., Szolek, A., Lübke, M., Wagner, P., Engler, T., Matovina, S., Wang, J., Hauri-Hohl, M., Martin, R., Kapolou, K., Walz, J. S., Velz, J., Moch, H., Regli, L., Silginer, M., Weller, M., Löffler, M. W., Erhard, F., Schlosser, A., Kohlbacher, O., Stevanović, S., Rammensee, H.-G. & Neidert, M. C. The HLA Ligand Atlas - A Resource of Natural HLA Ligands Presented on Benign Tissues. *bioRxiv*, 778944 (2020).
2. Kubiniok, P., Marcu, A., Bichmann, L., Kuchenbecker, L., Schuster, H., Hamelin, D., Despault, J., Kovalchik, K., Weissling, L., Kohlbacher, O., Stevanovic, S., Rammensee, H.-G., Neidert, M. C., Sirois, I. & Caron, E. *The Global Architecture Shaping the Heterogeneity and Tissue-Dependency of the MHC Class I Immunopeptidome is Evolutionary Conserved* submitted. 2020.
3. Ghosh, M., Hartmann, H., Jakobi, M., März, L., Bichmann, L., Freudenmann, L. K., Mühlenbruch, L., Segan, S., Rammensee, H.-G., Schneiderhan-Marra, N., Shipp, C., Stevanović, S. & Joos, T. O. *The impact of biomaterial cell contact on the immunopeptidome* submitted. 2020.

Articles in peer-reviewed journals:

1. Lübke, M., Spalt, S., Kowalewski, D. J., Zimmermann, C., Bauersfeld, L., Nelde, A., Bichmann, L., Marcu, A., Peper, J. K., Kohlbacher, O., Walz, J. S., Le-Trilling, V. T. K., Hengel, H., Rammensee, H.-G., Stevanović, S. & Halenius, A. Identification of HCMV-Derived T Cell Epitopes in Seropositive Individuals through Viral Deletion Models. *Journal of Experimental Medicine* **217** (2020).
2. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanović, S., Rammensee, H.-G. & Kohlbacher, O. MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. eng. *Journal of Proteome Research* **18**, 3876 (2019).
3. Löffler, M. W., Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., Schroeder, C. M., Trautwein, N., Hilke, F. J., Zinser, R. S., Mühlenbruch, L., Kowalewski, D. J., Schuster, H., Sturm, M., Matthes, J., Riess, O., Czemmel, S., Nahnsen, S., Königsrainer, I., Thiel, K., Nadalin, S., Beckert, S., Bösmüller, H., Fend, F., Velic, A., Maček, B., Haen, S. P., Buonaguro, L., Kohlbacher, O., Stevanović, S., Königsrainer, A., Rammensee, H.-G. & HEPAVAC Consortium. Multi-Omics Discovery of Exome-Derived Neoantigens in Hepatocellular Carcinoma. *Genome Medicine* **11**, 28 (2019).
4. Bilich, T., Nelde, A., Bichmann, L., Roerden, M., Salih, H. R., Kowalewski, D. J., Schuster, H., Tsou, C.-C., Marcu, A., Neidert, M. C., Lübke, M., Rieth, J., Schemionek, M., Brümmendorf, T. H., Vucinic, V., Niederwieser, D., Bauer, J., Märklin, M., Peper, J. K., Klein, R., Kohlbacher, O., Kanz, L., Rammensee, H.-G., Stevanović, S. & Walz, J. S. The HLA Ligandome Landscape of Chronic Myeloid Leukemia Delineates Novel T-Cell Epitopes for Immunotherapy. *Blood*, blood (2018).
5. Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., Schneidman-Duhovny, D., Demerdash, O. N., Mitchell, J. C., Wells, J. A., Fraser, J. S. & Sali, A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology*. *Computation Resources for Molecular Biology* **428**, 709 (2016).

6. Samsonov, S. A., Bichmann, L. & Pisabarro, M. T. Coarse-Grained Model of Glycosaminoglycans. *Journal of Chemical Information and Modeling* **55**, 114 (2014).
7. Bichmann, L., Wang, Y.-T. & Fischer, W. B. Docking Assay of Small Molecule Antivirals to P7 of HCV. ENG. *Computational Biology and Chemistry* **53PB**, 308 (2014).

Book chapters:

1. Alka, O., Sachsenberg, T., Bichmann, L., Pfeuffer, J., Weisser, H., Wein, S., Netz, E., Rurik, M., Kohlbacher, O. & Röst, H. in *Processing Metabolomics and Proteomics Data with Open Software* 201 (2020).

Conference contributions:

1. Schmidt, T., Samaras, P., Dorfer, V., Panse, C., Kockmann, T., Bichmann, L., Van Puyvelde, B., Perez-Riverol, Y., Deutsch, E. W., Bittremieux, W., Kuster, B. & Wilhelm, M. *Interactive Spectrum Validator as an inter-resource tool for fragment ion spectrum comparison between experimental and (predicted) reference spectra* in *American Society for Mass Spectrometry Conference, Houston, Texas, USA* (2020), MP 123.
2. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanović, S., Rammensee, H.-G. & Kohlbacher, O. *MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics* in *American Society for Mass Spectrometry Conference, Atlanta, Georgia, USA* (2019), MP 694.
3. Bauer, J., Zieger, N., Nelde, A., Bichmann, L., Salih, H. R., Kanz, L., Rammensee, H.-G., Stevanović, S. & Walz, J. S. *Mass Spectrometry-Based Immunopeptidome Analysis of Acute Myeloid Leukemia Cells Under Decitabine Treatment Delineates Induced Presentation of Cancer/Testis Antigens on HLA Class I Molecules* in *American Society of Hematology* (Blood, 2019), 5223.
4. Kapolou, K., Freudenmann, L. K., Friebel, E., Bichmann, L., Becher, B., Stevanović, S., Rammensee, H.-G., Weller, M. & Neidert, M. C. *The Antigenic Landscape of Glioblastoma-Refining the Targets for Immunotherapy* in *24th Annual Scientific Meeting and Education Day of the Society for Neuro-Oncology, Phoenix, Arizona, USA* (Neuro-Oncology, 2019), IMMU08.
5. Laban, S., Eziç, J., Bichmann, L., Mytilineos, D., Fürstberger, A., Kestler, H., Schuler, P., Hoffmann, T., Rammensee, H.-G., Stevanović, S. & Mühlenbruch, L. *HLA-Ligandome Analysis Reveals Target Antigens of Oropharyngeal Squamous Cell Carcinoma* in *European Society for Medical Oncology Congress 2019, Barcelona, Spain* (Annals of Oncology, 2019), 4030.
6. Bichmann, L., Marcu, A., Backert, L., Kowalewski, D. J., Freudenmann, L. K., Kohlbacher, O., Rammensee, H.-G., Stevanović, S. & Neidert, M. C. *Immunopeptidomics of human tissues using DDA and DIA* in *European Bioinformatics Winter School, Zakopane, Poland* (2019).
7. Marcu, A., Bichmann, L., Backert, L., Kowalewski, D. J., Freudenmann, L. K., Kohlbacher, O., Rammensee, H.-G., Stevanović, S. & Neidert, M. C. *The HLA-presented peptidome of healthy human organs: possible implications for immunotherapy* in *Association for Cancer Immunotherapy Annual Meeting, Mainz, Germany* (2019), Therapeutic vaccination session.

8. Bichmann, L., Nelde, A., Ghosh, M., Kowalewski, D., Mohr, C., Sachsenberg, T., Stevanović, S., Rammensee, H.-G. & Kohlbacher, O. *A versatile, high-throughput HLA peptidomics pipeline for cancer neoepitope discovery* in *Association for Cancer Immunotherapy Annual Meeting, Mainz, Germany* (2018), Therapeutic vaccination session.
9. Bichmann, L., Nelde, A., Ghosh, M., Kowalewski, D., Mohr, C., Sachsenberg, T., Stevanović, S., Rammensee, H.-G. & Kohlbacher, O. *A versatile, high-throughput HLA peptidomics pipeline for cancer neoepitope discovery* in *Human Immunopeptidome Project Summer School, Madrid, Spain* (2018).

Bibliography

1. Cooper, M. D. & Alder, M. N. The Evolution of Adaptive Immune Systems. *Cell* **124**, 815 (2006).
2. Janeway, C. A. & Medzhitov, R. Innate Immune Recognition. eng. *Annual Review of Immunology* **20**, 197 (2002).
3. Finlay, B. B. & McFadden, G. Anti-Immunology: Evasion of the Host Immune System by Bacterial and Viral Pathogens. *Cell* **124**, 767 (2006).
4. Powers, J. H. Antimicrobial Drug Development–the Past, the Present, and the Future. eng. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **10 Suppl 4**, 23 (2004).
5. Netea, M. G., Schlitzer, A., Placek, K., Joosten, L. A. B. & Schultze, J. L. Innate and Adaptive Immune Memory: An Evolutionary Continuum in the Host's Response to Pathogens. *Cell Host & Microbe* **25**, 13 (2019).
6. Barquet, N. & Domingo, P. Smallpox: The Triumph over the Most Terrible of the Ministers of Death. *Annals of Internal Medicine* **127**, 635 (1997).
7. Parijs, L. V. & Abbas, A. K. Homeostasis and Self-Tolerance in the Immune System: Turning Lymphocytes Off. *Science* **280**, 243 (1998).
8. Fitzmaurice, C. *et al.* Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncology* **3**, 524 (2017).
9. Coulie, P. G., Van den Eynde, B. J., van der Bruggen, P. & Boon, T. Tumour Antigens Recognized by T Lymphocytes: At the Core of Cancer Immunotherapy. *Nature Reviews Cancer* **14**, 135 (2014).
10. Sahin, U. & Türeci, Ö. Personalized Vaccines for Cancer Immunotherapy. *Science* **359**, 1355 (2018).
11. Hu, Z., Ott, P. A. & Wu, C. J. Towards Personalized, Tumour-Specific, Therapeutic Vaccines for Cancer. *Nature reviews. Immunology* **18**, 168 (2018).
12. Toussaint, N. C. & Kohlbacher, O. Towards in Silico Design of Epitope-Based Vaccines. eng. *Expert Opinion on Drug Discovery* **4**, 1047 (2009).
13. Li Pira, G., Ivaldi, F., Moretti, P. & Manca, F. *High Throughput T Epitope Mapping and Vaccine Development* https://www.hindawi.com/journals/bmri/2010/325720/. Review Article. 2010.
14. Caron, E., Kowalewski, D. J., Koh, C. C., Sturm, T., Schuster, H. & Aebersold, R. Analysis of MHC Immunopeptidomes Using Mass Spectrometry. *Molecular & Cellular Proteomics* **14**, 3105 (2015).
15. Vizcaíno, J. A., Kubiniok, P., Kovalchik, K., Ma, Q., Duquette, J. D., Mongrain, I., Deutsch, E. W., Peters, B., Sette, A., Sirois, I. & Caron, E. The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. *Molecular & Cellular Proteomics* (2019).

16. Weisser, H., Nahnsen, S., Grossmann, J., Nilse, L., Quandt, A., Brauer, H., Sturm, M., Kenar, E., Kohlbacher, O., Aebersold, R. & Malmström, L. An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *Journal of Proteome Research* **12**, 1628 (2013).

17. Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., Wojnar, D., Wolski, W. E., Schilling, O., Choudhary, J. S., Malmström, L., Aebersold, R., Reinert, K. & Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. En. *Nature Methods* **13**, 741 (2016).

18. Sinitcyn, P., Tiwary, S., Rudolph, J., Gutenbrunner, P., Wichmann, C., Yılmaz, Ş., Hamzeiy, H., Salinas, F. & Cox, J. MaxQuant Goes Linux. *Nature Methods* **15**, 401 (2018).

19. Mohr, C., Friedrich, A., Wojnar, D., Kenar, E., Polatkan, A. C., Codrea, M. C., Czemmel, S., Kohlbacher, O. & Nahnsen, S. qPortal: A Platform for Data-Driven Biomedical Research. *PLOS ONE* **13**, e0191603 (2018).

20. Murphy, J. P., Konda, P., Kowalewski, D. J., Schuster, H., Clements, D., Kim, Y., Cohen, A. M., Sharif, T., Nielsen, M., Stevanovic, S., Lee, P. W. & Gujar, S. MHC-I Ligand Discovery Using Targeted Database Searches of Mass Spectrometry Data: Implications for T-Cell Immunotherapies. eng. *Journal of Proteome Research* **16**, 1806 (2017).

21. Andreatta, M., Nicastri, A., Peng, X., Hancock, G., Dorrell, L., Ternette, N. & Nielsen, M. MS-Rescue: A Computational Pipeline to Increase the Quality and Yield of Immunopeptidomics Experiments. *PROTEOMICS* **19**, e1800357.

22. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer Neoantigen Prioritization through Sensitive and Reliable Proteogenomics Analysis. *Nature Communications* **11**, 1759 (2020).

23. Erhard, F., Dölken, L., Schilling, B. & Schlosser, A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunology Research* (2020).

24. Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J. & Jemal, A. Global Cancer Statistics, 2012. eng. *CA: a cancer journal for clinicians* **65**, 87 (2015).

25. Of the Liver, E. A. f. t. S. & of Cancer, E. O. f. R. a. T. EASL–EORTC Clinical Practice Guidelines: Management of Hepatocellular Carcinoma. *Journal of Hepatology* **56**, 908 (2012).

26. Jr, C. A. J., Travers, P., Walport, M., Shlomchik, M. J., Jr, C. A. J., Travers, P., Walport, M. & Shlomchik, M. J. *Immunobiology* Fifth (Garland Science, 2001).

27. Nicholson, L. B. The Immune System. *Essays in Biochemistry* **60**, 275 (2016).

28. Chaplin, D. D. Overview of the Immune Response. *The Journal of allergy and clinical immunology* **125**, S3 (2010).

29. Curtis, J. L. Cell-Mediated Adaptive Immune Defense of the Lungs. *Proceedings of the American Thoracic Society* **2**, 412 (2005).

30. Gonzalez-Galarza, F. F., Christmas, S., Middleton, D. & Jones, A. R. Allele Frequency Net: A Database and Online Repository for Immune Gene Frequencies in Worldwide Populations. *Nucleic Acids Research* **39**, D913 (2011).

31. Rammensee, H.-G., Friede, T. & Stevanović, S. MHC Ligands and Peptide Motifs: First Listing. *Immunogenetics* **41**, 178 (1995).

32. Rammensee, H. .-.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A. & Stevanović, S. SYFPEITHI: Database for MHC Ligands and Peptide Motifs. *Immunogenetics* **50**, 213 (1999).

33. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G. & Gfeller, D. Deciphering HLA-I Motifs across HLA Peptidomes Improves Neo-Antigen Predictions and Identifies Allostery Regulating HLA Specificity. *PLOS Computational Biology* **13**, e1005725 (2017).

34. McMahon, R. M., Friis, L., Siebold, C., Friese, M. A., Fugger, L. & Jones, E. Y. Structure of HLA-A*0301 in Complex with a Peptide of Proteolipid Protein: Insights into the Role of HLA-A Alleles in Susceptibility to Multiple Sclerosis. *Acta Crystallographica Section D: Biological Crystallography* **67**, 447 (2011).

35. Li, Y., Depontieu, F. R., Sidney, J., Salay, T. M., Engelhard, V. H., Hunt, D. F., Sette, A., Topalian, S. L. & Mariuzza, R. A. Structural Basis for the Presentation of Tumor-Associated MHC Class II-Restricted Phosphopeptides to CD4+ T Cells. *Journal of Molecular Biology* **399**, 596 (2010).

36. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. UCSF Chimera–a Visualization System for Exploratory Research and Analysis. eng. *Journal of Computational Chemistry* **25**, 1605 (2004).

37. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B. & Nielsen, M. NetMHCpan 4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of immunology (Baltimore, Md. : 1950)* **199**, 3360 (2017).

38. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data. *Nucleic Acids Research.*

39. Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A. & Peters, B. A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. *PLOS Computational Biology* **4**, e1000048 (2008).

40. Racle, J., Michaux, J., Rockinger, G. A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., Bassani-Sternberg, M. & Gfeller, D. Robust Prediction of HLA Class II Epitopes by Deep Motif Deconvolution of Immunopeptidomes. *Nature Biotechnology* **37**, 1283 (2019).

41. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a Systems Understanding of MHC Class I and MHC Class II Antigen Presentation. eng. *Nature Reviews. Immunology* **11**, 823 (2011).

42. Callaway, E. The Race for Coronavirus Vaccines: A Graphical Guide. *Nature* **580**, 576 (2020).

43. Purcell, A. W., McCluskey, J. & Rossjohn, J. More than One Reason to Rethink the Use of Peptides in Vaccine Design. *Nature Reviews Drug Discovery* **6**, 404 (2007).

44. Rammensee, H.-G. & Singh-Jasuja, H. HLA Ligandome Tumor Antigen Discovery for Personalized Vaccine Approach. eng. *Expert Review of Vaccines* **12**, 1211 (2013).

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

46. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., Busch, D. H., Peschel, C., Rad, R., Cox, J., Mann, M. & Krackhardt, A. M. Direct Identification of Clinically Relevant Neoepitopes Presented on Native Human Melanoma Tissue by Mass Spectrometry. *Nature Communications* **7**, 13404 (2016).

47. Shen, L., Zhang, J., Lee, H., Batista, M. T. & Johnston, S. A. RNA Transcription and Splicing Errors as a Source of Cancer Frameshift Neoantigens for Vaccines. *Scientific Reports* **9**, 14184 (2019).

48. Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research Network & Rätsch, G. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. eng. *Cancer Cell* **34**, 211 (2018).

49. Zajac, P., Schultz-Thater, E., Tornillo, L., Sadowski, C., Trella, E., Mengus, C., Iezzi, G. & Spagnoli, G. C. MAGE-A Antigens and Cancer Immunotherapy. *Frontiers in Medicine* **4** (2017).

50. Rosenberg, S. A., Yang, J. C. & Restifo, N. P. Cancer Immunotherapy: Moving beyond Current Vaccines. *Nature Medicine* **10**, 909 (2004).

51. Gouttefangeas, C. & Rammensee, H.-G. Personalized Cancer Vaccines: Adjuvants Are Important, Too. eng. *Cancer immunology, immunotherapy: CII* **67**, 1911 (2018).

52. Wei, L., Zhao, Y., Hu, X. & Tang, L. Redox-Responsive Polycondensate Neoepitope for Enhanced Personalized Cancer Vaccine. *ACS Central Science* **6**, 404 (2020).

53. June, C. H., Warshauer, J. T. & Bluestone, J. A. Is Autoimmunity the Achilles' Heel of Cancer Immunotherapy? *Nature Medicine* **23**, 540 (2017).

54. Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., Litzky, L., Bagg, A., Carreno, B. M., Cimino, P. J., Binder-Scholl, G. K., Smethurst, D. P., Gerry, A. B., Pumphrey, N. J., Bennett, A. D., Brewer, J. E., Dukes, J., Harper, J., Tayton-Martin, H. K., Jakobsen, B. K., Hassan, N. J., Kalos, M. & June, C. H. Cardiovascular Toxicity and Titin Cross-Reactivity of Affinity-Enhanced T Cells in Myeloma and Melanoma. eng. *Blood* **122**, 863 (2013).

55. Joura, E. A., Giuliano, A. R., Iversen, O.-E., Bouchard, C., Mao, C., Mehlsen, J., Moreira, E. D., Ngan, Y., Petersen, L. K., Lazcano-Ponce, E., Pitisuttithum, P., Restrepo, J. A., Stuart, G., Woelber, L., Yang, Y. C., Cuzick, J., Garland, S. M., Huh, W., Kjaer, S. K., Bautista, O. M., Chan, I. S. F., Chen, J., Gesser, R., Moeller, E., Ritter, M., Vuocolo, S., Luxembourg, A. & Broad Spectrum HPV Vaccine Study. A 9-Valent HPV Vaccine against Infection and

Intraepithelial Neoplasia in Women. eng. *The New England Journal of Medicine* **372**, 711 (2015).

56. Harper, D. M. & DeMars, L. R. HPV Vaccines – A Review of the First Decade. *Gynecologic Oncology* **146**, 196 (2017).

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

58. Bezu, L., Kepp, O., Cerrato, G., Pol, J., Fucikova, J., Spisek, R., Zitvogel, L., Kroemer, G. & Galluzzi, L. Trial Watch: Peptide-Based Vaccines in Anticancer Therapy. *Oncoimmunology* **7** (2018).

59. Mant, C. T., Chen, Y., Yan, Z., Popa, T. V., Kovacs, J. M., Mills, J. B., Tripet, B. P. & Hodges, R. S. HPLC Analysis and Purification of Peptides. eng. *Methods in Molecular Biology (Clifton, N.J.)* **386**, 3 (2007).

60. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization–Principles and Practice. *Mass Spectrometry Reviews* **9**, 37 (1990).

61. Sinha, A. & Mann, M. A Beginner's Guide to Mass Spectrometry–Based Proteomics. *The Biochemist*.

62. Bhardwaj, C. & Hanley, L. Ion Sources for Mass Spectrometric Identification and Imaging of Molecular Species. *Natural Product Reports* **31**, 756 (2014).

63. Mellon, F. A. in *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (ed Caballero, B.) 3739 (Academic Press, Oxford, 2003).

64. Perry, R. H., Cooks, R. G. & Noll, R. J. Orbitrap Mass Spectrometry: Instrumentation, Ion Motion and Applications. eng. *Mass Spectrometry Reviews* **27**, 661 (2008 Nov-Dec).

65. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. & Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. eng. *Nature Methods* **4**, 709 (2007).

66. Steen, H. & Mann, M. The Abc's (and Xyz's) of Peptide Sequencing. *Nature Reviews Molecular Cell Biology* **5**, 699 (2004).

67. Senko, M. W., Beu, S. C. & McLafferty, F. W. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *Journal of the American Society for Mass Spectrometry* **6**, 229 (1995).

68. Windig, W., Phalp, J. M. & Payne, A. W. A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry* **68**, 3602 (1996).

69. Hu, A., Noble, W. S. & Wolf-Yadlin, A. Technical Advances in Proteomics: New Developments in Data-Independent Acquisition. eng. *F1000Research* **5** (2016).

70. Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L. & Aebersold, R. OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nature Biotechnology* **32**, 219 (2014).

71. Backert, L. & Kohlbacher, O. Immunoinformatics and Epitope Prediction in the Age of Genomic Medicine. eng. *Genome Medicine* **7**, 119 (2015).

72. Mei, S., Li, F., Leier, A., Marquez-Lago, T. T., Giam, K., Croft, N. P., Akutsu, T., Smith, A. I., Li, J., Rossjohn, J., Purcell, A. W. & Song, J. A Comprehensive Review and Performance Evaluation of Bioinformatics Tools for HLA Class I Peptide-Binding Prediction. *Briefings in Bioinformatics*.

73. Parker, K. C., Bednarek, M. A. & Coligan, J. E. Scheme for Ranking Potential HLA-A2 Binding Peptides Based on Independent Binding of Individual Peptide Side-Chains. *The Journal of Immunology* **152**, 163 (1994).

74. Dönnes, P. & Elofsson, A. Prediction of MHC Class I Binding Peptides, Using SVMHC. eng. *BMC bioinformatics* **3**, 25 (2002).

75. Gulukota, K., Sidney, J., Sette, A. & DeLisi, C. Two Complementary Methods for Predicting Peptides Binding Major Histocompatibility Complex Molecules. eng. *Journal of Molecular Biology* **267**, 1258 (1997).

76. Koch, C. P., Perna, A. M., Pillong, M., Todoroff, N. K., Wrede, P., Folkers, G., Hiss, J. A. & Schneider, G. Scrutinizing MHC-I Binding Peptides and Their Limits of Variation. *PLOS Computational Biology* **9**, e1003088 (2013).

77. Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I. K. A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., Riemer, A. B., Velculescu, V. E., Anagnostou, V., Pagel, K. A. & Karchin, R. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunology Research* **8**, 396 (2020).

78. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O. & Buus, S. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PLOS ONE* **2**, e796 (2007).

79. Chen, B., Khodadoust, M. S., Olsson, N., Wagar, L. E., Fast, E., Liu, C. L., Muftuoglu, Y., Sworder, B. J., Diehn, M., Levy, R., Davis, M. M., Elias, J. E., Altman, R. B. & Alizadeh, A. A. Predicting HLA Class II Antigen Presentation through Integrated Deep Learning. eng. *Nature Biotechnology* **37**, 1332 (2019).

80. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B. & Wilhelm, M. Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nature Methods* **16**, 509 (2019).

81. Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P. & Cox, J. High-Quality MS/MS Spectrum Prediction for Data-Dependent and Data-Independent Acquisition Data Analysis. *Nature Methods* **16**, 519 (2019).

82. Eng, J. K., McCormack, A. L. & Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* **5**, 976 (1994).

83. Eng, J. K., Hoopmann, M. R., Jahan, T. A., Egertson, J. D., Noble, W. S. & MacCoss, M. J. A Deeper Look into Comet – Implementation and Features. *Journal of the American Society for Mass Spectrometry* **26**, 1865 (2015).

84. The, M., Tasnim, A. & Käll, L. How to Talk about Protein-level False Discovery Rates in Shotgun Proteomics. *Proteomics* **16**, 2461 (2016).

85. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. eng. *Journal of Proteome Research* **7**, 40 (2008).

86. Granholm, V., Navarro, J. C., Noble, W. S. & Käll, L. Determining the Calibration of Confidence Estimation Procedures for Unique Peptides in Shotgun Proteomics. *Journal of proteomics* **0**, 123 (2013).

87. Ma, K., Vitek, O. & Nesvizhskii, A. I. A Statistical Model-Building Perspective to Identification of MS/MS Spectra with PeptideProphet. *BMC Bioinformatics* **13**, S1 (2012).

88. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. eng. *Nature Methods* **4**, 923 (2007).

89. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. eng. *Nature Methods* **4**, 207 (2007).

90. Granholm, V., Navarro, J. C., Noble, W. S. & Käll, L. Determining the Calibration of Confidence Estimation Procedures for Unique Peptides in Shotgun Proteomics. *Journal of proteomics* **0**, 123 (2013).

91. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O. & Aebersold, R. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Molecular & Cellular Proteomics : MCP* **8**, 2405 (2009).

92. Bertsch, A., Gröpl, C., Reinert, K. & Kohlbacher, O. OpenMS and TOPP: Open Source Software for LC-MS Data Analysis. eng. *Methods in Molecular Biology (Clifton, N.J.)* **696**, 353 (2011).

93. Pfeuffer, J., Sachsenberg, T., Alka, O., Walzer, M., Fillbrunn, A., Nilse, L., Schilling, O., Reinert, K. & Kohlbacher, O. OpenMS – A Platform for Reproducible Analysis of Mass Spectrometry Data. *Journal of Biotechnology. Bioinformatics Solutions for Big Data Analysis in Life Sciences Presented by the German Network for Bioinformatics Infrastructure* **261**, 142 (2017).

94. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. & Aebersold, R. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics* **10**, 1150 (2010).

95. Aiche, S., Sachsenberg, T., Kenar, E., Walzer, M., Wiswedel, B., Kristl, T., Boyles, M., Duschl, A., Huber, C. G., Berthold, M. R., Reinert, K. & Kohlbacher, O. Workflows for Automated Downstream Data Analysis and Visualization in Large-Scale Computational Mass Spectrometry. *PROTEOMICS* **15**, 1443 (2015).

96. Tyanova, S., Temu, T. & Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. eng. *Nature Protocols* **11**, 2301 (2016).

97. Michael R. Berthold et. al. *KNIME: The Konstanz Information Miner* (Springer, 2007).

98. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. & Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nature Biotechnology* **35**, 316 (2017).

99. Köster, J. & Rahmann, S. Snakemake—a Scalable Bioinformatics Workflow Engine. *Bioinformatics* **28**, 2520 (2012).

100. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* **3** (2016).

101. Ewels, P. A., Peltzer, A., Fillinger, S., Alneberg, J., Patel, H., Wilm, A., Garcia, M. U., Tommaso, P. D. & Nahnsen, S. Nf-Core: Community Curated Bioinformatics Pipelines. *bioRxiv*, 610741 (2019).

102. Shahin, M., Ali Babar, M. & Zhu, L. Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices. *IEEE Access* **5**, 3909 (2017).

103. Smart, J. F. *Jenkins: The Definitive Guide* (O'Reilly Media, Inc., 2011).

104. Docker: Lightweight Linux Containers for Consistent Development and Deployment (2014).

105. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific Containers for Mobility of Compute. *PLOS ONE* **12**, e0177459 (2017).

106. da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I. & Perez-Riverol, Y. BioContainers: An Open-Source and Community-Driven Framework for Software Standardization. *Bioinformatics* **33**, 2580 (2017).

107. Chong, C., Marino, F., Pak, H.-S., Racle, J., Daniel, R. T., Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. High-Throughput and Sensitive Immunopeptidomics Platform Reveals Profound IFNγ-Mediated Remodeling of the HLA Ligandome. *Molecular & Cellular Proteomics* **17**, 533 (2017).

108. Klug, F., Miller, M., Schmidt, H.-H. & Stevanović, S. Characterization of MHC Ligands for Peptide Based Tumor Vaccination. eng. *Current Pharmaceutical Design* **15**, 3221 (2009).

109. Hammerbacher, J. & Snyder, A. Informatics for Cancer Immunotherapy. *Annals of Oncology* **28**, xii56 (2017).

110. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational Genomics Tools for Dissecting Tumour–Immune Cell Interactions. En. *Nature Reviews Genetics* **17**, 441 (2016).

111. Laumont, C. M., Daouda, T., Laverdure, J.-P., Bonneil, É., Caron-Lizotte, O., Hardy, M.-P., Granados, D. P., Durette, C., Lemieux, S., Thibault, P. & Perreault, C. Global Proteogenomic Analysis of Human MHC Class I-Associated Peptides Derived from Non-Canonical Reading Frames. *Nature Communications* **7**, 10238 (2016).

112. Rolfs, Z., Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion. *Journal of Proteome Research* **18**, 349 (2018).

113. Faridi, P., Purcell, A. W. & Croft, N. P. In Immunopeptidomics We Need a Sniper Instead of a Shotgun. *PROTEOMICS* **18**, e1700464 (2018).

114. Mylonas, R., Beer, I., Iseli, C., Chong, C., Pak, H.-S., Gfeller, D., Coukos, G., Xenarios, I., Müller, M. & Bassani-Sternberg, M. Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome. *Molecular & Cellular Proteomics* **17**, 2347 (2018).

115. Liepe, J., Marino, F., Sidney, J., Jeko, A., Bunting, D. E., Sette, A., Kloetzel, P. M., Stumpf, M. P. H., Heck, A. J. R. & Mishto, M. A Large Fraction of HLA Class I Ligands Are Proteasome-Generated Spliced Peptides. eng. *Science (New York, N.Y.)* **354**, 354 (2016).

116. Caron, E., Aebersold, R., Banaei-Esfahani, A., Chong, C. & Bassani-Sternberg, M. A Case for a Human Immuno-Peptidome Project Consortium. *Immunity* **47**, 203 (2017).

117. Sticker, A., Martens, L. & Clement, L. Mass Spectrometrists Should Search for All Peptides, but Assess Only the Ones They Care About. *Nature Methods* **14**, 643 (2017).

118. Wu, T., Guan, J., Handel, A., Tscharke, D. C., Sidney, J., Sette, A., Wakim, L. M., Sng, X. Y. X., Thomas, P. G., Croft, N. P., Purcell, A. W. & Gruta, N. L. L. Quantification of Epitope Abundance Reveals the Effect of Direct and Cross-Presentation on Influenza CTL Responses. En. *Nature Communications* **10**, 2846 (2019).

119. Vizcaíno, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R. & Hermjakob, H. 2016 Update of the PRIDE Database and Its Related Tools. eng. *Nucleic Acids Research* **44**, D447 (2016).

120. Shao, W., Pedrioli, P. G. A., Wolski, W., Scurtescu, C., Schmid, E., Vizcaíno, J. A., Courcelles, M., Schuster, H., Kowalewski, D., Marino, F., Arlehamn, C. S. L., Vaughan, K., Peters, B., Sette, A., Ottenhoff, T. H. M., Meijgaarden, K. E., Nieuwenhuizen, N., Kaufmann, S. H. E., Schlapbach, R., Castle, J. C., Nesvizhskii, A. I., Nielsen, M., Deutsch, E. W., Campbell, D. S., Moritz, R. L., Zubarev, R. A., Ytterberg, A. J., Purcell, A. W., Marcilla, M., Paradela, A., Wang, Q., Costello, C. E., Ternette, N., van Veelen, P. A., van Els, C. A. C. M., Heck, A. J. R., de Souza, G. A., Sollid, L. M., Admon, A., Stevanovic, S., Rammensee, H.-G., Thibault, P., Perreault, C., Bassani-Sternberg, M., Aebersold, R. & Caron, E. The SysteMHC Atlas Project. *Nucleic Acids Research* **46**, D1237 (2018).

121. Khodadoust, M. S., Olsson, N., Wagar, L. E., Haabeth, O. A. W., Chen, B., Swaminathan, K., Rawson, K., Liu, C. L., Steiner, D., Lund, P., Rao, S., Zhang, L., Marceau, C., Stehr, H., Newman, A. M., Czerwinski, D. K., Carlton, V. E. H., Moorhead, M., Faham, M., Kohrt, H. E., Carette, J., Green, M. R., Davis, M. M., Levy, R., Elias, J. E. & Alizadeh, A. A. Antigen Presentation Profiling Reveals Recognition of Lymphoma Immunoglobulin Neoantigens. *Nature* **543**, 723 (2017).

122. Schubert, B., de la Garza, L., Mohr, C., Walzer, M. & Kohlbacher, O. ImmunoNodes – Graphical Development of Complex Immunoinformatics Workflows. *BMC Bioinformatics* **18**, 242 (2017).

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

124. Berlin, C., Kowalewski, D. J., Schuster, H., Mirza, N., Walz, S., Handel, M., Schmid-Horch, B., Salih, H. R., Kanz, L., Rammensee, H.-G., Stevanović, S. & Stickel, J. S. Mapping the HLA Ligandome Landscape of Acute Myeloid Leukemia: A Targeted Approach toward Peptide-Based Immunotherapy. *Leukemia* **29**, 647 (2015).

125. Nelde, A., Kowalewski, D. J., Backert, L., Schuster, H., Kanz, L., Salih, H. R., Rammensee, H.-G., Stevanovic, S. & Stickel, J. S. HLA Ligandome Analysis of Primary Chronic Lymphocytic Leukemia (CLL) Cells Under In Vitro Lenalidomide Treatment Confirms Lenalidomide as a Suitable Combination Partner for T-Cell Based Immunotherapy. *Blood* **128**, 3234 (2016).

126. Barnstable, C. J., Bodmer, W. F., Brown, G., Galfre, G., Milstein, C., Williams, A. F. & Ziegler, A. Production of Monoclonal Antibodies to Group A Erythrocytes, HLA and Other Human Cell Surface Antigens-New Tools for Genetic Analysis. eng. *Cell* **14**, 9 (1978).

127. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A. & Deutsch, E. W. mzML—a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics : MCP* **10** (2011).

128. Schubert, B., Walzer, M., Brachvogel, H.-P., Szolek, A., Mohr, C. & Kohlbacher, O. FRED 2: An Immunoinformatics Framework for Python. *Bioinformatics* **32**, 2044 (2016).

129. Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I. M., Trevanion, S. J., Flicek, P. & Cunningham, F. Ensembl Variation Resources. *Database* **2018** (2018).

130. Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P. & Flicek, P. Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space. eng. *Database: The Journal of Biological Databases and Curation* **2011**, bar030 (2011).

131. Weisser, H. & Choudhary, J. S. Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *Journal of Proteome Research* **16**, 2964 (2017).

132. Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q.-W., Del Toro, N., Pérez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaíno, J. A. & Hermjakob, H. The mzTab Data Exchange Format: Communicating Mass-Spectrometry-Based Proteomics and Metabolomics Experimental Results to a Wider Audience. eng. *Molecular & cellular proteomics: MCP* **13**, 2765 (2014).

133. O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U. & Hammerbacher, J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Systems* **7**, 129 (2018).

134. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and Contaminants Encountered in Modern Mass Spectrometry. *Analytica Chimica Acta*. *Mass Spectrometry* **627**, 71 (2008).

135. Kim, S. & Pevzner, P. A. Universal Database Search Tool for Proteomics. *Nature communications* **5**, 5277 (2014).

136. Tran, N. H., Rahman, M. Z., He, L., Xin, L., Shan, B. & Li, M. Complete De Novo Assembly of Monoclonal Antibody Sequences. eng. *Scientific Reports* **6**, 31730 (2016).

137. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *ELECTROPHORESIS* **20**, 3551 (1999).

138. Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S. & Lund, O. Reliable Prediction of T-Cell Epitopes Using Neural Networks with Novel Sequence Representations. eng. *Protein Science: A Publication of the Protein Society* **12**, 1007 (2003).

139. Zhang, H., Lund, O. & Nielsen, M. The PickPocket Method for Predicting Binding Specificities for Receptors Based on Receptor Pocket Similarities: Application to MHC-Peptide Binding. eng. *Bioinformatics (Oxford, England)* **25**, 1293 (2009).

140. Bardou, P., Mariette, J., Escudié, F., Djemiel, C. & Klopp, C. Jvenn: An Interactive Venn Diagram Viewer. *BMC Bioinformatics* **15**, 293 (2014).

141. Pfeifer, N., Leinenbach, A., Huber, C. G. & Kohlbacher, O. Statistical Learning of Peptide Retention Behavior in Chromatographic Separations: A New Kernel-Based Approach for Computational Proteomics. *BMC Bioinformatics* **8**, 468 (2007).

142. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New Support Vector Algorithms. *Neural Computation* **12**, 1207 (2000).

143. Schölkopf, B., Bartlett, P., Smola, A. & Williamson, R. *Shrinking the Tube: A New Support Vector Regression Algorithm* in *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II* (MIT Press, Cambridge, MA, USA, 1999), 330.

144. Millman, K. J. & Aivazis, M. Python for Scientists and Engineers. *Computing in Science Engineering* **13**, 9 (2011).

145. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825 (2011).

146. Mester, G., Hoffmann, V. & Stevanović, S. Insights into MHC Class I Antigen Processing Gained from Large-Scale Analysis of Class I Ligands. *Cellular and Molecular Life Sciences* **68**, 1521 (2011).

147. Caron, E., Espona, L., Kowalewski, D. J., Schuster, H., Ternette, N., Alpízar, A., Schittenhelm, R. B., Ramarathinam, S. H., Lindestam Arlehamn, C. S., Chiek Koh, C., Gillet, L. C., Rabsteyn, A., Navarro, P., Kim, S., Lam, H., Sturm, T., Marcilla, M., Sette, A., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Purcell, A. W., Rammensee, H.-G., Stevanovic, S. & Aebersold, R. An Open-Source Computational and Data Resource to Analyze Digital Maps of Immunopeptidomes. eng. *eLife* **4** (2015).

148. Chong, C., Müller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B. J., Michaux, J., Bilic, I., Hirsekorn, A., Calviello, L., Simó-Riudalbas, L., Planet, E., Lubiński, J., Bryśkiewicz, M., Wiznerowicz, M., Xenarios, I., Zhang, L., Trono, D., Harari, A., Ohler, U., Coukos, G. & Bassani-Sternberg, M. Integrated Proteogenomic Deep Sequencing and Analytics Accurately Identify Non-Canonical Peptides in Tumor Immunopeptidomes. *Nature Communications* **11**, 1293 (2020).

149. Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., Martens, L. & Barsnes, H. PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets. *Nature Biotechnology* **33**, 22 (2015).

150. Gabriels, R., Martens, L. & Degroeve, S. Updated MS^{2}PIP Web Server Delivers Fast and Accurate MS^2 Peak Intensity Prediction for Multiple Fragmentation Methods, Instruments and Labeling Techniques. *Nucleic Acids Research* **47**, W295 (2019).

151. Meier, F., Beck, S., Grassl, N., Lubeck, M., Park, M. A., Raether, O. & Mann, M. Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *Journal of Proteome Research* **14**, 5378 (2015).

152. Meier, F., Brunner, A.-D., Frank, M., Ha, A., Bludau, I., Voytik, E., Kaspar-Schoenefeld, S., Lubeck, M., Raether, O., Aebersold, R., Collins, B. C., Röst, H. L. & Mann, M. Parallel Accumulation – Serial Fragmentation Combined with Data-Independent Acquisition (diaPASEF): Bottom-up Proteomics with near Optimal Ion Usage. *bioRxiv*, 656207 (2020).

153. Doerr, A. DIA Mass Spectrometry. *Nature Methods* **12**, 35 (2015).

154. Bekker-Jensen, D. B., Bernhardt, O. M., Hogrebe, A., Martinez-Val, A., Verbeke, L., Gandhi, T., Kelstrup, C. D., Reiter, L. & Olsen, J. V. Rapid and Site-Specific Deep Phosphoproteome Profiling by Data-Independent Acquisition without the Need for Spectral Libraries. *Nature Communications* **11**, 787 (2020).

155. Meyer, J. G. & Schilling, B. Clinical Applications of Quantitative Proteomics Using Targeted and Untargeted Data-Independent Acquisition Techniques. *Expert review of proteomics* **14**, 419 (2017).

156. Canterbury, J. D., Merrihew, G. E., Goodlett, D. R., MacCoss, M. J. & Shaffer, S. A. Comparison of Data Acquisition Strategies on Quadrupole Ion Trap Instrumentation for Shotgun Proteomics. *Journal of the American Society for Mass Spectrometry* **25**, 2048 (2014).

157. Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C. & Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nature Methods* **12**, 258 (2015).

158. Ritz, D., Kinzi, J., Neri, D. & Fugmann, T. Data Independent Acquisition of HLA Class I Peptidomes on the Q Exactive Mass Spectrometer Platform. *Proteomics* **17** (2017).

159. Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., MacLean, B. & Aebersold, R. Building High-Quality Assay Libraries for Targeted Analysis of SWATH MS Data. eng. *Nature Protocols* **10**, 426 (2015).

160. Rosenberger, G., Koh, C. C., Guo, T., Röst, H. L., Kouvonen, P., Collins, B. C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., Faini, M., Schubert, O. T., Faridi, P., Ebhardt, H. A., Matondo, M., Lam, H., Bader, S. L., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Tate, S. & Aebersold, R. A Repository of Assays to Quantify 10,000 Human Proteins by SWATH-MS. *Scientific Data* **1**, 140031 (2014).

161. Noble, W. S. Mass Spectrometrists Should Search Only for Peptides They Care About. *Nature Methods* **12**, 605 (2015).

162. Schuster, H., Shao, W., Weiss, T., Pedrioli, P. G. A., Roth, P., Weller, M., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Planz, O., Rammensee, H.-G., Aebersold, R. & Caron, E. A Tissue-Based Draft Map of the Murine MHC Class I Immunopeptidome. *Scientific Data* **5**, 180157 (2018).

163. Marcu, A., Bichmann, L., Kuchenbecker, L., Backert, L., Kowalewski, D. J., Freudenmann, L. K., Löffler, M. W., Lübke, M., Walz, J. S., Velz, J., Moch, H., Regli, L., Silginer, M., Weller, M., Schlosser, A., Kohlbacher, O., Stevanović, S., Rammensee, H.-G. & Neidert, M. C. The HLA Ligand Atlas. A Resource of Natural HLA Ligands Presented on Benign Tissues. *bioRxiv*, 778944 (2019).

164. Rosenberger, G., Bludau, I., Schmitt, U., Heusel, M., Hunter, C., Liu, Y., MacCoss, M. J., MacLean, B. X., Nesvizhskii, A. I., Pedrioli, P. G. A., Reiter, L., Röst, H. L., Tate, S., Ting, Y. S., Collins, B. C. & Aebersold, R. Statistical Control of Peptide and Protein Error Rates in Large-Scale Targeted DIA Analyses. *Nature methods* **14**, 921 (2017).

165. Gupta, S., Ahadi, S., Zhou, W. & Rost, H. DIAlignR Provides Precise Retention Time Alignment across Distant Runs in DIA and Targeted Proteomics. *Molecular & Cellular Proteomics* (2019).

166. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. & Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nature Biotechnology* **35**, 316 (2017).

167. Yu, F., Haynes, S. E., Teo, G. C., Avtonomov, D. M., Polasky, D. A. & Nesvizhskii, A. I. Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Molecular & Cellular Proteomics* (2020).

168. Röst, H. L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B. C., Gillet, L., Testa, G., Malmström, L. & Aebersold, R. TRIC: An Automated Alignment Strategy for Reproducible Protein Quantification in Targeted Proteomics. *Nature methods* **13**, 777 (2016).

169. Gupta, S., Ahadi, S., Zhou, W. & Röst, H. DIAlignR Provides Precise Retention Time Alignment Across Distant Runs in DIA and Targeted Proteomics. *Molecular & Cellular Proteomics* **18**, 806 (2019).

170. Gupta, S. & Röst, H. Automated Workflow For Peptide-Level Quantitation From DIA/SWATH-MS Data. *bioRxiv*, 2020.01.21.914788 (2020).

171. Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B. & Vitek, O. MSstats: An R Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Bioinformatics* **30**, 2524 (2014).

172. *Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function Using NetworkX* http://conference.scipy.org/proceedings/SciPy2008.

173. Lander, E. S. *et al.* Initial Sequencing and Analysis of the Human Genome. eng. *Nature* **409**, 860 (2001).

174. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304 (2001).

175. Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G. & Guigó, R. The Human Transcriptome across Tissues and Individuals. *Science* **348**, 660 (2015).

176. Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach,

S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H. & Pandey, A. A Draft Map of the Human Proteome. *Nature* **509**, 575 (2014).

177. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. & Pontén, F. Tissue-Based Map of the Human Proteome. *Science* **347** (2015).

178. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. Mass-Spectrometry-Based Draft of the Human Proteome. *Nature* **509**, 582 (2014).

179. Fortier, M.-H., Caron, E., Hardy, M.-P., Voisin, G., Lemieux, S., Perreault, C. & Thibault, P. The MHC Class I Peptide Repertoire Is Molded by the Transcriptome. eng. *The Journal of Experimental Medicine* **205**, 595 (2008).

180. Schuster, H., Peper, J. K., Bösmüller, H.-C., Röhle, K., Backert, L., Bilich, T., Ney, B., Löffler, M. W., Kowalewski, D. J., Trautwein, N., Rabsteyn, A., Engler, T., Braun, S., Haen, S. P., Walz, J. S., Schmid-Horch, B., Brucker, S. Y., Wallwiener, D., Kohlbacher, O., Fend, F., Rammensee, H.-G., Stevanović, S., Staebler, A. & Wagner, P. The Immunopeptidomic Landscape of Ovarian Carcinomas. eng. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E9942 (2017).

181. Weinzierl, A. O., Lemmel, C., Schoor, O., Müller, M., Krüger, T., Wernet, D., Hennenlotter, J., Stenzl, A., Klingel, K., Rammensee, H.-G. & Stevanovic, S. Distorted Relation between mRNA Copy Number and Corresponding Major Histocompatibility Complex Ligand Density on the Cell Surface. eng. *Molecular & cellular proteomics: MCP* **6**, 102 (2007).

182. Boegel, S., Löwer, M., Bukur, T., Sorn, P., Castle, J. C. & Sahin, U. HLA and Proteasome Expression Body Map. *BMC Medical Genomics* **11** (2018).

183. Finotello, F., Rieder, D., Hackl, H. & Trajanoski, Z. Next-Generation Computational Tools for Interrogating Cancer Immunity. *Nature Reviews Genetics* **20**, 724 (2019).

184. Hilf, N., Kuttruff-Coqui, S., Frenzel, K., Bukur, V., Stevanović, S., Gouttefangeas, C., Platten, M., Tabatabai, G., Dutoit, V., van der Burg, S. H., thor Straten, P., Martínez-Ricarte, F., Ponsati, B., Okada, H., Lassen, U., Admon, A., Ottensmeier, C. H., Ulges, A., Kreiter, S., von Deimling, A., Skardelly, M., Migliorini, D., Kroep, J. R., Idorn, M., Rodon, J., Piró, J., Poulsen, H. S., Shraibman, B., McCann, K., Mendrzyk, R., Löwer, M., Stieglbauer, M., Britten, C. M., Capper, D., Welters, M. J. P., Sahuquillo, J., Kiesel, K., Derhovanessian, E., Rusch, E., Bunse, L., Song, C., Heesch, S., Wagner, C., Kemmer-Brück, A., Ludwig, J., Castle, J. C., Schoor, O., Tadmor, A. D., Green, E., Fritsche, J., Meyer, M., Pawlowski, N., Dorner, S., Hoffgaard, F., Rössler, B., Maurer, D., Weinschenk, T., Reinhardt, C., Huber, C., Rammensee, H.-G., Singh-Jasuja, H., Sahin, U., Dietrich, P.-Y. & Wick, W. Actively Personalized Vaccination Trial for Newly Diagnosed Glioblastoma. En. *Nature* **565**, 240 (2018).

185. Löffler, M. W., Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., Schroeder, C. M., Trautwein, N., Hilke, F. J., Zinser, R. S., Mühlenbruch, L., Kowalewski, D. J., Schuster, H., Sturm, M., Matthes, J., Riess, O., Czemmel, S., Nahnsen, S., Königsrainer, I., Thiel, K., Nadalin, S., Beckert, S., Bösmüller, H., Fend, F., Velic, A., Maček, B., Haen, S. P., Buonaguro, L., Kohlbacher, O., Stevanović, S., Königsrainer, A., Rammensee, H.-G. & HEPAVAC Consortium. Multi-Omics Discovery of Exome-Derived Neoantigens in Hepatocellular Carcinoma. *Genome Medicine* **11**, 28 (2019).

186. Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M. & Butte, A. J. Comprehensive Analysis of Normal Adjacent to Tumor Transcriptomes. *Nature Communications* **8**, 1077 (2017).

187. Cameron, B. J., Gerry, A. B., Dukes, J., Harper, J. V., Kannan, V., Bianchi, F. C., Grand, F., Brewer, J. E., Gupta, M., Plesa, G., Bossi, G., Vuidepot, A., Powlesland, A. S., Legg, A., Adams, K. J., Bennett, A. D., Pumphrey, N. J., Williams, D. D., Binder-Scholl, G., Kulikovskaya, I., Levine, B. L., Riley, J. L., Varela-Rohena, A., Stadtmauer, E. A., Rapoport, A. P., Linette, G. P., June, C. H., Hassan, N. J., Kalos, M. & Jakobsen, B. K. Identification of a Titin-Derived HLA-A1-Presented Peptide as a Cross-Reactive Target for Engineered MAGE A3-Directed T Cells. eng. *Science Translational Medicine* **5**, 197ra103 (2013).

188. Iacobuzio-Donahue, C. A., Michael, C., Baez, P., Kappagantula, R., Hooper, J. E. & Hollman, T. J. Cancer Biology as Revealed by the Research Autopsy. eng. *Nature Reviews. Cancer* **19**, 686 (2019).

189. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science (New York, N.Y.)* **348**, 648 (2015).

190. Freudenmann, L. K., Marcu, A. & Stevanović, S. Mapping the Tumour Human Leukocyte Antigen (HLA) Ligandome by Mass Spectrometry. *Immunology* **154**, 331 (2018).

191. Fritsche, J., Rakitsch, B., Hoffgaard, F., Römer, M., Schuster, H., Kowalewski, D. J., Priemer, M., Stos-Zweifel, V., Hörzer, H., Satelli, A., Sonntag, A., Goldfinger, V., Song, C., Mahr, A., Ott, M., Schoor, O. & Weinschenk, T. Front Cover: Translating Immunopeptidomics to Immunotherapy-Decision-Making for Patient and Personalized Target Selection. *PROTEOMICS* **18**, 1870101 (2018).

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

193. Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W. H., Peters, B. & Nielsen, M. Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. eng. *Journal of Proteome Research* **19**, 2304 (2020).

194. Bilich, T., Nelde, A., Bichmann, L., Roerden, M., Salih, H. R., Kowalewski, D. J., Schuster, H., Tsou, C.-C., Marcu, A., Neidert, M. C., Lübke, M., Rieth, J., Schemionek, M., Brümmendorf, T. H., Vucinic, V., Niederwieser, D., Bauer, J., Märklin, M., Peper, J. K., Klein, R., Kohlbacher, O., Kanz, L., Rammensee, H.-G., Stevanović, S. & Walz, J. S. The HLA Ligandome Landscape of Chronic Myeloid Leukemia Delineates Novel T-Cell Epitopes for Immunotherapy. *Blood* **133**, 550 (2019).

195. Reustle, A., Di Marco, M., Meyerhoff, C., Nelde, A., Walz, J. S., Winter, S., Kandabarau, S., Büttner, F., Haag, M., Backert, L., Kowalewski, D. J., Rausch, S., Hennenlotter, J., Stühler, V., Scharpf, M., Fend, F., Stenzl, A., Rammensee, H.-G., Bedke, J., Stevanović, S., Schwab, M. & Schaeffeler, E. Integrative -Omics and HLA-Ligandomics Analysis to Identify Novel Drug Targets for ccRCC Immunotherapy. eng. *Genome Medicine* **12**, 32 (2020).

196. Granados, D. P., Rodenbrock, A., Laverdure, J.-P., Côté, C., Caron-Lizotte, O., Carli, C., Pearson, H., Janelle, V., Durette, C., Bonneil, E., Roy, D. C., Delisle, J.-S., Lemieux, S., Thibault, P. & Perreault, C. Proteogenomic-Based Discovery of Minor Histocompatibility Antigens with Suitable Features for Immunotherapy of Hematologic Cancers. eng. *Leukemia* **30**, 1344 (2016).

197. Laumont, C. M. & Perreault, C. Exploiting Non-Canonical Translation to Identify New Targets for T Cell-Based Cancer Immunotherapy. eng. *Cellular and molecular life sciences: CMLS* **75**, 607 (2018).

198. Ouspenskaia, T., Law, T., Clauser, K. R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B. A., Le, P. M., Hartigan, C. R., Keshishian, H., Apffel, A., Oliveira, G., Zhang, W., Chow, Y. T., Ji, Z., Shukla, S. A., Bachireddy, P., Getz, G., Hacohen, N., Keskin, D. B., Carr, S. A., Wu, C. J. & Regev, A. Thousands of Novel Unannotated Proteins Expand the MHC I Immunopeptidome in Cancer. *bioRxiv*, 2020.02.12.945840 (2020).

199. Faridi, P., Li, C., Ramarathinam, S. H., Vivian, J. P., Illing, P. T., Mifsud, N. A., Ayala, R., Song, J., Gearing, L. J., Hertzog, P. J., Ternette, N., Rossjohn, J., Croft, N. P. & Purcell, A. W. A Subset of HLA-I Peptides Are Not Genomically Templated: Evidence for Cis- and Trans-Spliced Peptide Ligands. eng. *Science Immunology* **3** (2018).

200. Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization. *Frontiers in Immunology* **8** (2017).

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

202. Zalocusky, K. A., Kan, M. J., Hu, Z., Dunn, P., Thomson, E., Wiser, J., Bhattacharya, S. & Butte, A. J. The 10,000 Immunomes Project: Building a Resource for Human Immunology. eng. *Cell Reports* **25**, 513 (2018).

203. Granot, T., Senda, T., Carpenter, D. J., Matsuoka, N., Weiner, J., Gordon, C. L., Miron, M., Kumar, B. V., Griesemer, A., Ho, S.-H., Lerner, H., Thome, J. J. C., Connors, T., Reizis, B. & Farber, D. L. Dendritic Cells Display Subset and Tissue-Specific Maturation Dynamics over Human Life. *Immunity* **46**, 504 (2017).

204. Goldman, J. M., Hibbin, J., Kearney, L., Orchard, K. & Th'ng, K. H. HLA-DR Monoclonal Antibodies Inhibit the Proliferation of Normal and Chronic Granulocytic Leukaemia Myeloid Progenitor Cells. eng. *British Journal of Haematology* **52**, 411 (1982).

205. Pawelec, G., Ziegler, A. & Wernet, P. Dissection of Human Allostimulatory Determinants with Cloned T Cells: Stimulation Inhibition by Monoclonal Antibodies TU22, 34, 35, 36, 37, 39, 43, and 58 against Distinct Human MHC Class II Molecules. eng. *Human Immunology* **12**, 165 (1985).

206. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. eng. *Molecular & cellular proteomics: MCP* **14**, 2394 (2015).

207. Marco, M. D., Schuster, H., Backert, L., Ghosh, M., Rammensee, H.-G. & Stevanović, S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. *The Journal of Immunology* (2017).

208. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D. J., Freudenmann, L. K., Backert, L., Mühlenbruch, L., Szolek, A., Lübke, M., Wagner, P., Engler, T., Matovina, S., Wang, J., Hauri-Hohl, M., Martin, R., Kapolou, K., Walz, J. S., Velz, J., Moch, H., Regli, L., Silginer, M., Weller, M., Löffler, M. W., Erhard, F., Schlosser, A., Kohlbacher, O., Stevanović, S., Rammensee, H.-G. & Neidert, M. C. The HLA Ligand Atlas - A Resource of Natural HLA Ligands Presented on Benign Tissues. *bioRxiv*, 778944 (2020).

209. Toprak, U. H., Gillet, L. C., Maiolica, A., Navarro, P., Leitner, A. & Aebersold, R. Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics. eng. *Molecular & cellular proteomics: MCP* **13**, 2056 (2014).

210. Kalaora, S., Wolf, Y., Feferman, T., Barnea, E., Greenstein, E., Reshef, D., Tirosh, I., Reuben, A., Patkar, S., Levy, R., Quinkhardt, J., Omokoko, T., Qutob, N., Golani, O., Zhang, J., Mao, X., Song, X., Bernatchez, C., Haymaker, C., Forget, M.-A., Creasy, C., Greenberg, P., Carter, B. W., Cooper, Z. A., Rosenberg, S. A., Lotem, M., Sahin, U., Shakhar, G., Ruppin, E., Wargo, J. A., Friedman, N., Admon, A. & Samuels, Y. Combined Analysis of Antigen Presentation and T-Cell Recognition Reveals Restricted Immune Responses in Melanoma. eng. *Cancer Discovery* **8**, 1366 (2018).

211. Jiang, L., Wang, M., Lin, S., Jian, R., Li, X., Chan, J., Fang, H., Dong, G., Consortium, G., Tang, H. & Snyder, M. P. A Quantitative Proteome Map of the Human Body. *bioRxiv*, 797373 (2019).

212. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L.-H., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F., Uhlen, M., Gagneur, J., Hahne, H. & Kuster, B. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. eng. *Molecular Systems Biology* **15**, e8503 (2019).

213. Kalaora, S., Lee, J. S., Barnea, E., Levy, R., Greenberg, P., Alon, M., Yagel, G., Bar Eli, G., Oren, R., Peri, A., Patkar, S., Bitton, L., Rosenberg, S. A., Lotem, M., Levin, Y., Admon, A., Ruppin, E. & Samuels, Y. Immunoproteasome Expression Is Associated with Better Prognosis and Response to Checkpoint Therapies in Melanoma. *Nature Communications* **11**, 896 (2020).

214. Ishii, N., Chiba, M., Iizuka, M., Watanabe, H., Ishioka, T. & Masamune, O. Expression of MHC Class II Antigens (HLA-DR, -DP, and -DQ) on Human Gastric Epithelium. *Gastroenterologia Japonica* **27**, 23 (1992).

215. Chan, T., Wiltrout, R. H. & Weiss, J. M. Immunotherapeutic Modulation of the Suppressive Liver and Tumor Microenvironments. eng. *International Immunopharmacology* **11**, 879 (2011).

216. Butterfield, L. H., Ribas, A., Potter, D. M. & Economou, J. S. Spontaneous and Vaccine Induced AFP-Specific T Cell Phenotypes in Subjects with AFP-Positive Hepatocellular Cancer. eng. *Cancer immunology, immunotherapy: CII* **56**, 1931 (2007).

217. Yao, W., He, J.-C., Yang, Y., Wang, J.-M., Qian, Y.-W., Yang, T. & Ji, L. The Prognostic Value of Tumor-Infiltrating Lymphocytes in Hepatocellular Carcinoma: A Systematic Review and Meta-Analysis. eng. *Scientific Reports* **7**, 7525 (2017).

218. Unitt, E., Marshall, A., Gelson, W., Rushbrook, S. M., Davies, S., Vowler, S. L., Morris, L. S., Coleman, N. & Alexander, G. J. M. Tumour Lymphocytic Infiltrate and Recurrence of Hepatocellular Carcinoma Following Liver Transplantation. eng. *Journal of Hepatology* **45**, 246 (2006).

219. Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J. J., Cowey, C. L., Lao, C. D., Schadendorf, D., Dummer, R., Smylie, M., Rutkowski, P., Ferrucci, P. F., Hill, A., Wagstaff, J., Carlino, M. S., Haanen, J. B., Maio, M., Marquez-Rodas, I., McArthur, G. A., Ascierto, P. A., Long, G. V., Callahan, M. K., Postow, M. A., Grossmann, K., Sznol, M., Dreno, B., Bastholt, L., Yang, A., Rollin, L. M., Horak, C., Hodi, F. S. & Wolchok, J. D. Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. eng. *The New England Journal of Medicine* **373**, 23 (2015).

220. Schadendorf, D., Hodi, F. S., Robert, C., Weber, J. S., Margolin, K., Hamid, O., Patt, D., Chen, T.-T., Berman, D. M. & Wolchok, J. D. Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. eng. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **33**, 1889 (2015).

221. Schumacher, T. N. & Schreiber, R. D. Neoantigens in Cancer Immunotherapy. eng. *Science (New York, N.Y.)* **348**, 69 (2015).

222. Balachandran, V. P., Łuksza, M., Zhao, J. N., Makarov, V., Moral, J. A., Remark, R., Herbst, B., Askan, G., Bhanot, U., Senbabaoglu, Y., Wells, D. K., Cary, C. I. O., Grbovic-Huezo, O., Attiyeh, M., Medina, B., Zhang, J., Loo, J., Saglimbeni, J., Abu-Akeel, M., Zappasodi, R., Riaz, N., Smoragiewicz, M., Kelley, Z. L., Basturk, O., Australian Pancreatic Cancer Genome Initiative, Garvan Institute of Medical Research, Prince of Wales Hospital, Royal North Shore Hospital, University of Glasgow, St Vincent's Hospital, QIMR Berghofer Medical Research Institute, University of Melbourne, Centre for Cancer Research, University of Queensland, Institute for Molecular Bioscience, Bankstown Hospital, Liverpool Hospital, Royal Prince Alfred Hospital, Chris O'Brien Lifehouse, Westmead Hospital, Fremantle Hospital, St John of God Healthcare, Royal Adelaide Hospital, Flinders Medical Centre, Envoi Pathology, Princess Alexandria Hospital, Austin Hospital, Johns Hopkins Medical Institutes, ARC-Net Centre for Applied Research on Cancer, Gönen, M., Levine, A. J., Allen, P. J., Fearon, D. T., Merad, M., Gnjatic, S., Iacobuzio-Donahue, C. A., Wolchok, J. D., DeMatteo, R. P., Chan, T. A., Greenbaum, B. D., Merghoub, T. & Leach, S. D. Identification of Unique Neoantigen Qualities in Long-Term Survivors of Pancreatic Cancer. eng. *Nature* **551**, 512 (2017).

223. Gubin, M. M. & Schreiber, R. D. CANCER. The Odds of Immunotherapy Success. eng. *Science (New York, N.Y.)* **350**, 158 (2015).

224. Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., Hollmann, T. J., Bruggeman, C., Kannan, K., Li, Y., Elipenahli, C., Liu, C., Harbison, C. T., Wang, L., Ribas, A., Wolchok, J. D. & Chan, T. A. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. eng. *The New England Journal of Medicine* **371**, 2189 (2014).

225. Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., Garon, E. B., Merghoub, T., Wolchok, J. D., Schumacher, T. N. & Chan, T. A. Cancer Immunology. Mutational Landscape Determines Sensitivity to PD-1 Blockade in Non-Small Cell Lung Cancer. eng. *Science (New York, N.Y.)* **348**, 124 (2015).

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

227. van Rooij, N., van Buuren, M. M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., van Dijk, L. J. A., Behjati, S., Hilkmann, H., El Atmioui, D., Nieuwland, M., Stratton, M. R., Kerkhoven, R. M., Kesmir, C., Haanen, J. B., Kvistborg, P. & Schumacher, T. N. Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. eng. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **31**, e439 (2013).

228. Tran, E., Turcotte, S., Gros, A., Robbins, P. F., Lu, Y.-C., Dudley, M. E., Wunderlich, J. R., Somerville, R. P., Hogan, K., Hinrichs, C. S., Parkhurst, M. R., Yang, J. C. & Rosenberg, S. A. Cancer Immunotherapy Based on Mutation-Specific CD4+ T Cells in a Patient with Epithelial Cancer. eng. *Science (New York, N.Y.)* **344**, 641 (2014).

229. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J. & Stratton, M. R. Signatures of Mutational Processes in Human Cancer. eng. *Nature* **500**, 415 (2013).

230. Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S. & Getz, G. Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types. eng. *Nature* **505**, 495 (2014).

231. Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T. D., Gong, Z., Gao, H., Hao, K., Willard, M. D., Xu, J., Hauptschein, R., Rejto, P. A., Fernandez, J., Wang, G., Zhang, Q., Wang, B., Chen, R., Wang, J., Lee, N. P., Zhou, W., Lin, Z., Peng, Z., Yi, K., Chen, S., Li, L., Fan, X., Yang, J., Ye, R., Ju, J., Wang, K., Estrella, H., Deng, S., Wei, P., Qiu, M., Wulur, I. H., Liu, J., Ehsani, M. E., Zhang, C., Loboda, A., Sung, W. K., Aggarwal, A., Poon, R. T., Fan, S. T., Wang, J., Hardwick, J., Reinhard, C., Dai, H., Li, Y., Luk, J. M. & Mao, M. Whole-Genome Sequencing Identifies Recurrent Mutations in Hepatocellular Carcinoma. eng. *Genome Research* **23**, 1422 (2013).

232. El-Khoueiry, A. B., Sangro, B., Yau, T., Crocenzi, T. S., Kudo, M., Hsu, C., Kim, T.-Y., Choo, S.-P., Trojan, J., Welling, T. H., Meyer, T., Kang, Y.-K., Yeo, W., Chopra, A., Anderson, J., Dela Cruz, C., Lang, L., Neely, J., Tang, H., Dastani, H. B. & Melero, I. Nivolumab in Patients with Advanced Hepatocellular Carcinoma (CheckMate 040): An Open-Label,

Non-Comparative, Phase 1/2 Dose Escalation and Expansion Trial. eng. *Lancet (London, England)* **389**, 2492 (2017).

233. Kowalewski, D. J. & Stevanović, S. Biochemical Large-Scale Identification of MHC Class I Ligands. eng. *Methods in Molecular Biology (Clifton, N.J.)* **960**, 145 (2013).

234. Löffler, M. W., Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., Schroeder, C. M., Trautwein, N., Hilke, F. J., Zinser, R. S., Mühlenbruch, L., Kowalewski, D. J., Schuster, H., Sturm, M., Matthes, J., Riess, O., Czemmel, S., Nahnsen, S., Königsrainer, I., Thiel, K., Nadalin, S., Beckert, S., Bösmüller, H., Fend, F., Velic, A., Maček, B., Haen, S. P., Buonaguro, L., Kohlbacher, O., Stevanović, S., Königsrainer, A., Rammensee, H.-G. & HEPAVAC Consortium. Multi-Omics Discovery of Exome-Derived Neoantigens in Hepatocellular Carcinoma. *Genome Medicine* **11**, 28 (2019).

235. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C. & MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. eng. *Bioinformatics (Oxford, England)* **26**, 966 (2010).

236. Pino, L. K., Searle, B. C., Bollinger, J. G., Nunn, B., MacLean, B. & MacCoss, M. J. The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics. eng. *Mass Spectrometry Reviews* **39**, 229 (2020).

237. Sturm, M., Schroeder, C. & Bauer, P. SeqPurge: Highly-Sensitive Adapter Trimming for Paired-End NGS Data. *BMC Bioinformatics* **17**, 208 (2016).

238. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*. Comment: 3 pages and 1 color figure (2013).

239. Faust, G. G. & Hall, I. M. SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction. *Bioinformatics* **30**, 2503 (2014).

240. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: Improved Coding Indel Detection via Assembly-Based Realignment. *Bioinformatics* **30**, 2813 (2014).

241. Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J. & Cheetham, R. K. Strelka: Accurate Somatic Small-Variant Calling from Sequenced Tumor-Normal Sample Pairs. eng. *Bioinformatics (Oxford, England)* **28**, 1811 (2012).

242. Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P. & Saunders, C. T. Strelka2: Fast and Accurate Calling of Germline and Somatic Variants. eng. *Nature Methods* **15**, 591 (2018).

243. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3. eng. *Fly* **6**, 80 (2012 Apr-Jun).

244. Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M. & Lu, X. Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. eng. *Frontiers in Genetics* **3**, 35 (2012).

245. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. *Human Mutation* **32**, 894 (2011).

246. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* **29**, 15 (2013).

247. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: A Fast and Accurate Adapter Trimmer for next-Generation Sequencing Paired-End Reads. *BMC Bioinformatics* **15**, 182 (2014).

248. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

249. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M. & Kohlbacher, O. OptiType: Precision HLA Typing from next-Generation Sequencing Data. eng. *Bioinformatics (Oxford, England)* **30**, 3310 (2014).

250. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. L. TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions. *Genome Biology* **14**, R36 (2013).

251. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python Framework to Work with High-Throughput Sequencing Data. *Bioinformatics* **31**, 166 (2015).

252. Qian, G.-S., Kuang, S.-Y., He, X., Groopman, J. D. & Jackson, P. E. Sensitivity of Electrospray Ionization Mass Spectrometry Detection of Codon 249 Mutations in the P53 Gene Compared with RFLP. eng. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* **11**, 1126 (2002).

253. Almeida, L. G., Sakabe, N. J., deOliveira, A. R., Silva, M. C. C., Mundstein, A. S., Cohen, T., Chen, Y.-T., Chua, R., Gurung, S., Gnjatic, S., Jungbluth, A. A., Caballero, O. L., Bairoch, A., Kiesler, E., White, S. L., Simpson, A. J. G., Old, L. J., Camargo, A. A. & Vasconcelos, A. T. R. CTdatabase: A Knowledge-Base of High-Throughput and Curated Data on Cancer-Testis Antigens. eng. *Nucleic Acids Research* **37**, D816 (2009).

254. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.-P., Gendron, P., Courcelles, M., Hardy, M.-P., Côté, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., Haddad, E., Lemieux, S., Thibault, P. & Perreault, C. Noncoding Regions Are the Main Source of Targetable Tumor-Specific Antigens. *Science Translational Medicine* **10**, eaau5516 (2018).

255. Wen, B., Wang, X. & Zhang, B. PepQuery Enables Fast, Accurate, and Convenient Proteomic Validation of Novel Genomic Alterations. eng. *Genome Research* **29**, 485 (2019).

256. Haen, S. P., Löffler, M. W., Rammensee, H.-G. & Brossart, P. Towards New Horizons: Characterization, Classification and Implications of the Tumour Antigenic Repertoire. *Nature Reviews Clinical Oncology*, 1 (2020).