

Exploring Alternative Sources Of Tumor Antigens Using Large-Scale Immunopeptidomics.

Georges Bedran

PhD Thesis



Unraveling Tumor Antigens



Exploring alternative sources of tumor antigens using large-scale immunopeptidomics

By Georges **Bedran**

Supervisors

Primary supervisor: *Prof.* Theodore Hupp

Auxiliary supervisor: *Dr.* Javier Alfaro

A dissertation submitted to the Intercollegiate Faculty of Biotechnology UG&MUG in partial fulfillment of the requirements for the degree of Doctor of Philosophy (bioinformatics),
University of Gdańsk.

International Centre for Cancer Vaccine Science

Intercollegiate Faculty of Biotechnology UG & MUG

University of Gdańsk

March 17, 2023



Georges Bedran

georges.bedran@phdstud.ug.edu.pl

gbadran_90@live.com

ORCID ID: 0000-0002-3086-3742

© Georges Bedran 2023

Preface

I am delighted to present this dissertation, which comprises an in-depth investigation of the immunopeptidome. The research presented in here draws on my passion for cancer research and my commitment to advancing our understanding of the mechanisms underlying tumor-immune interactions.

The presented work builds upon two previous publications of mine, which have been incorporated into Chapters 2 and 3, respectively. In accordance with the plagiarism check policies of the University of Gdansk, both chapters were embedded into the dissertation as is, and each has its own bibliographic references.

Chapter 2, titled "The Immunopeptidome from a Genomic Perspective: Establishing the Non-Canonical Landscape of MHC Class I-Associated Peptides" has been accepted for publication in Cancer Immunology Research. This chapter delves into the non-canonical landscape of MHC class I-associated peptides from a genomic perspective.

Meanwhile, Chapter 3, titled "HLA-Glyco: A Large-Scale Interrogation of the Glycosylated Immunopeptidome" was previously published as a pre-print on bioRxiv (DOI: <https://doi.org/10.1101/2022.12.05.519200>). This chapter explores the glycosylation landscape of MHC class II-associated peptides, providing a comprehensive analysis of the glycosylated immunopeptidome.

I am grateful to my supervisors, colleagues, and collaborators for their guidance and support throughout this research endeavor. I hope that this dissertation will contribute to the advancement of cancer immunology and inspire future research in this exciting field.

Georges Bedran.

TABLE OF CONTENT

ACKNOWLEDGEMENTS

Abstract in English

Abstract in Polish

Chapter 1: Introduction	1
The major histocompatibility complex (MHC)	2
<i>MHC class I</i>	3
<i>MHC class II</i>	4
<i>MHC molecules and MHC-associated peptides</i>	6
Cancer and the immune system	6
Selection of neoantigen candidates	7
<i>Indirect identification</i>	9
<i>Direct identification</i>	10
Promising sources of antigens	10
<i>Genomic variants</i>	12
<i>Transcriptomic variants</i>	12
<i>Proteomic variants</i>	13
Thesis outline	13
<i>Key technical aims</i>	14
<i>Key biological aims</i>	22
References	24
Chapter 2	25
Abstract	27
Introduction	27
Materials and methods	27
<i>Dataset selection</i>	28
<i>Proteogenomic database generation</i>	31
<i>MS computational analysis</i>	31
<i>Alignment of immunopeptides to the genome</i>	31
<i>Open reading frame analysis</i>	32
<i>Intron retention analysis</i>	
<i>Frameshift mutation analysis</i>	
<i>Comparison of the identified non-canonical MHC class I-associated peptides between studies</i>	33
<i>Cancer selectivity of the non-canonical MHC class I-associated peptides</i>	33
<i>Code availability</i>	35
<i>Data availability</i>	35
Results	35
<i>Immunopeptidomic MS datasets</i>	35
<i>Closed Open De novo – deep immunopeptidomics pipeline (COD-dipp)</i>	36
<i>ptmMAPs</i>	37
<i>ncMAPs</i>	37

Discussion	43
Acknowledgments	46
Author Contributions	47
References	47
Tables	53
Figures	54
<i>Figure 1</i>	54
<i>Figure 2</i>	55
<i>Figure 3</i>	56
<i>Figure 4</i>	57
<i>Figure 5</i>	58
<i>Figure 6</i>	59
Supplementary Figures S1-S7	60
<i>Figure S1</i>	60
<i>Figure S2</i>	61
<i>Figure S3</i>	62
<i>Figure S4</i>	63
<i>Figure S5</i>	64
<i>Figure S6</i>	65
<i>Figure S7</i>	66
Supplementary Notes S1-2	67
<i>Supplementary Notes</i>	67
<i>Note 1: Dataset selection</i>	67
<i>Note 2: Correctness of the identified peptides</i>	68
Chapter 3	71
Abstract	71
Introduction	73
Results	74
<i>Computational glyco-immunopeptidomics workflow</i>	74
<i>Large multi-tissue MHC immunopeptidome dataset</i>	76
<i>Enrichment of N-glycosylation in the class II immunopeptidome</i>	77
<i>Glycosylation of MAPs does not influence the HLA binding motif</i>	79
<i>Deconvolution of peptides using a semi-supervised approach</i>	79
<i>Deconvolution of peptides using a fully unsupervised approach</i>	80
<i>The HLA class II N-glycosylation characteristics</i>	81
Discussion	82
Methods	84
<i>Dataset selection</i>	84
<i>Mass spectrometry N-glycan search</i>	84
<i>FDR control</i>	85
<i>Deconvolution of the MHC associated peptides</i>	86
<i>Figure generation</i>	86
Authorship contribution	87

Acknowledgments	87
References	87
Figures	94
<i>Figure 1</i>	94
<i>Figure 2</i>	95
<i>Figure 3</i>	96
<i>Figure 3</i>	96
<i>Figure 4</i>	97
<i>Figure 5</i>	98
<i>Figure 6</i>	99
Supplementary materials	100
<i>Supplementary Figure 1</i>	100
Chapter 4	101
Chapter 4: Summary, milestones, and future directions	101
<i>Summary and highlights of the presented work</i>	101
<i>Beyond antigen presentation, towards antigen recognition</i>	103
<i>References</i>	110
Appendix 1: A technical guide for the COD-dipp pipeline	117
General description	117
<i>Applications of the method</i>	119
<i>Experimental design</i>	119
<i>Expertise needed to implement the protocol</i>	119
<i>Hardware requirements</i>	119
<i>Software requirements</i>	120
PROCEDURE	120
<i>Annotation generation step</i>	120
<i>Raw data conversion</i>	121
<i>Environment setup</i>	122
<i>Launching the analysis</i>	124
TROUBLESHOOTING	125
TIMING	125
ANTICIPATED RESULTS	126
Supporting document 1: Contribution statement from the co-authors of chapter 2	129
Supporting document 2: Acceptance letter from cancer immunology research	130
Supporting document 3: Contribution statement from the co-authors of chapter 3	132

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my PhD advisor, Dr. Javier A. Alfaro, for his help, advice, and motivation in many aspects of my PhD. Dr. Alfaro has been a constant source of encouragement, has pushed me out of my comfort zone, and has contributed, to a great extent, to the quality of this scientific work. I would also like to thank my advisor Prof. Ted Hupp for allowing this project to take place within the broad vision of the International Centre for Cancer Vaccine Science (ICCVS). I wish to extend my gratitude to my research colleagues for their input, expert advice, and support through in-person/virtual meetings, casual conversations, and emails. On the same note, a special acknowledgment for all the research participants who helped revise and shape the final manuscripts.

I would like to express my appreciation to the ICCVS for funding this PhD. The ICCVS is a project within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. I would also like to thank CI-TASK, Gdansk, and PL-Grid Infrastructure, Poland, for providing hardware and software resources.

I am also thankful for the opportunity of performing a 5-month internship at the University of Michigan Medical School within Prof. Alexey I. Nesvizhskii's lab. Many thanks go to Prof. Nesvizhskii and Dr. Daniel A. Polasky, who instructed me on the intricacies of glyco-proteomics. I wish to extend my gratitude to both the ICCVS and Prof. Nesvizhskii for their funding.

This endeavor would not have been possible without the tremendous understanding, encouragement, mental support, and sympathetic ear of my beloved wife. Finally, I would be remiss in not mentioning my parents for their support throughout this journey and my friend Gordon for providing happy distractions that rested my mind outside the research zone.

Abstract in English

The identification of cancer neoantigens is propelling a new era of vaccines and antigen-specific T cell therapies. Mass spectrometry has been the sole high-throughput approach for characterizing the physical presence of neoantigens in cancer. Early efforts to investigate antigen presentation focused on combining publicly available studies to query canonical MHC-associated peptides (MAPs). However, the profiling of non-conventional antigens, such as non-canonical (i.e., translation of non-coding regions) and post-translationally modified MHC-associated peptides, remains limited and is rarely clearly understood.

In Chapter Two, I developed a proteogenomic pipeline based on deep learning *de novo* mass spectrometry to enable the discovery of non-canonical MHC-associated peptides (ncMAPs) from non-coding regions. Considering that the emergence of tumor antigens can also involve post-translational modifications, an open search component was included in the pipeline. Leveraging the wealth of mass spectrometry-based immunopeptidomics, I analyzed 26 MHC class I immunopeptidomic studies of eleven different cancer types. I validated the *de novo* identified ncMAPs, along with the most abundant post-translational modifications, using spectral matching and controlled their false discovery rate (FDR) to 1%. Interestingly, the non-canonical presentation appeared to be 5 times enriched for the A03 HLA supertype, with a projected population coverage of 54.85%. I revealed an atlas of 8,601 ncMAPs with varying levels of cancer selectivity and suggested 17 cancer-selective ncMAPs as attractive targets according to a stringent cutoff.

In Chapter Three, I developed a glyco-immunopeptidomics method using the ultrafast glycopeptide search of MSFragger and several layers of stringent control of false discovery rates. I performed a harmonized large-scale analysis of eight publicly available studies to produce a resource containing over 3,400 HLA class II glycopeptides from 1,049 distinct

protein-glycosylation sites. I revealed characteristics in HLA glycopeptides, including high levels of truncated glycans, conserved HLA-binding cores across the 72 studied HLA class II alleles, and a different glycosylation positional specificity between the classical allele groups. With the goal of supporting further development in the nascent field of glyco-immunopeptidomics, I provided a reproducible glyco-immunopeptidomics pipeline within the fragpipe suite along with a web resource for ease of access.

In Chapter Four, I conclude this thesis with a summary of my findings, a discussion of the unmet needs in the field, and my vision of the research to come.

The establishment of both the non-canonical and glycosylated landscapes of MHC-associated peptides within the framework of my PhD represents a milestone towards understanding the complexity of the immunopeptidome and paves the way for broader therapeutic research against cancer.

Abstract in Polish

Identyfikacja antygenów nowotworowych rozpoczyna nową erę szczepionek przeciwnowotworowych i terapii z wykorzystaniem antygenowo-specyficzných limfocytów T. Spektrometria mas jest natomiast obecnie jedyną metodą, która umożliwia scharakteryzowanie fizycznej obecności antygenów nowotworowych.

Wczesne badania prezentacji antygenów wykorzystywały głównie publicznie dostępne dane w celu identyfikacji kanonicznych peptydów prezentowanych przez cząsteczki MHC (MAP). Jednakże profilowanie niekonwencjonalnych antygenów, takich jak peptydy niekanoniczne (np. będące produktem translacji regionów niekodujących) czy peptydy zmodyfikowane potranslacyjnie, pozostaje ograniczone i nie jest w pełni scharakeryzowane.

W rozdziale drugim opisałem zaprojektowany przeze mnie proteogenomiczny system przetwarzania potokowego oparty na spektrometrii mas de novo z głębokim uczeniem, który umożliwia wykrycie niekanonicznych peptydów pochodzących z regionów niekodujących prezentowanych przez cząsteczki MHC (ncMAP). Biorąc pod uwagę, że antygeny nowotworowe mogą również powstawać w wyniku modyfikacji potranslacyjnych, w systemie tym uwzględniono element wyszukiwania otwartego. Wykorzystując szerokie zasoby publicznych baz danych, przeanalizowałem 26 badań, które z zastosowaniem spektrometrii mas identyfikowały peptydy prezentowane przez MHC klasy I w 9 różnych typach nowotworów. Zweryfikowałem zidentyfikowane de novo ncMAP, wraz z najliczniejszymi modyfikacjami potranslacyjnymi, używając dopasowania widmowego i ograniczając oczekiwaną proporcję błędów I rodzaju wśród wyników istotnych statystycznie (ang. false discovery rate; FDR) do 1%. Warty podkreślenia jest fakt, że niekanoniczna prezentacja była 5-krotnie częstsza w przypadku HLA- A03, przy przewidywanym pokryciu w populacji na poziomie 54,85%. Ponadto, przedstawiłem zbiór 8601 ncMAP o różnych poziomach specyficzności dla nowotworów i wskazałem, zgodnie z rygorystycznym punktem odcięcia, 17 ncMAP specyficznych dla nowotworów, które stanowią potencjalne cele terapeutyczne.

W rozdziale trzecim przedstawiłem nową metodę glikoimmunopeptydomiczną wykorzystującą ultra szybkie wyszukiwanie glikopeptydów za pomocą narzędzia MSFragger oraz przedstawiłem kilka etapów zapewniających ścisłą kontrolę błędów I rodzaju wśród wyników istotnych statystycznie (FDR). Przeprowadziłem zharmonizowaną, zakrojoną na szeroką skalę analizę 8 publicznie dostępnych badań, aby utworzyć zasób zawierający ponad 3400 glikopeptydów prezentowanych przez anygeny HLA klasy II wywodzących się z 1049 różnych regionów glikozylacji białek. Przedstawiłem ponadto cechy charakterystyczne dla glikopeptydów prezentowanych przez HLA, wśród których często obserwuje się skrócone glikany, peptydy z konserwatywnym rdzeniem wiążącym HLA (zidentyfikowane w 72 badanych allelach HLA klasy II) oraz różną swoistość pozycji glikozylacji. Mając na celu wspieranie dalszego rozwoju glikoimmunopeptydomiki, udostępniłem system włączony do pakietu fragpipe umożliwiający powtarzalną analizę glikoimmunopeptydomu. System jest połączony z zasobami internetowymi, co ułatwia dostęp.

W rozdziale czwartym zakończyłem dysertację podsumowaniem wszystkich obserwacji, dyskusją na temat niezaspokojonych potrzeb medycznych w przedstawionej dziedzinie oraz wizją przyszłych badań. Jestem przekonany, że opracowana w ramach niniejszej pracy doktorskiej sygnatura niekanonicznych oraz glikozylowanych peptydów prezentowanych przez cząsteczki MHC stanowi kamień milowy w kierunku zrozumienia złożoności immunopeptydomu oraz toruje drogę do szerszych badań nad terapiami przeciwnowotworowymi.

Chapter 1: Introduction

The major histocompatibility complex (MHC)

The capacity of the immune system to differentiate between self and non-self is crucial owing to the continuous exposure of our bodies to diseases and pathogens. The ability to differentiate between the two is governed by the presentation of antigens and their recognition by the immune cells. The Major Histocompatibility Complex (MHC), also known as the Human Leukocyte Antigen (HLA) in humans, is a group of genes that when translated into proteins, bind intra/extracellular components for immune monitoring¹. The MHC system is responsible for antigen presentation through two classical pathways, termed class I and II.

MHC class I

To begin with MHC class I, all nucleated cells present peptides derived from cytosolic protein turnover at the cell surface (see **Figure 1a**). Under healthy and diseased conditions, these proteins are degraded by the proteasome. The resultant peptides can be further trimmed by several cytoplasmic peptidases such as tripeptidyl peptidase II, leucine aminopeptidase, and bleomycin hydrolase. These peptides are then transported to the endoplasmic reticulum (ER) by a transporter associated with antigen processing (TAP). Through transient interactions with the chaperones calnexin, calreticulin, and tapasin, the peptides are further processed by ER-resident aminopeptidase ERAP1 and ERAP2 and loaded onto the nascent HLA class I heavy chain. This MHC-I-peptide complex passes through the Golgi apparatus for glycosylation, enters a secretory vesicle, and fuses with the cell membrane. This process is referred to as MHC class I presentation and serves as immune monitoring of the self, where CD8+ T lymphocytes circulate and attack cells presenting foreign MHC-associated peptides, referred to as antigens².

The MHC Class I system is composed of three classical genes (HLA-A, -B, and -C) and some non-classical genes such as HLA-E and -G. Unlike classical genes, HLA-E is an oligomorphic HLA molecule that has just few alleles³ (*i.e.*, not polymorphic). HLA-G is primarily expressed in trophoblasts, which develop in the placenta during pregnancy. Particularly in pregnancy, HLA-G prevents the fetus from being identified as a foreign entity by the mother's immune system. HLA-G is also expressed in some cancers and has been associated with tumor immune escape⁴. In summary, both HLA-G and -E have a small peptide-binding repertoire and are involved in regulating the immune response.

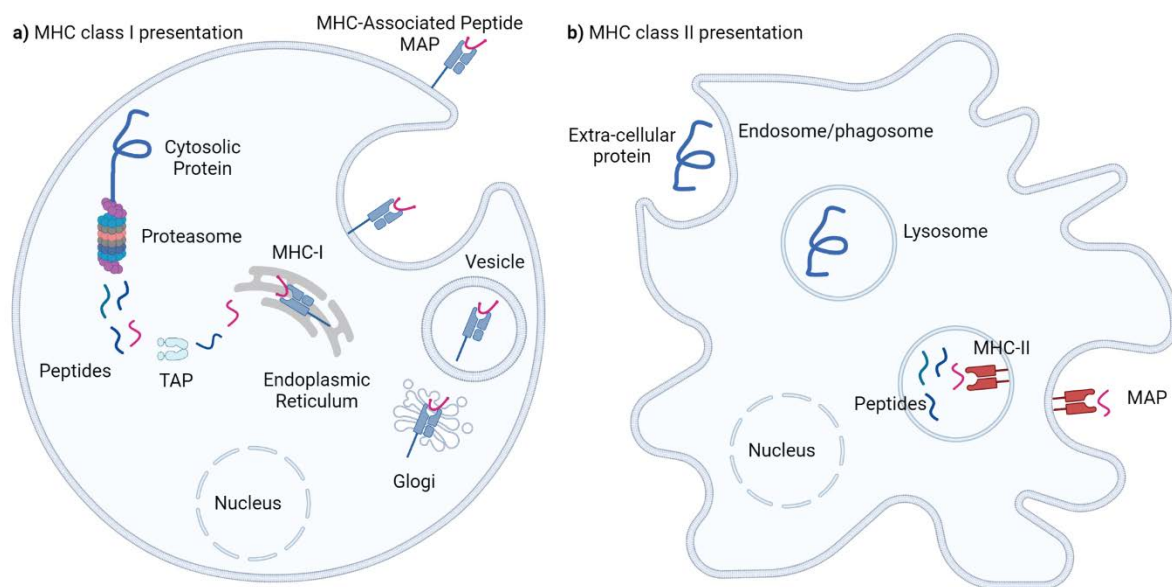


Figure 1: Antigen processing and presentation pathways. a) MHC class I processing and presentation pathway within nucleated cells. **b)** MHC class II processing and presentation pathway within antigen presenting cells such as monocytes, macrophages, and dendritic cells. Created with [BioRender.com](https://www.biorender.com).

MHC class II

In contrast, antigen-presenting cells (APC), such as monocytes, macrophages, and dendritic cells, incorporate extracellular proteins through phagocytosis or endocytosis of B cells (see **Figure 1b**). In the endosome, proteins are trimmed into peptides by either proteases or non-enzymatic cleavage and interact with MHC II molecules. MHC class II molecules are initially

synthesized in the endoplasmic reticulum, where they assemble with the invariant chain. This chain is important for stability, appropriate folding, blocking cellular peptide binding, and export into specialized endosomes/lysosomes. The class II-associated invariant chain peptide (CLIP) is created by proteolytic degradation of the invariant chain by cathepsin and resides in the MHC groove as a replacement for the associated peptides. The interchange of CLIP with high-affinity peptides is facilitated by several conditions such as low pH, endosomal proteases, and support from an unconventional MHC class II protein, HLA-DM. The HLA class II-peptide complex is then delivered to the cell surface by the trans-Golgi apparatus for presentation to the cognate CD4⁺ T lymphocytes⁵.

The expression of MHC class II-associated antigens in various normal tissues has been examined with the help of specific monoclonal antibodies, demonstrating that these antigens are more widely distributed in normal tissues than previously thought. In addition to APC, various tissues can constitutively express MHC class II antigens, as evidenced by the weak to moderate expression of HLA class II antigens in skin, breast, lung, and renal tissues⁶⁻⁹.

MHC molecules and MHC-associated peptides

HLA class I and class II molecules are composed of two polypeptide chains that come together to form a structure that can accommodate short peptides. In class I, the immunoglobulin superfamily α -chain and β 2-microglobulin form a heterodimer known as a class I molecular complex. The HLA molecule is composed of three extracellular domains (α 1, α 2, and α 3) with a cytoplasmic anchor that traverses the cell membrane. These α chains are encoded by the highly polymorphic A, B, and C genes. In class II, two α - and two β -noncovalently linked and non-identical transmembrane glycoprotein chains make up the molecule. These chains are coded by the highly polymorphic DP, DQ, and DR genes¹.

MHC-associated peptides are set in the peptide-binding grooves on the top surface of HLA molecules. They range from 8 to 11 amino acids long for class I and 13 to 25 amino acids long

for class II. They are held within the grooves by conserved binding motifs that can vary between HLA molecules. The major structural variations between the two MHC classes reside in the peptide-binding groove being closed at both ends for class I compared to being open at both ends for class II.

Cancer and the immune system

Cancer is a disease characterized by adaptive evolutionary growth. As the human body contains trillions of cells, cancer can develop in nearly all tissues. Human cells typically proliferate and divide to create new cells as the body requires them. Cells die when they age or become damaged and are replaced by new ones. When this routine process fails, aberrant cells begin to proliferate and develop into tumors that can be either cancerous (malignant) or non-cancerous (benign). Cancer is a dynamic disease characterized by genetic and epigenetic mutations as well as hereditary changes in gene expression that are transmitted to subsequent generations of cells as the tumor progresses. The most well-understood changes involve alterations in the genes that control cellular behavior, particularly how they grow, divide, and survive in their local microenvironment. Cancer-causing genetic alterations can result from many different environmental factors including (I) cell division mistakes; (II) DNA damage from exogenous exposure to substances such as tobacco smoke, bile acid reflux, or UV light; and (III) pathogens such as Human Papilloma Virus, Epstein Barr Virus, and fungal infections. Furthermore, cancer has a hereditary component, in which the passed-on alleles of certain genes can confer a predisposition to the disease.

The eight hallmarks of cancer comprise the acquired capabilities for sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing/accessing vasculature, activating invasion and metastasis, reprogramming cellular metabolism, unlocking phenotypic plasticity, non-mutational epigenetic reprogramming, polymorphic microbiomes, senescence, and avoiding immune destruction¹⁰⁻¹². Cancer

treatment remains challenging owing to a variety of factors, such as the immune-dependent remodeling of cancer tissue into an adaptive state, significant heterogeneity, and numerous genetic alterations. In addition, cancer can affect a variety of organs and is not static. Instead, it develops and advances over time, accumulating additional mutations. Surgery, radiation, and chemotherapy are the three types of traditional cancer treatments. Despite their adverse effects on healthy tissues, radiation therapy and chemotherapy remain crucial parts of cancer treatment today. Several novel approaches have emerged that offer significant promise for cancer therapy. These include photodynamic therapy (destroying tumor cells using a photosensitizing drug activated by specific wavelengths of light), photothermal therapy (using a photothermal agent activated by light-producing heat to damage tumor cells), nanoparticle drug therapy (tumor-directed drug delivery), and gene therapy (immunotherapy and vaccines)¹³.

Prior to the advent of immune checkpoint inhibitors (ICI), immunotherapy was based on very toxic and mostly ineffective immunocytokines such as interleukin-2 and alpha-interferon¹⁴. Immune checkpoints are essential for the preservation of self-tolerance (i.e., prevention of autoimmunity) and tissue protection, while the immune system responds to pathogenic infection under normal physiological conditions¹⁵. Malignancies found ways to evade antitumor immune responses by dysregulating immune checkpoint proteins forcing immune resistance¹⁶. The 2018 Nobel Prize in Medicine was awarded to James Allison and Tasuku Honjo, two immunologists who were responsible for drafting the idea of ICI-based immunotherapy, illustrating its major success¹⁷. Allison observed that cytotoxic T cell antigen 4 (CTLA-4), a protein encoded on the surface of T lymphocytes, blocks T cell function, and that once it is blocked, cancer cells are successfully eliminated in mice¹⁸. Similarly, Honjo showed that programmed cell death protein 1 (PD-1), like CTLA-4, functions as a T cell down-regulator, but operates via a different mechanism¹⁹. Unlike previous studies, this new principle targets the immune system, instead of cancer cells, by reactivating the immune response. The

first approved drug against CTLA-4 was released in 2011 (ipilimumab²⁰) with very convincing outcomes. Next, anti PD-1²¹ monoclonal antibodies (nivolumab and pembrolizumab) and anti-PDL1 antibodies (atezolizumab and durvalumab) were developed, marking a major step in cancer treatment.

ICI have made significant advancements in cancer treatment; however, few challenges continue to limit its development, since the response rate varies from 10% to 50% with certain solid tumor types. These challenges stem in part from the tumor microenvironment. As T cells are often the primary targets of immune checkpoint inhibitors, effector T cell infiltration in solid tumors is a unique characteristic of patients who respond well to therapy. As a result, only a small percentage of patients with solid tumors benefit from immune checkpoint inhibitors. The remaining cancer patients are unlikely to respond to single-agent therapy due to a scarcity of targets, as their tumors appear to be depleted of effector immune cells. Immunotherapy based on cancer vaccines may overcome the resistance of some malignancies to immune checkpoint inhibitors. With cancer vaccination enhancing effector T-cell infiltration into tumors and ICI releasing the brakes, combination immunotherapy unites the best qualities of each immunotherapy technique^{22,23}. However, optimizing the set of 'neoantigens' to pursue vaccine development remains a challenge.

Selection of neoantigen candidates

Currently, antigenic peptides can be detected both indirectly through predictive genomics and directly through immunopeptidomics.

Indirect identification of MHC-associated peptides by next-generation sequencing

Owing to the significant decrease in cost and time required for next-generation sequencing (NGS), many studies focusing on cancer neoantigens have utilized *in silico* prediction tools.

The indirect identification of neoantigens can be succinctly described using the following steps: (I) NGS data comparison between normal and tumor samples to identify tumor-specific genetic variants^{24,25} (*i.e.*, mutation calling), (II) HLA typing^{26,27}, (III) neoantigen prioritization based on HLA binding affinity^{28–30}, and (IV) validation of immunogenicity³¹. Considering the highly polymorphic nature of the MHC system³², HLA typing is one of the most crucial steps for determining the mixture of HLA alleles present in a sample. Next, the identification and prioritization of neoantigens *in silico* heavily relies on predictive models to shortlist a presentable set of mutated peptides according to the HLA genotype of the sample. Lastly, in addition to MHC presentation, neoantigens need to be immunogenic, that is, effectively recognizable by T cells to trigger an immune response. These approaches are often validated using the ‘tetramer assay”, which can detect T cells present in the human body that can bind to such mutated peptides. Although fast and inexpensive, indirect methods lead to suboptimal neoantigen prioritization owing to the discrepancy between the theoretically possible MHC-associated peptides and the experimentally presented ones³³. Moreover, indirect methods do not allow going beyond translational events to monitor post-translational ones.

Direct identification of MHC-associated peptides by mass spectrometry

In addition to neoantigens resulting from DNA mutations, direct identification of MHC-associated peptides by mass spectrometry allows the detection of neoantigens from alternative sources, such as peptides bearing post-translational modifications^{34–36} and peptides originating from regions beyond the boundary of the known coding genome^{37–42}. Nevertheless, this method has not yet been widely adopted in vaccine development, for technical reasons. The fundamental idea behind MS is that by manipulating ions with electric and/or magnetic fields in vacuum, one may determine the masses of the analytes. The mass of an analyte can be determined in a variety of ways, such as by measuring the oscillating image current produced by ions orbiting in an electrostatic trap⁴³ (for Orbitrap-type analyzers)

or the amount of time it takes for an ion to travel a specific distance (for time-of-flight mass analyzers), which depends only on the mass-to-charge ratio (m/z) of the ion⁴⁴. However, m/z values are typically insufficient to identify an analyte. Tandem mass spectrometry (MS/MS or MS²) addresses this issue by carrying out numerous rounds of mass analysis to obtain more information regarding the analyte. Following the analysis of an intact analyte, the analyte is broken up most frequently by purposeful collisions with inert gas molecules, and the broken-up molecules are then mass-measured. A molecular fingerprint, known as the MS/MS spectrum, can provide details regarding the substructures of a molecule. Tandem MS facilitates peptide identification by providing sequence information in an MS/MS spectrum. Peptides are broken down into fragment ions that reveal the amino acid sequence, thus creating complementary ions that appear as charged entities. This occurs repeatedly at varying locations on the peptide backbone, yielding varying ion m/z values corresponding to amino acid masses. These predictable fragmentation paths enable peptide sequence analysis.

A technique that scans MS/MS spectra for peptide sequences is the foundation of peptide identification. These can be divided into two groups: *de novo* search algorithms and database search algorithms⁴⁵; however, many methods combine aspects from both. De novo algorithms, such as PEAKS⁴⁶ and DeepNovo^{47,48}, analyze the peaks in the spectrum, determine how far apart the peaks are from one another (which might correspond to the masses of the amino acids), and then determine how closely the spectrum matches the theoretical peptide. Search methods for databases such as MS-GF+⁴⁹ and MSFragger⁵⁰ rely on a reference database of protein sequences that are anticipated to be in the sample. Regardless of the algorithm, the MS/MS spectra are scored against hypothetical ones with known peptide sequences to evaluate their closeness. The most evident method for generating such a set of labeled theoretical spectra is by non-specific cleavage of the proteome^{51,52}, that is, non-specific cleavage of known proteins at every peptide bond.

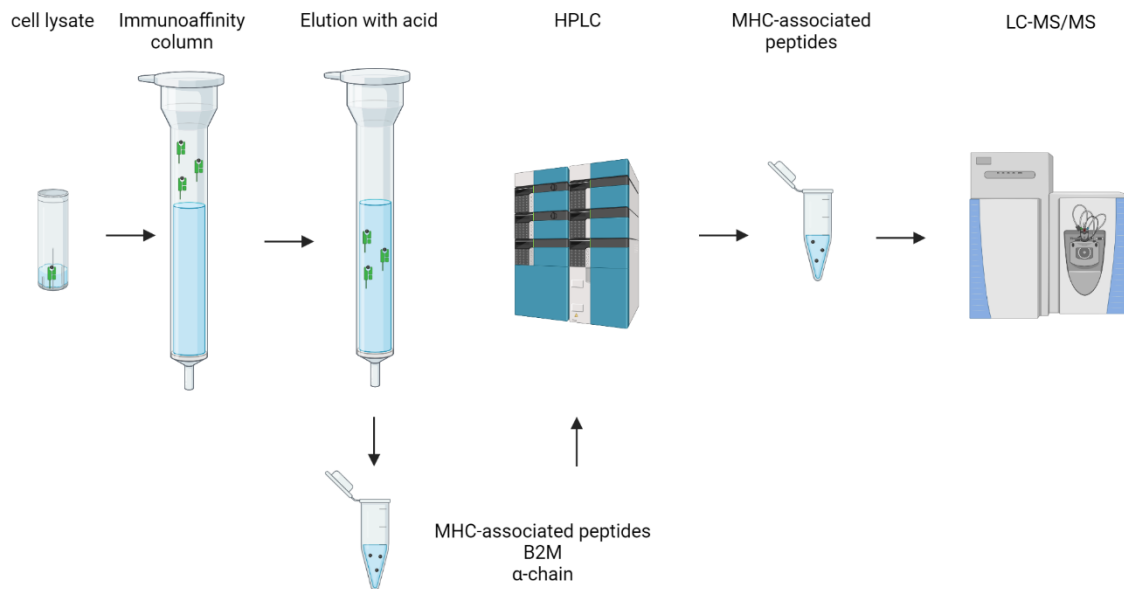


Figure 2: General workflow for purification of MHC-associated peptides through immunoprecipitation. Created with [BioRender.com](https://www.biorender.com).

Mass spectrometry is the most widely used technique for identifying MHC-associated peptides (MAPs) at the immunopeptidome level; that is, a subset of processed and presented subsequences of proteins at the cell surface⁵³. MAPs are typically isolated using immunoprecipitation (IP) columns after cell lysis, as illustrated in **Figure 2**. IP columns are usually loaded with a pan MHC class I or II antibody to capture the MHC-peptide complexes. Next, peptides are eluted with acid and purified by HPLC from other molecules including Beta-2-Microglobulin and α -chains, as the last step before tandem mass spectrometry profiling⁵⁴.

Promising sources of antigens

Antigens are the main components of cancer vaccines, with the aim of eliciting an immune response while limiting their toxicity to healthy tissues. A vital step in their development is to identify and focus on relevant epitopes or antigens that are present only in cancer cells. Tumor antigens can arise from multiple sources⁵⁵, including (I) genomic variants such as SNVs, INDELs, gene fusions, and structural variants; (II) transcriptomic variants such as alternative splicing and non-coding regions; (III) and proteomic variants such as PTMs.

Genomic variants

Single nucleotide variants (SNVs) are non-synonymous point mutations that arise from various causes, including errors in DNA replication, exposure to exogenous or endogenous mutagens, ineffective DNA repair, and errors in DNA replication. SNVs, which are the most widespread genomic level mutations, have been extensively studied, as they are thought to be the most promising source of tumor-specific antigens. However, it has been reported that (I) only a small fraction of SNVs in tumor cells can yield antigenic peptides, (II) they are mostly patient-specific, and (III) their landscape is highly variable between cancer types and even cancer stages.

INDELS refers to the insertion and/or deletion of nucleotides at the genomic DNA level, and can induce translation in alternative frames. Neoantigens derived from INDELS are more common in malignancies with high microsatellite instability (*e.g.*, colorectal and gastric) owing to deficiencies in the DNA mismatch repair (MMR) processes⁵⁶. In addition, they are excellent candidates for microsatellite unstable cancers because of their recurrency⁵⁷. Gene fusions and structural variants are similar to INDELS but operate on much larger scales. However, they require more sophisticated pipelines and are less well studied.

Transcriptomic variants

Post-transcriptional events have the potential to expand the neoantigen space. The variety of tumor neoantigens is influenced by many messenger RNA processing mechanisms, such as alternative splicing events, RNA editing, and the translation of non-coding regions.

Alternative splicing

Neoantigens could originate from alternative splicing through mutations⁵⁸ at either cis-acting elements in the precursor mRNA or trans-acting alterations in a splicing factor. This leads to

formation of sequences with alternative 5' and 3' splice site determination, intron retention, exon skipping, and mutually exclusive exons⁵⁹.

Non-coding regions

Screening for neoantigens primarily resulting from mutations in exonic areas is restricted to 2% of the complete human genome because 91% of tumor mutations occur in the non-coding regions of genes. Many regions previously classified as non-coding have been found to have coding functions. The majority of these non-canonical events result from atypical translation events, rather than mutations. These aberrantly expressed antigens can be shared between tumor patients and are more common than neoantigens derived from the coding regions.

The translation of supposedly non-coding sequences or coding sequences into a non-canonical reading frame is an example of a non-canonical translation event. These typically involve non-canonical initiation, elongation, and termination events. In summary, a non-canonical initiation event occurs when the ribosome initiates translation at a codon other than the primary AUG codon, such as a non-primary AUG codon, or at a start codon that is close by (CUG, UUG, or GUG), as a result of a start codon scan-through⁶⁰, translation re-initiation, or the presence of an internal ribosome entry site (IRES) on the messenger RNA^{61,62}. When a frameshift occurs incidentally during elongation and results in the translation of a portion of the protein in a non-canonical reading frame, it is referred to as a non-canonical elongation event. It has already been noted that some slippage-prone regions found in transcripts can facilitate a process known as programmed ribosomal frameshift^{63,64}. Although uncommon, non-canonical termination events are possible and include a stop codon read-through⁶⁵ (certain stop codons, such as UGA and UAG, seem to be leakier than UAA) or a ribosomal frameshift at the stop codon. These non-canonical translation products are have been detected to be presented by the MHC system and found to illicit immune response in tumors^{37-42,66,67}.

Proteomic variants

Proteins can undergo significant co- and post-translational modifications (PTM) to control their activity. These PTMs are crucial for each phase of the protein lifetime and result in many possible proteoforms influencing the dynamic interactions of the proteome. MHC-associated peptides carrying PTMs, such as phosphorylation^{35,68}, citrullination, ubiquitination⁶⁹, and glycosylation⁷⁰, have been reported to alter antigen presentation and recognition. The ability of T lymphocytes to distinguish between modified and unmodified epitopes may be due to T cell escape from central tolerance in the thymus⁷¹. PTMs may also modify proteolytic activity and, in turn, affect how the MHC system presents peptides³⁶.

Thesis outline

One milestone in cancer vaccine development is the identification of effective tumor antigens that elicit tumor rejection in clinical settings. With this aim in mind, I implemented state-of-the-art mass spectrometry pipelines for the characterization of alternative sources of tumor antigens. In Chapter 2, I present the computational development of a mass spectrometry-based pipeline for characterizing non-canonical MHC-associated peptides (ncMAPs). This chapter introduces the importance of ncMAPs and the challenges hindering their large-scale assessment. Next, it describes the collected data from online available studies and, subsequently, the revealed characteristics of ncMAPs across cancer types along with their cancer selectivity. With the same aim in mind, I optimized a state-of-the-art mass spectrometry pipeline for the characterization of glycosylated MHC-associated peptides. In Chapter 3, I elaborate on the computational difficulties in analyzing glycosylated MHC-associated peptides and the bottlenecks in their large-scale assessment. Next, I describe the data collected from online-available studies and, subsequently, the characteristics of the revealed glycosylated peptides.

Chapters 2 and 3 serve the purpose of understanding different landscapes of MAPs across tissues and genomic origins. They shared a commonality of prerequiring few technical aims in the order below:

Key technical aims

Aim 1: Collect publicly available MS datasets and their metadata.

Aim 2: Develop large-scale computational MS pipelines to enable the interrogation of many immunopeptidomic datasets.

Aim 3: Benchmark the pipelines against publicly available datasets for quality control.

Aim 4: Perform data exploration and visualization of the MS search results.

Key biological aims

With the technical aims achieved, this thesis fills a series of gaps in our knowledge of the immunopeptidome outlined below:

1. **gap 1:** Establishing the landscape of non-canonical MHC class I-associated peptides (ncMAPs) in a plethora of cancer types and assessing their tumor selectivity.
 - I. What are the sources of ncMAPs in terms of their biological mechanisms and genomic origin?
 - II. Are ncMAPs preferentially presented by specific HLA types or subtypes?
 - III. Are ncMAPs common or shared among different cancer types?
 - IV. To what degree are ncMAPs present in healthy tissues versus in tumors?
2. **gap 2:** Establishing the landscape of post-translationally modified MHC class I-associated peptides (ptmMAPs) in various cancer types.
 - I. To what degree is the immunopeptidome post-translational modified?
 - II. Can PTMs be potential targets for cancer treatment?

3. **gap 3:** Establishing the landscape of glycosylated MAPs.
 - I. What are the most common glycan types on MAPs?
 - II. Where are glycans located relative to the MHC peptide-binding pocket?
 - III. Are glycosylated MAPs preferentially presented by specific HLA types or subtypes?

Chapter 4 concludes this thesis with a summary of the findings, discussion around the unmet needs in the field and my vision of the research course over the next decade.

References

1. Nesmiyanov, P. P. Antigen Presentation and Major Histocompatibility Complex. in *Encyclopedia of Infection and Immunity* (ed. Rezaei, N.) 90–98 (Elsevier, 2022). doi:10.1016/B978-0-12-818731-9.00029-X.
2. Jhunjhunwala, S., Hammer, C. & Delamarre, L. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. *Nat. Rev. Cancer* **21**, 298–312 (2021).
3. O’Callaghan, C. A. *et al.* Structural Features Impose Tight Peptide Binding Specificity in the Nonclassical MHC Molecule HLA-E. *Mol. Cell* **1**, 531–541 (1998).
4. Carosella, E. D., Favier, B., Rouas-Freiss, N., Moreau, P. & LeMaout, J. Beyond the increasing complexity of the immunomodulatory HLA-G molecule. *Blood* **111**, 4862–4870 (2008).
5. Roche, P. A. & Furuta, K. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat. Rev. Immunol.* **15**, 203–216 (2015).
6. Seliger, B., Kloor, M. & Ferrone, S. HLA class II antigen-processing pathway in tumors: Molecular defects and clinical relevance. *OncolImmunology* **6**, e1171447 (2017).

7. Drozina, G., Kohoutek, J., Jabrane-Ferrat, N. & Peterlin, B. M. Expression of MHC II genes. *Curr. Top. Microbiol. Immunol.* **290**, 147–170 (2005).
8. Natali, P. G. *et al.* Analysis of the antigenic profile of uveal melanoma lesions with anti-cutaneous melanoma-associated antigen and anti-HLA monoclonal antibodies. *Cancer Res.* **49**, 1269–1274 (1989).
9. Real, F. X. *et al.* Surface antigens of melanomas and melanocytes defined by mouse monoclonal antibodies: specificity analysis and comparison of antigen expression in cultured cells and tissues. *Cancer Res.* **45**, 4401–4411 (1985).
10. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
11. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
12. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *cell* **100**, 57–70 (2000).
13. Bidram, E. *et al.* A concise review on cancer treatment methods and delivery systems. *J. Drug Deliv. Sci. Technol.* **54**, 101350 (2019).
14. Robert, C. A decade of immune-checkpoint inhibitors in cancer therapy. *Nat. Commun.* **11**, 3801 (2020).
15. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
16. Park, Y.-J., Kuen, D.-S. & Chung, Y. Future prospects of immune checkpoint blockade in cancer: from response prediction to overcoming resistance. *Exp. Mol. Med.* **50**, 1–13 (2018).

17. Allison, J. & Honjo, T. Cancer immunologists scoop medicine Nobel prize. <https://www.nature.com/articles/d41586-018-06751-0>.
18. Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of Antitumor Immunity by CTLA-4 Blockade. *Science* **271**, 1734–1736 (1996).
19. Ishida, Y., Agata, Y., Shibahara, K. & Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO J.* **11**, 3887–3895 (1992).
20. Hodi, F. S. *et al.* Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *N. Engl. J. Med.* **363**, 711–723 (2010).
21. Wang, X. *et al.* Effectiveness and safety of PD-1/PD-L1 inhibitors in the treatment of solid tumors: a systematic review and meta-analysis. *Oncotarget* **8**, 59901–59914 (2017).
22. Kleponis, J., Skelton, R. & Zheng, L. Fueling the engine and releasing the break: combinational therapy of cancer vaccines and immune checkpoint inhibitors. *Cancer Biol. Med.* **12**, 201–208 (2015).
23. Grenier, J. M., Yeung, S. T. & Khanna, K. M. Combination Immunotherapy: Taking Cancer Vaccines to the Next Level. *Front. Immunol.* **9**, 610–610 (2018).
24. Richters, M. M. *et al.* Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med.* **11**, 56 (2019).
25. Hundal, J. *et al.* pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol. Res.* **8**, 409–420 (2020).
26. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).

27. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
28. 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 Res.* **48**, W449–W454 (2020).
29. O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* **11**, 42–48 (2020).
30. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
31. Diao, K. *et al.* Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction. *Int. J. Mol. Sci.* **23**, 11624 (2022).
32. Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Res.* gkz950 (2019) doi:10.1093/nar/gkz950.
33. Ebrahimi-Nik, H. *et al.* Mass spectrometry–driven exploration reveals nuances of neoepitope-driven tumor rejection. *JCI Insight* **4**, e129152 (2019).
34. Yi, X. *et al.* caAtlas: An immunopeptidome atlas of human cancer. *iScience* **24**, 103107 (2021).
35. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 1–16 (2016).

36. Kacen, A. *et al.* Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01464-2.
37. Erhard, F., Dölken, L., Schilling, B. & Schlosser, A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunol. Res.* **8**, 1018–1026 (2020).
38. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
39. Smart, A. C. *et al.* Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.* **36**, 1056–1058 (2018).
40. Ruiz Cuevas, M. V. *et al.* Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815 (2021).
41. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
42. Ouspenskaia, T. *et al.* Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* **40**, 209–217 (2022).
43. Scigelova, M. & Makarov, A. Orbitrap Mass Analyzer – Overview and Applications in Proteomics. *PROTEOMICS* **6**, 16–21 (2006).
44. Haag, A. M. Mass Analyzers and Mass Spectrometers. in *Modern Proteomics – Sample Preparation, Analysis and Practical Applications* (eds. Mirzaei, H. & Carrasco, M.) 157–169 (Springer International Publishing, 2016). doi:10.1007/978-3-319-41448-5_7.
45. Chen, C., Hou, J., Tanner, J. J. & Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int. J. Mol. Sci.* **21**, 2873 (2020).

46. Zhang, J. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111-010587 (2012).
47. Qiao, R. *et al.* DeepNovoV2: Better de novo peptide sequencing with deep learning. (2019).
48. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci.* **114**, 8247 (2017).
49. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
50. Kong, A. T., Lempvost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513 (2017).
51. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
52. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
53. Kote, S., Pirog, A., Bedran, G., Alfaro, J. & Dapic, I. Mass Spectrometry-Based Identification of MHC-Associated Peptides. *Cancers* **12**, 535 (2020).
54. Purcell, A. W., Ramarathinam, S. H. & Ternette, N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat. Protoc.* **14**, 1687–1707 (2019).
55. Xie, N. *et al.* Neoantigens: promising targets for cancer therapy. *Signal Transduct. Target. Ther.* **8**, 9 (2023).

56. Roudko, V. *et al.* Lynch Syndrome and MSI-H Cancers: From Mechanisms to ‘Off-The-Shelf’ Cancer Vaccines. *Front. Immunol.* **12**, 757804 (2021).
57. Ballhausen, A. *et al.* The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution. *Nat. Commun.* **11**, 4740 (2020).
58. Rivero-Hinojosa, S. *et al.* Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat. Commun.* **12**, 6689 (2021).
59. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Signal Transduct. Target. Ther.* **6**, 78 (2021).
60. Bullock, T. N. & Eisenlohr, L. C. Ribosomal scanning past the primary initiation codon as a mechanism for expression of CTL epitopes encoded in alternative reading frames. *J. Exp. Med.* **184**, 1319–1329 (1996).
61. Jang, S. K. *et al.* A segment of the 5’ nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J. Virol.* **62**, 2636–2643 (1988).
62. Pelletier, J. & Sonenberg, N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334**, 320–325 (1988).
63. Zook, M. B., Howard, M. T., Sinnathamby, G., Atkins, J. F. & Eisenlohr, L. C. Epitopes Derived by Incidental Translational Frameshifting Give Rise to a Protective CTL Response. *J. Immunol.* **176**, 6928–6934 (2006).
64. Saulquin, X. *et al.* +1 Frameshifting as a Novel Mechanism to Generate a Cryptic Cytotoxic T Lymphocyte Epitope Derived from Human Interleukin 10. *J. Exp. Med.* **195**, 353–358 (2002).

65. Goodenough, E. *et al.* Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR. *Proc. Natl. Acad. Sci.* **111**, 5670–5675 (2014).
66. Laumont, C. M. *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238 (2016).
67. Attig, J. *et al.* LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. *Genome Res.* **29**, 1578–1590 (2019).
68. Penny, S. A. *et al.* Tumor Infiltrating Lymphocytes Target HLA-I Phosphopeptides Derived From Cancer Signaling in Colorectal Cancer. *Front. Immunol.* **12**, 723566 (2021).
69. Gavali, S., Liu, J., Li, X. & Paolino, M. Ubiquitination in T-Cell Activation and Checkpoint Inhibition: New Avenues for Targeted Cancer Immunotherapy. *Int. J. Mol. Sci.* **22**, 10800 (2021).
70. Malaker, S. A. *et al.* Identification and Characterization of Complex Glycosylated Peptides Presented by the MHC Class II Processing Pathway in Melanoma. *J. Proteome Res.* **16**, 228–237 (2017).
71. Raposo, B. *et al.* T cells specific for post-translational modifications escape intrathymic tolerance induction. *Nat. Commun.* **9**, 353 (2018).

The immunopeptidome from a genomic perspective: Establishing the non-canonical landscape of MHC class I-associated peptides.

Georges Bedran¹, Hans-Christof Gasser², Kenneth Weke¹, Tongjie Wang², Dominika Bedran¹, Alexander Laird^{3,4}, Christophe Battail⁵, Fabio Massimo Zanzotto⁶, Catia Pesquita⁷, Håkan Axelsson⁸, Ajitha Rajan², David J. Harrison⁹, Aleksander Palkowski¹, Maciej Pawlik¹⁰, Maciej Parys¹¹, Robert O'Neill¹², Paul M. Brennan¹³, Stefan N. Symeonides⁴, David R. Goodlett^{1,14,15}, Kevin Litchfield^{16,17}, Robin Fahraeus^{1,18}, Ted R. Hupp^{1,4}, Sachin Kote^{1,*}, Javier A. Alfaro^{1,2,14,*}

1 International Centre for Cancer Vaccine Science, University of Gdansk, Gdansk, Poland

2 School of Informatics, University of Edinburgh, Edinburgh, UK

3 Urology Department, Western General Hospital, NHS Lothian, Edinburgh, UK

4 Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

5 CEA, Grenoble Alpes University, INSERM, IRIG, Biosciences and bioengineering for health laboratory (BGE) - UA13 INSERM-CEA-UGA, Grenoble, France

6 Department of Enterprise Engineering, University of Rome "Tor Vergata", Rome, Italy

7 LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

8 Division of Translational Cancer Research, Department of Laboratory Medicine, Lund University, Lund, Sweden

9 School of Medicine, University of St Andrews, St Andrews, UK

10 Academic Computer Centre CYFRONET, AGH University of Science and Technology, Cracow, Poland

11 Royal (Dick) School of Veterinary Studies and The Roslin Institute, University of Edinburgh, Edinburgh, UK

12 Cambridge Oesophagogastric Centre, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

13 Translational Neurosurgery, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

14 Department of Biochemistry and Microbiology, University of Victoria, Victoria, Canada

15 University of Victoria Genome BC Proteome Centre, Victoria, Canada

16 Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK

17 Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK

18 Inserm UMRS1131, Institut de Génétique Moléculaire, Université Paris 7, Paris, France

* **Correspondence** should be addressed to Javier A. Alfaro; mailing address: ul. Kładki 24 80-822 Gdańsk, Poland; e-mail address: javier.alfaro@proteogenomics.ca and Sachin Kote; mailing address: ul. Kładki 24 80-822 Gdańsk, Poland; e-mail address: sachin.kote@ug.edu.pl

Running title: The MHC class I non-canonical landscape.

Conflicts of interest statement: The authors declare no potential conflicts of interest.

Keywords: Cancer, tumor antigens, non-canonical MHC class I-associated peptides, mass spectrometry, shared antigens.

Financial support: G.B., D.B., K.W., A.P., R.F., T.R.H., S.K., and J.A.A. received support from Fundacja na rzecz Nauki Polskiej (FNP) (grant ID: MAB/3/2017). D.R.G. received support from Genome Canada & Genome BC (grant ID: 264PRO). D.J.H. received support from NuCana plc (grant ID: SMD0-ZIUN05). H.A. received support from Swedish Cancer Foundation (grant ID: 211709). H.G. received support from United Kingdom Research and Innovation (UKRI) (grant ID: EP/S02431X/1). C.P. received support from Fundação para a Ciência e a Tecnologia (FCT) through LASIGE Research Unit (grant ID: UIDB/00408/2020 and UIDP/00408/2020). A.L. F.M.Z., C.P., A.R., A.P., and J.A.A. received support from European Union's Horizon 2020 research and innovation programme (grant ID: 101017453). C.B. received support from Agence Nationale de la Recherche (ANR) through GRAL LabEX (grant ID: ANR-10-LABX-49-01) and CBH-EUR-GS 32 (grant ID: ANR-17-EURE-0003). S.N.S. received support from Cancer Research UK (CRUK) and the Chief Scientist's Office of Scotland (CSO): Experimental Cancer Medicine Centre (ECMC) (grant ID: ECMCQQR-2022/100017). A.L. received support from Chief Scientist's Office of Scotland (CSO) NRS Career Researcher Fellowship. R.O.N. received support from CRUK Cambridge Centre Thoracic Cancer Programme (grant ID: CTRQQR-2021\100012).

Synopsis: Identification of tumor-specific antigens is crucial for developing effective cancer treatments. The authors use MS *de novo* and proteogenomics to generate an atlas of non-canonical MHC class I-associated peptides, providing potential targets for cancer T-cell therapies or vaccines.

Abstract

Tumor antigens can emerge through multiple mechanisms, including translation of non-coding genomic regions. This non-canonical category of tumor antigens has recently gained attention; however, our understanding of how they recur within and between cancer types is still in its infancy. Therefore, we developed a proteogenomic pipeline based on deep learning *de novo* mass spectrometry to enable the discovery of non-canonical MHC class I-associated peptides (ncMAPs) from non-coding regions. Considering that the emergence of tumor antigens can also involve post-translational modifications, we included an open search component in our pipeline. Leveraging the wealth of mass spectrometry-based immunopeptidomics, we analyzed data from 26 MHC class I immunopeptidomic studies across 11 different cancer types. We validated the *de novo* identified ncMAPs, along with the most abundant post-translational modifications, using spectral matching and controlled their false discovery rate (FDR) to 1%. The non-canonical presentation appeared to be 5 times enriched for the A03 HLA supertype, with a projected population coverage of 54.85%. The data reveal an atlas of 8,601 ncMAPs with varying levels of cancer selectivity and suggest 17 cancer-selective ncMAPs as attractive therapeutic targets according to a stringent cutoff. In summary, the combination of the open-source pipeline and the atlas of ncMAPs reported herein could facilitate the identification and screening of ncMAPs as targets for T-cell therapies or vaccine development.

Introduction

The accelerated adoption of mass spectrometry (MS) for high-throughput profiling of immunopeptidomes in cancer has led to several discoveries. Leveraging these studies to improve cancer immunotherapy involves connecting the wealth of immunopeptidomic data to immunogenomics, where the goal is to carefully choose effective targets for T-cell therapies or vaccine development.

The discovery of cancer antigens has mainly focused on mutated tumor-specific antigens (neoantigens) arising from patient-specific somatic mutations. It has been shown that only a small percentage of the numerous non-synonymous mutations in a tumor actually produce neoantigens (1,2). The challenging task of identifying those that can evoke a suitable tumor rejection was addressed by Ebrahimi-Nik et al. (3). Using a combination of genomics, shotgun MS immunopeptidomics, and targeted MS, they found that (I) MS-identified neopeptides are a rich source of tumor rejection–mediating antigens, (II) neoantigens derive from passenger mutations, and (III) binding affinity and CD8⁺ T-cell responses in tumor-bearing hosts are poor predictors of antitumor activity *in vivo*. Although neoantigens confer an advantage to patients undergoing immunotherapy (4), their patient-specific nature is a major bottleneck when producing off-the-shelf treatments for a large number of individuals. Alternatively, shared neoantigens (5) (*i.e.*, recurrent mutations in cancer) could offer a new line of population-level immunotherapy. However, high-throughput experimental profiling of such broadly presented neoantigens across the human population is a long-term goal with many milestones to be achieved.

Recently, tumor antigens that exceed the exome boundaries (*i.e.*, non-canonical) have attracted attention as potential targets as a result of their immunogenicity and recurrence among cancer patients (6). These antigens find their way to the cell surface through rapid degradation (7) of “non-coding” translation products stemming from novel open reading frames

(nORF) (8). In addition, “non-coding” translation products can originate from other sources (9), including intron retention (IR) (10), ribosomal slippage (11), and frameshift mutations (12). In 2016, Laumont *et al.* (13) demonstrated their association with MHC molecules using a reductionist approach based on 6-frame translation and subsequently their recurrence between patients (14). Ribo-Seq has proven to be an immensely valuable tool for identifying non-canonical MHC class I-associated peptides (ncMAPs) as it provides experimental evidence for their non-canonical translation and MHC class I presentation when combined with MS immunopeptidomics (6,15,16). Despite previous efforts to study non-canonical immunopeptidomes, the requirements of such multi-level experimental data (Ribo-seq and/or RNA-Seq) or computational struggles when dealing with large MS databases have hindered their large-scale profiling in a harmonized manner across multiple cancer types from hundreds of samples.

With these considerations in mind, we developed COD-dipp (Closed Open *De novo* – deep immunopeptidomics pipeline), a pipeline based on deep learning *de novo* MS to enable the discovery of ncMAPs. Owing to the potential involvement of post-translational modifications (PTMs) in this process (1), we added an open search component for their discovery. We applied COD-dipp to a large-scale dataset using immunopeptidome profiles of over 772 samples from 26 (1,13,14,17–40) published studies and 11 cancer types. We identified a range of PTMs of potential interest from a therapeutic standpoint and tackled the non-canonical immunopeptidome. We validated the *de novo* identified ncMAPs and controlled their false discovery rate (FDR) to 1% using a second-round search with tuned PTM parameters, in addition to a series of quality-control steps. Our large-scale analysis revealed 8,601 ncMAPs, accounting for 1.7% of immunopeptidomes. These peptides had varying levels of tumor selectivity, defined by their parent gene expression levels in normal tissues. We suggest 17 ncMAPs as attractive therapeutic targets using a stringent tumor-selectivity cutoff.

Materials and methods

Dataset selection

Twenty-four studies were selected based on a list of keywords related to immunopeptidomics (**Supplementary Method S1**). Low-resolution analyses were eliminated, and MHC class I-related datasets conducted with at least one of the following instruments were kept: Q Exactive, Q Exactive plus/HF/HFX, LTQ Orbitrap Velos, LTQ Orbitrap Elite, Orbitrap Fusion, and Orbitrap Fusion Lumos (**Supplementary Table S1**). An additional study was considered from the MassIVE (RRID:SCR_013665) database, as it incorporates 95 HLA-A, -B, -C, and -G mono-allelic cell lines (28,40). An auxiliary immunopeptidomic dataset (39) covering 30 healthy tissues from 21 healthy individuals was also used to partly assess cancer selectivity.

Proteogenomic database generation

Canonical protein database for MS database search

A protein database was downloaded using ENSEMBL r94 BioMart (RRID:SCR_002344); decoy sequences were appended by reversing the target sequences, and 116 contaminant proteins were added (41).

Non-canonical protein database for alignment using BLAST-like alignment tool (BLAT)

A pre-mRNA 3-frame translation (3FT) database was generated from genes with a protein-coding biotype based on ENSEMBL r94 (RRID:SCR_002344) using the AnnotationHub and Biostrings (RRID:SCR_016949) R packages.

COSMIC mutated protein database for BLAT alignment

COSMIC (RRID:SCR_002260) coding Mutants (42) VCF v95 was downloaded along with ENSEMBL v94 CDS and GTF files. An in-house Python (RRID:SCR_008394) package was used along with the previously mentioned inputs to generate a FASTA file containing the corresponding mutated protein sequences.

MS computational analysis

Algorithms representing three main philosophies of peptide-spectrum matching including open search, *de novo* sequencing, and closed search were used. The open search approach allowed the identification of distantly related peptides and could identify PTMs and single amino acid variations. The *de novo* sequencing approach derived sequences from first-principle analysis of the MS² spectra. The closed search approach, used as a validation step, assumed a specific set of reference protein sequences and allowed for limited post-translational modifications. Although each approach has its own limitations, our strategy addressed them by combining a closed search with a *de novo* sequencing approach and implementing multiple filtering steps for accuracy control and quality control checkpoints (see **Supplementary Figure S1**).

Data conversion

The proprietary RAW files acquired from the selected instruments were converted to mzML and MGF formats using `msconvert` (ProteoWizard version 3.0.19295. c8b8b470d, RRID:SCR_012056) with the peak-picking and TPP compatibility filters.

Open search analysis

The MSFragger (43) v2.2 search engine was used to conduct an open search analysis against the ENSEMBL r94 protein database in combination with PTMiner (44) v1.1.2, to apply a transfer FDR and a false localization rate of 1% (FLR, the rate of falsely localizing the site of

modification). Unspecific cleavage generating peptides 8 to 25 amino acids long with no fixed/variable PTMs was considered. Further analysis revealed that the frequent unexplained mass shifts observed during the open-search annotations were caused by non-specific cleavage. To address this issue, an open-search post-processing algorithm, PTMiner, was employed to effectively corrects for mass shifts introduced by in-source fragmentation, nonspecific digestion, or missed cleavage, by adding or deleting amino acids from the peptide N- or C-termini. For instance, a deviation of -128.1 to -128.08 Dalton on lysine residues was frequently detected on the first 2 or last 2 amino acids of peptides. The deviation was caused by non-specific cleavage during the open search and resulted in an incorrect assignment of a negative mass shift of a lysine due to the presence of an additional lysine in the sequence. As these cases are not biologically meaningful, unexplained mass shifts were removed from the final results of the study.

De novo analysis

DeepNovoV2 (45) is a neural-network-based *de novo* peptide sequencing model that integrates convolutional neural networks (CNNs) and long short-term memory (LSTM) architectures. This deep-learning design extracts features from both the spectrum and the language of the presented peptides. DeepNovo has demonstrated improved performance compared to the state-of-the-art *de novo* sequencing algorithms by large margins (45). The model can be tuned on a restricted peptide space to improve its performance. The training, testing, and validation sets were derived from MS-GF+ (v2019.04.18, RRID:SCR_015646) database search results for each sample. The search used the ENSEMBL v94 protein database and 8 to 25 amino acid peptides with unspecific cleavage, no fixed/variable PTMs and an FDR of 1% applied by Scavenger (46). The trained models were used to perform *de novo* (prediction) on the remaining unmatched spectra of each sample (from MS-GF+ after 1% FDR control). Accuracy was calculated by comparing the *de novo* predicted sequences and MS-GF+ results on the validation set. A *de novo* score threshold that controlled the

accuracy at 90% within the validation set was applied to the predicted sequence in a sample-specific manner.

***De novo* peptide annotation**

De novo peptides from canonical human proteins were identified using BLAT (47) (RRID:SCR_011919) alignment against the target-decoy protein database. Sequences perfectly matching any protein sequence were considered exonic (one mismatch allowed for the isobaric amino acids leucine and isoleucine). All remaining sequences unexplained by proteins were considered potential non-canonical peptides and were aligned against the pre-mRNA 3FT database. Stringently, peptides perfectly matching a 3FT sequence without any mismatch were required to have at least three mismatches with any known protein sequence before being considered non-canonical. Since peptide-spectrum matches (PSMs) can be assigned without complete sequencing accuracy, requiring a 3 amino acid difference alongside the 90% accuracy cutoff above increases the confidence that the peptides assigned fall far outside the standard human proteome. Remaining *de novo* peptides without any canonical or non-canonical annotation were labeled as 'unmapped peptides' and discarded.

Second-round search

A second-round search was performed using the FragPipe (41,43) headless pipeline, which includes MSFragger v3.4, MSBooster (bioRxiv 2022.10.19.512904), and Philosopher (41). Non-canonical peptides from all samples were concatenated with the ENSEMBL v94 protein into a custom database. Only four of the most abundant PTMs were considered to avoid a large search space complexity, inflated FDR, and decreased sensitivity. The following variable PTMs were included: methionine oxidation, N-terminal acetylation, cysteinylolation, and cysteine carbamidomethylation (for samples treated with iodoacetamide). Unspecific cleavage generating peptides 7 to 15 amino acids long was considered. The ion, PSM, and peptide-level FDR were maintained at 1%.

Alignment of immunopeptides to the genome

Second-round search non-canonical peptide coordinates were retrieved from the 3FT database FASTA headers and stored in BED format.

Open reading frame analysis

Upstream genomic sequences of ncMAPs were scanned for start codons up to the first encounter with a stop codon. Sequences were centered around the detected start codons and stretches of 100 nucleotides from each side were extracted. Translation initiation site (TIS) scores were predicted for each sequence using TITER (48), a deep-learning-based framework for accurately predicting TIS on a genome-wide scale based on QTI-seq data. A TIS score greater than 0.5 was considered a positive prediction.

Intron retention analysis

For each intron in the UCSC hg38 KnownGene table (RRID:SCR_005780), the first codon coordinates of the corresponding upstream exon in-frame with the canonical translation were extracted and stored in BED format (see **Pseudocode 1**). Intronic coordinates from the generated BED file were intersected with the ncMAPs BED file using pybedtools (49) (RRID:SCR_021018). Intronic retention events were considered possible when ncMAPs within introns were in-frame with their upstream exons (see **Pseudocode 2**).

```

// Pseudo-code 1: extracts the start coordinate of the first in-frame codon for
each exon (inframeCoordinate variable)
for each transcript
  remainderValue = 0
  for each exon
    if strand is positive
      if downstream intron exists
        leftoverBases = remainder of (ExonEndCoordinate - remainderValue - ExonStart +
1) / 3
        if remainderValue is equal to 0
          inframeCoordinate = ExonStartCoordinate
        else
          inframeCoordinate = ExonStartCoordinate - remainderValue
        if leftoverBases is greater than 0
          remainderValue = 3 - leftoverBases
        addToTable(transcript, chromosome, ExonStart, ExonEnd, inframeCoordinate,
IntronStart, IntronEnd)
    if strand is negative
      if downstream intron exists
        leftoverBases = remainder of (ExonStart - ExonEndCoordinate +
remainderValue + 1) / 3
        if remainderValue is equal to 0
          inframeCoordinate = ExonEndCoordinate
        else
          inframeCoordinate = ExonEndCoordinate + remainderValue
        if leftoverBases is greater than 0
          remainderValue = 3 - leftoverBases
        addToTable(transcript, chromosome, ExonStart, ExonEnd, inframeCoordinate,
IntronStart, IntronEnd)

// Pseudo-code 2: checks if each intronic ncMAP is in-frame with its upstream
exon.
ncMAPIsInFrame = False
if strand is positive
  // firstCoordinate = start coordinate of ncMAP
  // secondCoordinate = start coordinate of the first inframe codon from
previous exon
  coordinateDifference = firstCoordinate - secondCoordinate
  if remainder of (coordinateDifference / 3) is equal to 0
    ncMAPIsInFrame = True
else:
  // firstCoordinate = start coordinate of first inframe codon from previous
exon
  // secondCoordinate = end coordinate of ncMAP
  coordinateDifference = firstCoordinate - secondCoordinate
  if remainder of (coordinateDifference / 3) is equal to 0
    ncMAPIsInFrame = True

```

Frameshift mutation analysis

The COSMIC (42) v95 coding mutations (RRID:SCR_002260) VCF file was downloaded and converted into a protein FASTA file using aVCF-to-Proteogenomics toolkit (<https://github.com/immuno-informatics/VCFtoProteogenomics>) ncMAPs were then aligned to

the resulting 16 GB FASTA using BLAT v35 (47). Only hits with exact matches to sequences from frameshift mutations were considered.

Comparison of the identified non-canonical MHC class I–associated peptides between studies

ncMAPs from 4 different studies (6,13,16,50) were collected. First, sequences were aligned to the human proteome (ENSEMBL v94) using BLAT v35 (47). Sequences found in human proteins were discarded, and the remaining sequences were aligned to the 3FT database with one mismatch allowance for the isobaric amino acids leucine and isoleucine, as allowed for COD-dipp ncMAPs. Genomic coordinates of the sequences found in the 3FT database were extracted and overlapped between studies using the ChIPpeakAnno (51) R package (RRID:SCR_012828). A minimum overlap of 21 nucleotides (7 amino acids) between two sequences was required.

Cancer selectivity of the non-canonical MHC class I–associated peptides

Tumor specificity has been previously implied when peptide parent genes are either completely absent or present in trace amounts in healthy tissues (6,14,16) since MHC class I presentation is preferentially derived from highly abundant transcripts (28,30). While tumor specificity implies the expression of an antigen solely in tumor samples, the experimental design of this study cannot guarantee this constraint. Instead, cancer-selective ncMAPs were conservatively identified through three iterative steps:

Step 1: Panel of normal immunopeptidomes

In addition to the 88 healthy MS samples from the initial set of the 25 considered studies, the HLA Ligand Atlas (39) was used to extend the panel of normal immunopeptidomes and partly

assess the cancer selectivity of the 8,601 identified ncMAPs. The HLA Ligand Atlas is a pan-tissue immunopeptidomic reference for 30 healthy tissue types obtained from 21 human subjects. The resulting 334 healthy samples (see **Supplementary Table S1**) were analyzed in the same manner as in the second-round search (see *Second-round search* above). ncMAPs identified in the panel of normal immunopeptidomes were labeled as non-cancer selective.

Dimensionality reduction of the HLA-binding motif space

Binding affinity prediction was employed to identify similarities and differences in HLA-binding motifs among the 65 healthy and 51 tumor-only HLA alleles. NetMHCpan-4.1 was utilized to evaluate the binding of 1,000,000 random peptides to each allele, which resulted in a binding matrix (BM) of 116 alleles and 1,000,000 peptides. A value of 1 was assigned to strong binders (EL rank $\leq 0.5\%$) in the BM; otherwise, a value of 0 was assigned. A pairwise cosine distance matrix (DM) was then calculated to assess the similarity of binding between alleles. The DM was then reduced using t-SNE to visualize the data in 2D with a perplexity of 20 and 500 iterations.

Step 2: Parental gene expression levels in healthy tissue

The gene expression levels of the identified ncMAPs were retrieved from the GTEx v8 (52) dataset, consisting of 29 tissues from 948 healthy donors and 17,382 overall samples. Considering all individuals, the 90th percentile value of normalized expression was assigned to each gene per tissue as a strict step to guarantee the upper-end gene expression in healthy tissues. A stringent cutoff for cancer selectivity was used to shortlist ncMAPs whose parent genes fell below a 1 TPM expression cutoff (excluding the testis tissue given its immune-privileged status). It is worth noting that this stringent threshold removes 92% of protein-coding genes that show expression above 1 TPM in any tissue within the GTEx v8.

Step 3: Protein expression levels in healthy tissue

The protein expression levels of ncMAPs passing the 1 TPM cutoff were retrieved from the Human Protein Atlas V22.0 database (53). ncMAPs without parent protein expression in healthy tissues were labeled as cancer-selective (excluding the testis tissue given its immune-privileged status).

Code availability

The COD-dipp code, intended for high-performance computing (HPC), is available on the GitHub repository: <https://github.com/immuno-informatics/COD-dipp>.

Data availability

The data analyzed in this study were obtained from [PRIDE](#) at PXD004746, PXD014017, PXD012308, PXD011628, PXD012083, PXD011766, PXD013057, PXD011723, PXD007203, PXD004233, PXD003790, PXD001898, PXD007860, PXD011257, PXD007935, PXD009749, PXD009753, PXD009750, PXD009751, PXD009752, PXD009754, PXD009755, PXD004023, PXD007596, PXD009531, PXD010808, PXD008937, PXD009738, PXD006939, PXD005231, PXD000394, PXD004894, PXD019643 and from [massIVE](#) at MSV000080527, MSV000084172, MSV000084442. The results of this study are available within the article and its supplementary data files and are accessible on the following figshare repository: <https://doi.org/10.6084/m9.figshare.16538097>.

Results

Immunopeptidomic MS datasets

We selected 25 immunopeptidomic MS studies (see **Supplementary Table S1**) to create a cancer-centered dataset of MHC class I presentation. Data-dependent acquisition (DDA)

studies covered eleven cancer types distributed across the brain (Glioblastoma and Meningioma), lung, skin, liver, blood (Leukemia and Lymphoma), colon, ovaries, kidneys, and breast. Moreover, tumor and healthy samples were derived from either cell lines or patient tissues (**Fig. 1a** and **Supplementary Method S1**). We selected publicly available studies with data generated using high-resolution MS instruments (LTQ Orbitrap, Q Exactive Plus/HF/HFX, and Fusion Lumos) to minimize the bias associated with older tandem MS instrumentation (**Fig. 1b**). Within our dataset, the most commonly used monoclonal antibody for HLA class I immunoprecipitation (IP) was W6/32 in comparison to the other antibodies (BB7.2 and G46-2.6) (**Fig. 1c**, see **Supplementary Table S1**). The selected studies covered five different HLA class I genes, with HLA-A, B, and C being the most studied compared to HLA-E and -G (**Fig. 1d**). Furthermore, the included MS samples covered 114 HLA alleles (**Fig. 1e**).

Closed Open *De novo* – deep immunopeptidomics pipeline (COD-dipp)

We present COD-dipp, an open-source high-throughput pipeline with novel post-processing steps, to deeply interrogate immunopeptidomic datasets (**Fig. 2**). We used this pipeline to screen for ncMAPs in datasets utilizing DDA due to its widespread use. To identify post-translationally modified MHC class I-associated peptides (ptmMAPs), we performed an open-search analysis with MSFragger (43) and controlled both FDR and the FLR to 1% with PTMiner (44). To identify ncMAPs, we used DeepNovoV2 (45) for *de novo* analysis. In combination with the PSM level information of MS-GF+ (54), DeepNovoV2 was trained to interpret the raw MS data in a sample-specific manner. The training step for such a deep learning approach is crucial for learning the features of tandem mass spectra, fragment ions, and leveraging sequence patterns in the immunopeptidome to impute missing MS² fragments. All high-quality *de novo* peptides (90% accuracy) were sequentially mapped (47) to (I) the human reference proteome to reveal the *de novo*-based canonical MHC class I-associated peptides, and (II) to a 3FT database to reveal the *de novo*-based ncMAPs. Finally, an

orthogonal validation step was performed by a second-round search to control a 1% FDR for the *de novo* identified ncMAPs while considering the most abundant PTMs found by the open-search strategy. Applying the COD-dipp pipeline across the dataset revealed the breadth of (I) post-translationally modified MHC class I-associated peptides referred to as ptmMAPs, and (II) non-canonical MHC class I-associated peptides referred to as ncMAPs.

ptmMAPs

The open search analysis reported 4.03% of the MS spectra showing post-translational modifications (**Fig. 3a**). Some identified PTMs were confirmatory, representing chemical modifications from sample preparation methods (cysteine carbamidomethylation) or common chemical derivatives (methionine oxidation and di-oxidation). We also observed PTMs that are extremely common in proteins, such as protein N-terminal acetylation, affecting multiple properties such as half-life time, folding, and interaction. On the other hand, some of the identified PTMs have been reported previously to increase immunogenicity against diseases (55) and protect against degradation (tri-oxidation of cysteine (56), cysteinylolation (57), and N-term serine acetylation, see **Fig. 3b** and **Supplementary Table S2**). Furthermore, 1.12% of spectra from open search showed unknown mass shifts, as illustrated in **Fig. 3a** (green and red). This category was partly populated by computational artifacts and was excluded from the final results. To validate these findings, we performed an independent post-search by crosschecking the identifications from our open search with those of the original studies. The results showed 96.1% agreement in peptide-spectrum matches, which are detailed in **Supplementary Method S2: Validation 1** and **Supplementary Figure S2**.

ncMAPs

We explored the ncMAP landscape in cancer using our workflow (**Fig. 2**) and identified 10,413 unique *de novo*-based ncMAPs from intragenic non-coding regions (before the second-round search validation), which accounted for 3.7% of the identified *de novo* sequences. We took

two additional validation steps, including checking the identification scores as well as the correlation between the experimental and theoretical liquid chromatography retention times, to guarantee the correctness of these identifications (see **Supplementary Method S2: Validation 2 and 3**, and **Supplementary Figure S2**). The *de novo* non-canonical peptides showed strong evidence of high-quality identification (*i.e.*, correctly predicted complete peptide sequences). Even with this strong evidence, it was possible that chromatic behavior remained unchanged in certain instances where neighboring amino acids were in flipped positions, or that a 90% accuracy rate still led to an uncertain FDR percentage. Hence, we confirmed the identified 10,413 *de novo*-based ncMAPs by performing a second-round search for additional validation and controlling the FDR at 1%. Several PTMs were also considered in the parameters from the *a priori* knowledge provided by the open search strategy. Of the 516,382 uniquely identified peptides in the second-round search, 1.7% (8,601) were non-canonical (**Fig. 3c** and **Supplementary Table S3**). The PTM profiles (**Fig. 3d**) of canonical (dark gray) and non-canonical (light gray) peptides appeared to be similar, with M oxidation being the most prevalent modification. The identified ncMAPs showed comparable spectra from patients within the same studies and from different studies (**Supplementary Figures S3, S4**, and **S5** provide examples of such similarities). The binding affinities of all 8,601 ncMAPs resulting from the second-round search were further investigated using NetMHCpan 4.1 (58). The binding prediction analysis showed a comparable binding rate for both the canonical (90%) and non-canonical (93%) MAPs, as depicted in **Fig. 3e**. We further took four additional independent post-search validation steps, including checking retention time shifts induced by PTMs, mass accuracy, and spectra comparison to those of the original studies, guaranteeing the correctness of the ncMAPs identified by the second-round search (see **Supplementary Method S2: Validation 4, 5, 6, and 7**, and **Supplementary Figure S2**).

Comparison of COD-dipp ncMAPs with the literature

To assess the performance of our COD-dipp method, we conducted a comparison with the results of peptide-PRISME by Erhard *et al.* 2020 (50). Our comparison was based on three

common studies (1,14,34) and resulted in 3,453 at 1% FDR from COD-dipp along with 4,576 ncMAPs at 10% FDR from Erhard *et al.* We first aligned Erhard *et al.*'s ncMAPs to the human proteome and eliminated a small fraction (1.4%) that matched the canonical protein sequences (**Fig. 4a**, left-hand side). Since the COD-dipp ncMAPs were restricted to the 3FT of protein-coding genes, we aligned the remaining ncMAPs from Erhard *et al.* to the same 3FT database for comparison purposes. **Fig. 4a (left-hand side)** shows that 68.25% of ncMAPs were successfully mapped to the 3FT database. The rest (30.35%) that did not align to any of the human proteome or the 3FT database are shown in yellow on **Fig. 4a** left-hand side. This unmapped fraction consisted of ncMAPs from regions of the genome not studied herein, such as intergenic regions, anti-sense translation, etc. The successfully mapped fraction to the 3FT database (navy) of 3,123 ncMAPs along with 3,453 ncMAPs from COD-dipp were then compared, as shown in **Fig. 4a** right-hand side (see **Supplementary Table S4**). peptide-PRISME shared 38% (1,197) of its ncMAPs (intersection) with COD-dipp (**Fig. 4a** right-hand side) and showed 62% (1,926) of exclusive ones. Adjusting the higher FDR used by peptide-PRISME from 10% to 1% increased the shared fraction to 48.9% (**Fig. 4b**), along with a ~ 2.4-fold decrease in total ncMAPs (from 4,576 to 1,916). At an FDR of 1%, COD-dipp identified 2.34 times more exclusive ncMAPs (2,298 vs. 979) from the 3FT of protein-coding genes.

To contextualize our findings from COD-dipp within the existing literature on ncMAPs, we compared our results with those of three previous studies: (I) Laumont *et al.* 2016 (13), (II) Chong *et al.* 2020 (6), and (III) Ouspenskaia *et al.* 2021 (16), as shown in Figure 4c. We used the same mapping procedure that was applied to peptide-PRISME results. We eliminated a fraction of sequences mapping to known proteins, which was 4%, 5%, and 3% of sequences for Chong *et al.* 2020, Laumont *et al.* 2016, and Ouspenskaia *et al.* 2021, respectively (see **Fig. 4c** left-hand side). **Fig. 4c** left-hand side shows in navy the fractions of ncMAPs that were successfully mapped to the 3FT database, which was 34.38% for Chong *et al.* 2020, 63.69% for Laumont *et al.* 2016, and 72.74% for Ouspenskaia *et al.* 2021. The remaining ncMAPs that did not align (**Fig. 4c** left-hand side in yellow) to any of the human proteome or the 3FT

database originate from sources not studied herein. For instance, Laumont *et al.* 2016 included 6-frame translation in their MS search database, which accounts for intergenic regions, anti-sense translation, long non-coding RNA, and retroelement sources. Both Chong *et al.* 2020 and Ouspenskaia *et al.* 2021 added Ribo-Seq detected proteins to their MS database searches, accounting for all possible nORFs, even those outside of known genes. The fractions successfully mapped to the 3FT database (navy) from these three studies, along with the 8,601 ncMAPs from COD-dipp, were then compared, as shown in **Fig. 4c** right-hand side (**Supplementary Table S4**). Intersections with COD-dipp were 31.42% for Chong *et al.* 2020, 38.3% for Ouspenskaia *et al.* 2021, and 45.8% for Laumont *et al.* 2016, respectively. In contrast, intersections with all other studies were 40% for Chong *et al.* 2020, 38.66% for Ouspenskaia *et al.* 2021, and 65.93% for Laumont *et al.* 2016. Hence, COD-dipp ncMAPs alone accounted for 78.55% of Chong *et al.* 2020's intersection, 96.07% of Ouspenskaia *et al.* 2021's intersection, and 69.47% of Laumont *et al.* 2016's intersection. COD-dipp ncMAPs accounted, on average, for 81.36% of the intersection when comparing three previously published ncMAP sets, thus validating our approach. With 2,168 ncMAPs (25%) shared with the literature and 6,433 new ncMAPs, we have revealed an atlas of non-canonical MHC class I presentation.

Properties and origins of ncMAPs

We compared the sequence lengths of canonical and non-canonical MAPs (**Fig. 5a**) and found them to be similar, with a slight skew of the non-canonical category toward longer lengths. This could be due either to an actual preference of ncMAPs toward longer sequence lengths or simply the consequence of requiring 3 amino acid differences from any known proteins favoring longer sequences. Next, we inspected ncMAPs according to their relative positions within protein-coding genes (**Fig. 5b**). Exonic regions translated in alternative frames were the main source of ncMAPs (19.2%). These events could arise from frameshift mutations, initiation codon readthrough (59), nORFs, or ribosomal slippage (11) during translation (i.e., ribosome frameshifting). Intronic regions were the second most abundant source of ncMAPs (12.2%).

These events can arise from frameshift mutations, nORFs, or IR. Interestingly, 5'-UTRs contributed to 10.2% of ncMAPs and have been shown to produce translation products through upstream ORFs along with a non-AUG start codon (60). Lastly, 3'-UTRs contributed the least toward ncMAPs (3.2%), potentially through stop codon read-through (61). It is important to note that these categorizations are not mutually exclusive and that an ncMAP sequence may have multiple assignments due to the overlapping nature of transcripts. We conducted three analyses to estimate how well the nORFs (I), IR (II), and frameshift mutations (III) could explain the detected ncMAPs. (I) ncMAPs with upstream start codons (AUG, CUG, UUG, GUG, and ACG) accounted for 63.4%, and 41.5% were predicted to be TIS using TITER (48) (**Fig. 5c** left-hand side). The breakdown of the TIS start codon distribution (**Fig. 5c**, right-hand side) showed CUG (L) as the most abundant nORF start codon, and 70% of the predicted TIS showed non-AUG start codons, in line with previous findings (15). (II) Translation frames of ncMAPs from intronic regions were checked for compatibility with upstream exons, and 49.4% were found in-frame with upstream exons, making IR events a possible source (**Fig. 5d**). (III) A total of 597 ncMAPs were found in aberrant proteins from frameshift mutations in cancer (42) (**Fig. 5e** and **Supplementary Table S5**). Eventually, 70.1% of ncMAPs were explicable by novel ORFs, IR, or frameshift mutations (**Fig. 5f**). ncMAPs were found to be presented by all 113 alleles in our dataset, except for the HLA-C*07:17 allele, mostly because of low sample coverage by MS for this allele (see **Supplementary Figure S6**). Furthermore, the average non-canonical presentation per HLA supertype (62) was 1%, except for A03, which was 5% (see **Supplementary Figure S6**).

Cancer selectivity of ncMAPs

Of the 8,601 identified ncMAPs, 2,758 were detected in the panel of normal healthy tissue by MS and were labeled as non-cancer-selective. The panel of normals originally consisted of healthy MS samples from all considered studies, which we extended by adding the HLA Ligand Atlas (39), a pan-tissue immunopeptidomic reference of 30 healthy tissue types obtained from 21 human subjects. **Fig. 6a** shows the ability of the extended panel of normals

to capture several more ncMAPs (12.85%) in healthy tissues that were not observed in our original panel of normals (19.22%). We assessed the coverage of tumor-only HLA alleles in healthy samples using the panel of normal samples. The 334 healthy samples covered 53% of the HLA alleles expressed in the tumor samples. Analysis of a subset of ncMAPs represented by the 57 shared alleles (i.e., present in both healthy and tumor samples) revealed a substantial overlap in HLA-binding motifs between the panel of normal samples and other samples. This was demonstrated by (I) the majority of identified ncMAPs being retained (7,513 out of 8,601) and (II) a comparable percentage of ncMAPs being detected in healthy samples through MS (36.46% with shared alleles versus 32% with all alleles) (see **Supplementary Figure S7**). To better understand the similarity between the HLA-binding motifs of the alleles represented in tumor-only samples and those represented in healthy samples, we generated a matrix of cosine distances of binding affinities and used t-SNE to reduce the dimensionality and visualize the data. Our results indicated a high level of similarity between the two, further supporting the notion that the 65 alleles in the panel of normal samples were representative of the tumor-only alleles (**Supplementary Figure S7**).

However, the lack of ncMAP detection in the panel of normals does not confirm cancer selectivity owing to the sensitivity limitations of MS. Proper cancer selectivity assessment should be performed at the gene expression level in healthy tissues. Hence, we retrieved the parent gene expression values (in TPM) of the remaining ncMAPs from the Genotype-Tissue Expression project (GTEx v8) (52). We first compared the gene expression levels of the following two groups: (I) ncMAPs detected in the panel of normals by MS (blue), and (II) remaining ncMAPs without detection in the panel of normals (red). **Fig. 6b** shows significantly higher gene expression for ncMAPs detected in healthy tissues (blue) than for those that were left undetected (red). Moreover, to ensure low toxicity levels in normal tissues, we filtered ncMAPs to retain those with parent genes expressed below 1 TPM and without evidence of protein expression in any healthy tissue except the testis (immune-privileged site) (**Fig. 6c**). By applying this stringent filter, we identified 24 ncMAPs derived from genes not expressed or

expressed only in trace amounts in healthy tissues. Of these, 17 were associated with proteins not detected in healthy tissues. **Table 1** provides a summary of these 17 cancer-selective ncMAPs, which we suggest as promising targets for clinical applications (see **Supplementary Table S3** for more details).

Discussion

The cartography of non-canonical antigen presentation revealed in our study arose from a harmonized large-scale analysis of immunopeptidomic data mapped to the human genome. Our innovations over the most recent trends in computational MS identified a diversity of peptides mapping to canonical and non-canonical translation products. We mapped deviations away from the reference proteome as mass shifts (PTMs) and applied a sequential approach to tackle the non-canonical immunopeptidome. Our proteogenomic pipeline allowed the identification of thousands of ncMAPs (8,601) derived from non-coding regions of protein-coding genes with an FDR of 1%. This was accomplished by analyzing a large collection of publicly available studies using COD-dipp, a highly modular large-scale pipeline that bypasses the challenge of multi-omics requirements and large MS databases when identifying ncMAPs.

Recent studies have suggested that the immunopeptidome is rich in PTMs (63), which can have profound effects on immune tolerance. T cells can discriminate between modified and non-modified epitopes, which has been demonstrated in the case of ubiquitination (64), glycosylation (65), phosphorylation (1,66). T-cell reactivity to PTMs is an effect of their central tolerance escape from the thymus (67). PTMs may also alter proteolytic activity, and consequently, peptide presentation by the MHC system (68). The open-search component sheds light on several PTMs implicated in immunogenicity (serine N-terminal acetylation, cysteinylolation, and cysteine tri-oxidation) and could provide insights for future studies on PTM-based epitopes. For instance, tri-oxidation of cysteine has the potential to alter the immune response (56); however, its mechanism of interaction with HLA molecules and T cells is still in

its infancy. Additionally, T cells can discriminate between cysteinylated and unmodified cysteine residues (57). Likewise, N-terminal serine acetylation is known for multifunctional regulation, acting as a protein degradation signal, inhibitor of endoplasmic reticulum (ER) translocation, and mediator of protein complex formation. Methionine sulfone (methionine dioxidation) has been found to occur *in vivo* in *Proteus mirabilis* (69), a Gram-negative bacterium present in malignant cancers (70), although it can result from the use of a strong oxidizing agent.

The validity of ncMAPs was rigorously tested using retention time correlation (experimental vs. theoretical), orthogonal second-round search, mass accuracy, PTM retention time shifts, HLA binding prediction, and PSM comparison with previously published results. Twenty-five percent of the identified ncMAPs accounted, on average, for 81.36% of intersections when compared with three other high-profile studies (6,13,16). In addition, COD-dipp revealed 6,433 new ncMAPs from protein-coding genes. Considering the high-quality and rigorous computational validation, the identification rate discrepancy is partly due to the performance of COD-dipp and the size of our dataset collection, making it the most exhaustive non-canonical library of MHC class I-associated peptides.

Our survey of the possible sources of ncMAPs revealed that 70.1% could be attributed to nORFs, IR, or frameshift mutations. We identified 597 ncMAPs downstream of known frameshift mutations in COSMIC, an understudied source of antigens in immunopeptidomic studies. Certainly, other biological processes not accounted for in this study could generate ncMAPs. For instance, mechanisms such as ribosomal slippage (11) and stop codon readthrough could explain some of the remaining ncMAPs (29.9%).

This study focuses on peptides from non-coding regions of the genome, referred to as non-canonical peptides. Unlike neoantigens, which derive from patient-specific mutations in cancer, these non-canonical peptides are not mutated and are present in both cancer and

healthy individuals. Although their presence in healthy samples makes their tumor specificity less clear, non-canonical peptides tend to be more abundant in cancer cells than in healthy cells. Over two decades ago, Ishii et al. (71) purified an octamer non-canonical antigen (IPGLPLSL or pRL1a) associated with heat shock proteins (HSPs) and validated their findings using MS. The isolated octamer non-canonical antigen pRL1a was derived from the 5'-untranslated region of the *AKT* gene in leukemia and induced tumor rejection. To the best of our knowledge, this was the first demonstration of a non-canonical antigen that confers immunity. Subsequent studies have shown that HSPs are beneficial for anticancer vaccines (72) because they bind canonical/non-canonical antigens with tumor rejection properties that end up being presented by MHC I and II molecules (73).

Numerous studies have suggested various possible candidates for cancer vaccines over the past 2–3 decades, and each has failed, at least partly, due to the issue of specificity. We used a conservative definition of cancer selectivity that follows three iterative steps. We searched for the identified ncMAPs over a panel of 334 normal MS samples and confirmed a fraction (32%) of non-cancer-selective ncMAPs. The remaining fraction (5,843, 68%) contained both cancer-selective ncMAPs and non-cancer-selective ncMAPs that were not detected by MS. We used the expression levels of the ncMAPs' parent genes across 29 healthy tissues as a means of prioritization (6,14,16). ncMAPs whose parent genes were expressed in any normal tissue above a threshold of 1 TPM were not considered cancer selective. However, we caution that this definition excludes the consideration of 92% of protein-coding genes. We revealed 17 rigidly defined candidates as cancer-selective ncMAPs, originating from genes and proteins that were completely absent or available in trace amounts in healthy tissues. We hope that this offers a sufficiently stringent approach to reducing toxicity in clinical applications. We provide a complete breakdown of all detected ncMAPs in **Supplementary Table S3**. We report the parent gene and protein expression values across healthy tissue types from the GTEx cohort and Human Protein Atlas, respectively. Moreover, we report their cancer-selectivity status conditioned on a gene expression cutoff (1 TPM) and lack of protein

expression in healthy tissues. This will allow the research community to make decisions regarding the peptides that should be retained or removed from their analyses. It is particularly important that we do not filter all peptides, as aberrant intron-retention and frame-shift mutations that are certainly cancer-specific may lie within these results and would not need this stringent filtering if found in subsequent studies.

Here, we provide a free and open-source informatics pipeline to study non-canonical peptides, along with a reservoir of potential targets that could be used in combination with T-cell therapies or cancer vaccines. We anticipate that this will help pave the way for future research on antigens from non-canonical sources and engage further oncology research on alternative sources of antigens.

We acknowledge that our study presents several limitations. First, our approach relies on a DDA MS, which is known for its dynamic range limitations. Thus, only the most abundant ncMAPs were identified. Moreover, owing to the technical limitations of MS, we require that our ncMAPs be at least 3 amino acids different from any known human protein. Thus, a substantial fraction could be eliminated, leading to underestimation of the non-canonical fraction. Second, because immunogenicity prediction is still in its infancy, the identified ncMAPs require further validation to qualify as tumor rejection–mediating antigens for clinical applications. Despite our efforts to identify cancer-selective targets, the toxicity of these peptides in healthy tissues requires further investigation.

Acknowledgments

This work was supported by the International Centre for Cancer Vaccine Science (Fundacja na rzecz Nauki Polskiej: MAB/3/2017) project is carried out within the International Research Agendas programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. This project has received funding from the

European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453. This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. The authors would like to thank 'CI-TASK, Gdansk', and the 'PLGrid Infrastructure, Poland' for providing their hardware and software resources.

Author Contributions

J. A. and G. B. conceived and initiated the project. J. A. and S. K. coordinated and supervised the project. G. B. and J. A. wrote the first draft of the manuscript. G. B. collected the online studies, developed the computational approach and software, processed the data, and coordinated the manuscript. G.B, J.A and H. G. created and revised the figures. The manuscript was reviewed and approved by all the authors. A. L., C. B., F. M. Z., C. P., H. A., A. R., D. J. H., T. R. H., and S. S., part of the 'KATY' consortium, and T. W., M. P., R. O., P. B. and K. L. revised the manuscript. G. B., D. B., K. W., A. P., D. R. G., R. F., S. K., J. A. Part of the 'International Centre for Cancer Vaccine Science' revised the manuscript.

References

1. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun.* 2016;7:1–16.
2. Newey A, Griffiths B, Michaux J, Pak HS, Stevenson BJ, Woolston A, et al. Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon α . *J Immunother Cancer.* 2019;7:309.
3. Ebrahimi-Nik H, Michaux J, Corwin WL, Keller GLJ, Shcheglova T, Pak H, et al. Mass spectrometry–driven exploration reveals nuances of neoepitope-driven tumor rejection. *JCI Insight.* 2019;4:e129152.

4. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat Rev Clin Oncol*. 2021;18:215–29.
5. Pearlman AH, Hwang MS, Konig MF, Hsiue EH-C, Douglass J, DiNapoli SR, et al. Targeting public neoantigens for cancer immunotherapy. *Nat Cancer*. 2021;2:487–97.
6. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun*. 2020;11:1293.
7. Malabat C, Feuerbach F, Ma L, Saveanu C, Jacquier A. Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *eLife*. 2015;4:e06722.
8. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife*. 2014;3:e03528.
9. Rivero-Hinojosa S, Grant M, Panigrahi A, Zhang H, Caisova V, Bollard CM, et al. Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat Commun*. 2021;12:6689.
10. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol*. 2018;36:1056–8.
11. Zook MB, Howard MT, Sinnathamby G, Atkins JF, Eisenlohr LC. Epitopes Derived by Incidental Translational Frameshifting Give Rise to a Protective CTL Response. *J Immunol*. 2006;176:6928–34.
12. Fang W, Wu C-H, Sun Q-L, Gu Z-T, Zhu L, Mao T, et al. Novel Tumor-Specific Antigens for Immunotherapy Identified From Multi-omics Profiling in Thymic Carcinomas. *Front Immunol*. 2021;12:748820.
13. Laumont CM, Daouda T, Laverdure J-P, Bonneil E, Caron-Lizotte O, Hardy M-P, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun*. 2016;7:10238.
14. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure J-P, et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med*. 2018;10.
15. Ruiz Cuevas MV, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep*. 2021;34:108815.
16. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol*. 2022;40:209–17.
17. Olsson N, Schultz LM, Zhang L, Khodadoust MS, Narayan R, Czerwinski DK, et al. T-Cell Immunopeptidomes Reveal Cell Subtype Surface Markers Derived From Intracellular Proteins. *PROTEOMICS*. 2018;18:1700410.
18. Demmers LC, Heck AJR, Wu W. Pre-fractionation Extends but also Creates a Bias in the Detectable HLA Class I Ligandome. *J Proteome Res*. 2019;18:1634–43.

19. Khodadoust MS, Olsson N, Chen B, Sworder B, Shree T, Liu CL, et al. B-cell lymphomas present immunoglobulin neoantigens. *Blood*. 2019;133:878–81.
20. Komov L, Kadosh DM, Barnea E, Milner E, Hendler A, Admon A. Cell Surface MHC Class I Expression Is Limited by the Availability of Peptide-Receptive “Empty” Molecules Rather than by the Supply of Peptide Ligands. *PROTEOMICS*. 2018;18:1700248.
21. Zeiner PS, Zinke J, Kowalewski DJ, Bernatz S, Tichy J, Ronellenfitsch MW, et al. CD74 regulates complexity of tumor cell HLA class II peptidome in brain metastasis and is a positive prognostic marker for patient survival. *Acta Neuropathol Commun*. 2018;6:18.
22. Bichmann L, Nelde A, Ghosh M, Heumos L, Mohr C, Peltzer A, et al. MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. *J Proteome Res*. 2019;18:3876–84.
23. Chong C, Marino F, Pak H, Racle J, Daniel RT, Müller M, et al. High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferony-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. *Mol Cell Proteomics*. 2018;17:533–48.
24. Koumantou D, Barnea E, Martin-Esteban A, Maben Z, Papakyriakou A, Mpakali A, et al. Editing the immunopeptidome of melanoma cells using a potent inhibitor of endoplasmic reticulum aminopeptidase 1 (ERAP1). *Cancer Immunol Immunother*. 2019;68:1245–61.
25. Marino F, Mommen GPM, Jeko A, Meiring HD, van Gaans-van den Brink JAM, Scheltema RA, et al. Arginine (Di)methylated Human Leukocyte Antigen Class I Peptides Are Favorably Presented by HLA-B*07. *J Proteome Res*. 2017;16:34–44.
26. Narayan R, Olsson N, Wagar LE, Medeiros BC, Meyer E, Czerwinski D, et al. Acute myeloid leukemia immunopeptidome reveals HLA presentation of mutated nucleophosmin. *PLoS One*. 2019;14:e0219547.
27. Shraibman B, Kadosh DM, Barnea E, Admon A. Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. *Mol Cell Proteomics*. 2016;15:3058–70.
28. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*. 2017;46:315–26.
29. Di Marco M, Schuster H, Backert L, Ghosh M, Rammensee H-G, Stevanovic S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. *J Immunol Baltim Md 1950*. 2017;199:2639–51.
30. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics MCP*. 2015;14:658–73.
31. Khodadoust MS, Olsson N, Wagar LE, Haabeth OAW, Chen B, Swaminathan K, et al. Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature*. 2017;543:723–7.

32. Andreatta M, Nicastrì A, Peng X, Hancock G, Dorrell L, Ternette N, et al. MS-Rescue: A Computational Pipeline to Increase the Quality and Yield of Immunopeptidomics Experiments. *Proteomics*. 2019;19:e1800357.
33. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol*. 2017;13:e1005725–e1005725.
34. Ternette N, Olde Nordkamp MJM, Muller J, Anderson AP, Nicastrì A, Hill AVS, et al. Immunopeptidomic Profiling of HLA-A2-Positive Triple Negative Breast Cancer Identifies Potential Immunotherapy Target Antigens. *Proteomics*. 2018;18:e1700465.
35. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, et al. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest*. 2016;126:4690–701.
36. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods*. 2018;15:363–6.
37. Newey A, Griffiths B, Michaux J, Pak HS, Stevenson BJ, Woolston A, et al. Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. *J Immunother Cancer*. 2019;7:309.
38. Löffler MW, Mohr C, Bichmann L, Freudenmann LK, Walzer M, Schroeder CM, et al. Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med*. 2019;11:28.
39. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer*. 2021;9:e002071.
40. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol*. 2020;38:199–209.
41. da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang H-Y, Shanmugam AK, Mellacheruvu D, et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods*. 2020;17:869–70.
42. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47:D941–7.
43. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods*. 2017;14:513.
44. An Z, Zhai L, Ying W, Qian X, Gong F, Tan M, et al. PTMiner: Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-translational Modification Characterization in Human Proteome*. *Mol Cell Proteomics*. 2019;18:391–405.

45. Qiao R, Tran NH, Xin L, Chen X, Li M, Shan B, et al. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat Mach Intell.* 2021;3:420–5.
46. Ivanov MV, Levitsky LI, Bubis JA, Gorshkov MV. Scavager: A Versatile Postsearch Validation Algorithm for Shotgun Proteomics Based on Gradient Boosting. *PROTEOMICS.* 2019;19:1800280.
47. Kent WJ. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* 2002;12:656–64.
48. Zhang S, Hu H, Jiang T, Zhang L, Zeng J. TITER: predicting translation initiation sites by deep learning. *Bioinformatics.* 2017;33:i234–42.
49. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics.* 2011;27:3423–4.
50. Erhard F, Dölken L, Schilling B, Schlosser A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunol Res.* 2020;8:1018–26.
51. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics.* 2010;11:237.
52. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation Biobanking.* 2015;13:311–9.
53. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science.* 2015;347:1260419.
54. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun.* 2014;5:5277.
55. Ojha R, Prajapati VK. Cognizance of posttranslational modifications in vaccines: A way to enhanced immunogenicity. *J Cell Physiol.* 2021;jcp.30483.
56. Trujillo JA, Croft NP, Dudek NL, Channappanavar R, Theodossis A, Webb AI, et al. The Cellular Redox Environment Alters Antigen Presentation. *J Biol Chem.* 2014;289:27979–91.
57. Parker R, Partridge T, Wormald C, Kawahara R, Stalls V, Aggelakopoulou M, et al. Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. *Cell Rep.* 2021;35:109179.
58. 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 Res.* 2020;48:W449–54.
59. Bullock TN, Eisenlohr LC. Ribosomal scanning past the primary initiation codon as a mechanism for expression of CTL epitopes encoded in alternative reading frames. *J Exp Med.* 1996;184:1319–29.
60. Starck SR, Tsai JC, Chen K, Shodiya M, Wang L, Yahiro K, et al. Translation from the 5' untranslated region shapes the integrated stress response. *Science.* 2016;351:aad3867.

61. Goodenough E, Robinson TM, Zook MB, Flanigan KM, Atkins JF, Howard MT, et al. Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR. *Proc Natl Acad Sci.* 2014;111:5670–5.
62. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. *BMC Immunol.* 2008;9:1.
63. Yi X, Liao Y, Wen B, Li K, Dou Y, Savage SR, et al. caAtlas: An immunopeptidome atlas of human cancer. *iScience.* 2021;24:103107.
64. Gavali S, Liu J, Li X, Paolino M. Ubiquitination in T-Cell Activation and Checkpoint Inhibition: New Avenues for Targeted Cancer Immunotherapy. *Int J Mol Sci.* 2021;22:10800.
65. Malaker SA, Ferracane MJ, Depontieu FR, Zarling AL, Shabanowitz J, Bai DL, et al. Identification and Characterization of Complex Glycosylated Peptides Presented by the MHC Class II Processing Pathway in Melanoma. *J Proteome Res.* 2017;16:228–37.
66. Penny SA, Abelin JG, Malaker SA, Myers PT, Saeed AZ, Steadman LG, et al. Tumor Infiltrating Lymphocytes Target HLA-I Phosphopeptides Derived From Cancer Signaling in Colorectal Cancer. *Front Immunol.* 2021;12:723566.
67. Raposo B, Merky P, Lundqvist C, Yamada H, Urbonaviciute V, Niaudet C, et al. T cells specific for post-translational modifications escape intrathymic tolerance induction. *Nat Commun.* 2018;9:353.
68. Kacen A, Javitt A, Kramer MP, Morgenstern D, Tsaban T, Shmueli MD, et al. Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. *Nat Biotechnol.* United States; 2022;
69. Buzy A, Bracchi V, Sterjiades R, Chroboczek J, Thibault P, Gagnon J, et al. Complete amino acid sequence of *Proteus mirabilis* PR catalase. Occurrence of a methionine sulfone in the close proximity of the active site. *J Protein Chem.* 1995;14:59–72.
70. Lin L, Jia L, Fu Y, Zhao R, Huang Y, Tang C, et al. A comparative analysis of infection in patients with malignant cancer: A clinical pharmacist consultation study. *J Infect Public Health.* 2019;12:789–93.
71. Ishii T, Udono H, Yamano T, Ohta H, Uenaka A, Ono T, et al. Isolation of MHC Class I-Restricted Tumor Antigen Peptide and Its Precursors Associated with Heat Shock Proteins hsp70, hsp90, and gp96. *J Immunol.* 1999;162:1303–9.
72. Bloch O, Lim M, Sughrue ME, Komotar RJ, Abrahams JM, O'Rourke DM, et al. Autologous Heat Shock Protein Peptide Vaccination for Newly Diagnosed Glioblastoma: Impact of Peripheral PD-L1 Expression on Response to Therapy. *Clin Cancer Res.* 2017;23:3575–84.
73. Binder RJ. Immunosurveillance of cancer and the heat shock protein-CD91 pathway. *Cell Immunol.* 2019;343:103814.

Tables

ID	Peptide	Gene name	Mean expression in healthy tissues (TPM)	Number of healthy tissues with protein expression	Annotation
1	AFAPFPTQF	CXorf49B	0.01	0 of 56	Cancer selective
1	AFAPFPTQF	CXorf49	0.01	0 of 56	Cancer selective
1	AFAPFPTQF	RP11-402P6.15	0.10	0 of 56	Cancer selective
2	DYIHFVHHF	RP11-325B23.2	0.00	0 of 56	Cancer selective
3	EALSASQALYTR	HIST1H4L	0.04	43 of 56	
4	ELIKAFSK	GNGT1	0.05	1 of 56	
5	ESAGLFQVPR	SUN3	0.13	3 of 56	
6	EVEKILIQY	KCNU1	0.05	0 of 56	Cancer selective
7	EVPGAQQQQGPR	CTAG2	0.15	0 of 56	Cancer selective
7	EVPGAQQQQGPR	CTAG1B	0.03	0 of 56	Cancer selective
7	EVPGAQQQQGPR	CTAG1A	0.06	0 of 56	Cancer selective
8	FPVDVDHAVL	CTAG2	0.15	0 of 56	Cancer selective
8	FPVDVDHAVL	CTAG1B	0.03	0 of 56	Cancer selective
8	FPVDVDHAVL	CTAG1A	0.06	0 of 56	Cancer selective
9	ILSDNIRNL	C1orf94	0.14	0 of 56	Cancer selective
10	IPKDKSKNK	C2orf83	0.02	0 of 56	Cancer selective
11	KLLELIKAFSK	GNGT1	0.05	1 of 56	
12	KNNIYAFKI	RP11-231113.2	0.01	0 of 56	Cancer selective
13	KTLHLTIVK	C12orf50	0.07	0 of 56	Cancer selective
14	KYLSRFRPK	TRPC5	0.08	0 of 56	Cancer selective
15	MVRSPEEGSLR	TEX19	0.13	0 of 56	Cancer selective
16	MVRSVSAAR	HIST1H2BB	0.26	44 of 56	
17	MVRSVSAAR	HIST1H2BB	0.26	44 of 56	
18	REEAPRGVRM	CTAG2	0.15	0 of 56	Cancer selective
18	REEAPRGVRM	CTAG1B	0.03	0 of 56	Cancer selective
18	REEAPRGVRM	CTAG1A	0.06	0 of 56	Cancer selective
19	SAGLFQVPR	SUN3	0.13	3 of 56	
20	SQVHKFFLL	OR9Q1	0.04	0 of 56	Cancer selective
21	SYGIYIYTY	SLC15A5	0.06	0 of 56	Cancer selective
22	TVSHQIIFY	EXD1	0.06	0 of 56	Cancer selective
23	VIQKVILVV	MGAT4D	0.03	0 of 56	Cancer selective
24	YYFILEHAKY	SOX30	0.29	0 of 56	Cancer selective

Table 1: List of cancer-selective non-canonical MHC-associated peptides. The mean parent gene expression in TPM was derived from 29 healthy tissues from the GTEx v8 dataset. The number of healthy tissues with protein expression was obtained from the Human Protein Atlas v22.0.

Figures

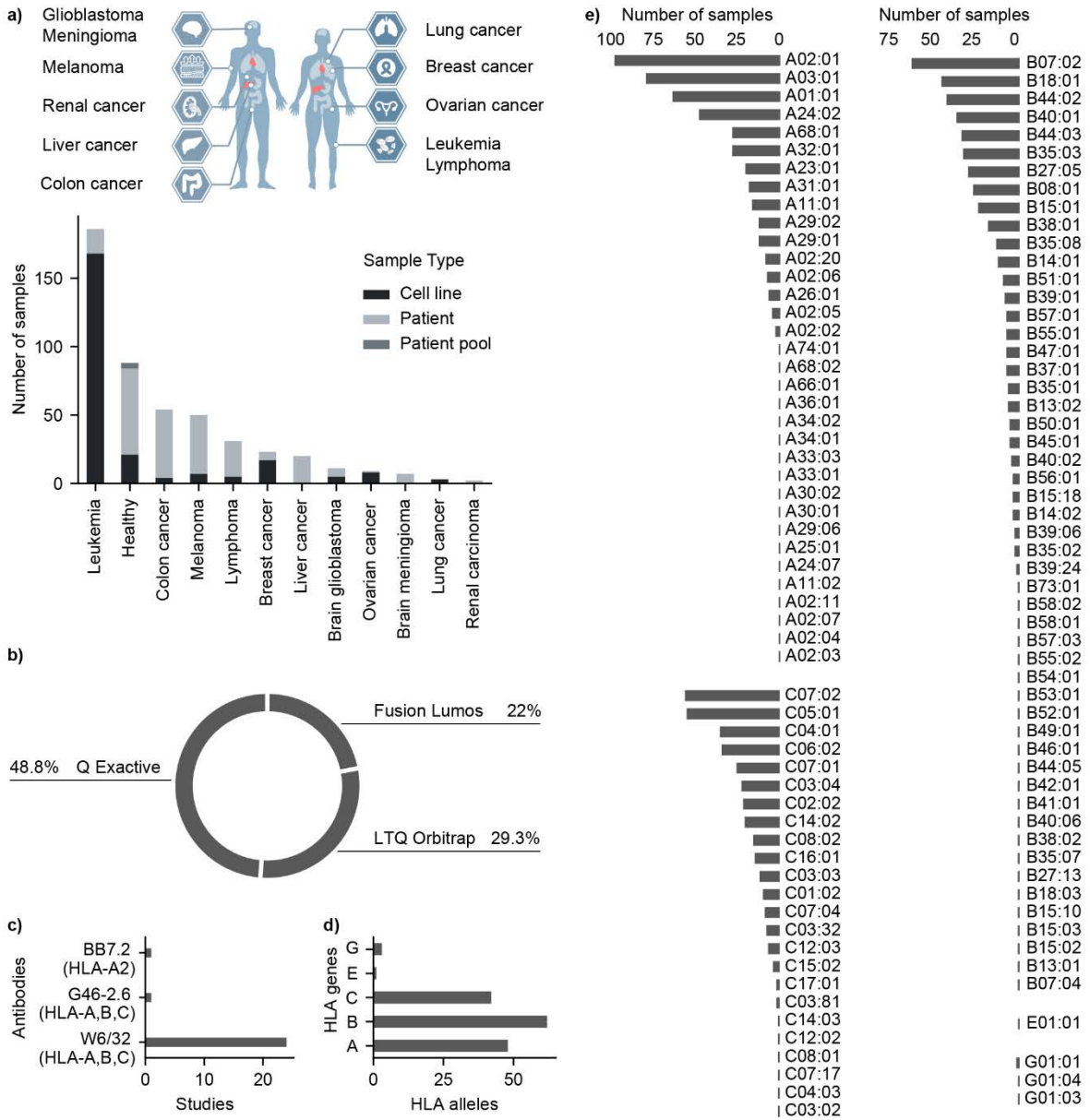


Figure 1: Infographics of immunopeptidomic datasets included in this study. a) Different types of cancer considered in this study with the number of samples and sample types per cancer type. **b)** Proportions of different mass spectrometry instruments used in this study. **c)** Antibodies used for immunoprecipitation (IP) **d)** Overall count of HLA alleles per HLA gene. **e)** Overall count of mass spectrometry immunopeptidomic samples per HLA allele.

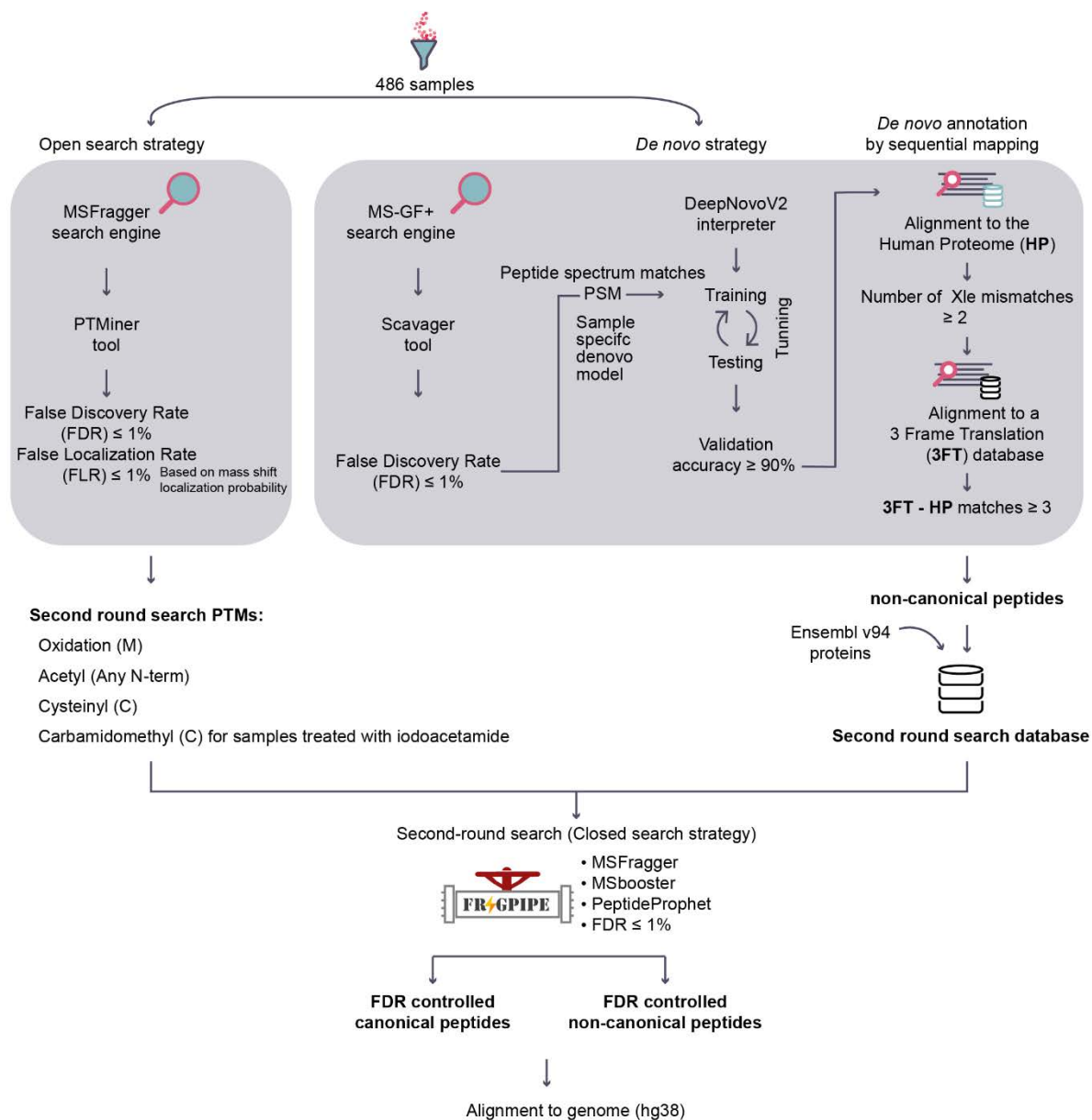
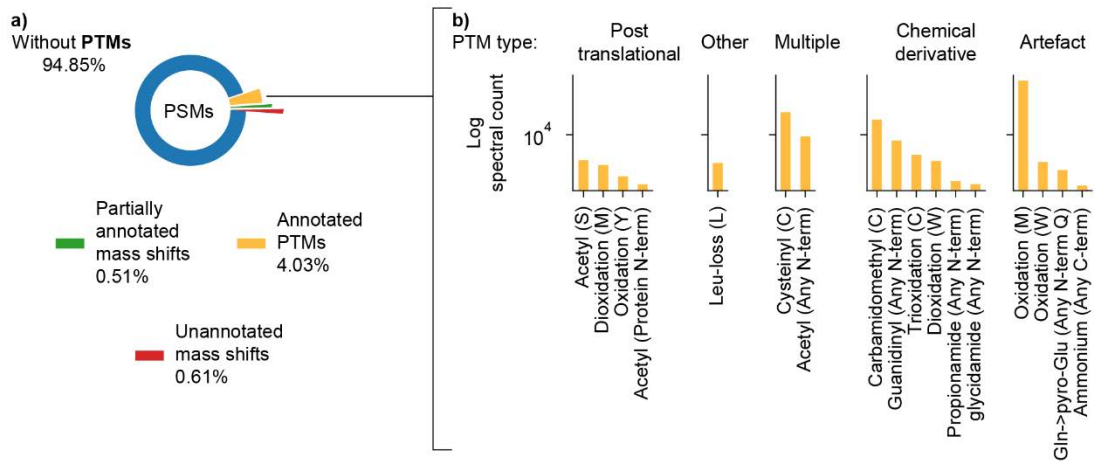


Figure 2: COD-dipp: A new high-throughput pipeline for a deep interrogation of immunopeptidomic datasets. Samples are first analyzed with an open search strategy to detect the landscape of post-translational modifications (PTMs). A false localization rate (FLR) for the PTMs and false discovery rate (FDR) of 1% are applied. Simultaneously, the samples are analyzed using a novel *de novo* approach to identify non-canonical peptides. The *de novo* strategy trains a model per sample using quality-controlled peptide-spectrum matches from the MS-GF+ search engine to learn the direct interpretation of sample-specific mass spectra. The MS-GF+ results are split into three groups: training and testing to tune the hyperparameters and account for overfitting, and a validation group to approximate the accuracy per sample. *De novo* predicted peptides with an accuracy of at least 90% are sequentially mapped against the Human proteome (HP) then a 3-frame translation (3FT) database of protein-coding genes (1 mismatch allowed between leucine/isoleucine, i.e., Xle). Predicted *de novo* peptides matching any known protein are labeled “canonical”. Peptides mapping to the 3FT database with at least 3 amino acids mismatches from any known protein sequence are labeled “non-canonical”. Lastly, a second-round search is performed as a validation approach. Four of the most abundantly identified PTMs and a custom database consisting of ENSEMBL proteins and non-canonical peptides are considered. The resulting canonical and non-canonical peptides are controlled to an FDR of 1% and aligned to the hg38 human genome.

Open search post-translational modifications (PTMs)



Second-round search (1% FDR controlled non/canonical peptides)

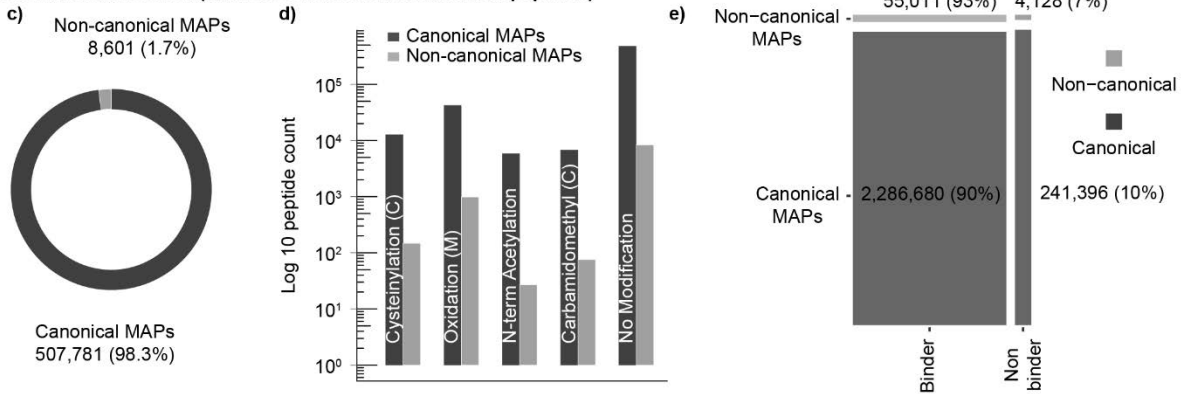


Figure 3: Landscape of post-translationally modified and non-canonical MHC class I-associated peptides (ncMAPs). Open search: a) Overview of post-translational modifications (PTMs) identified by open search (blue: spectra without PTMs, orange: spectra with a known UNIMOD PTM localized on a specific amino acid on the peptide. Green: The mass shift is localized, however the known PTM options do not fit the modified residue. Red: Otherwise). **b)** Most abundant “annotated PTMs” grouped by type. **Second-round search: c)** Fraction of canonical (dark gray) and non-canonical (light gray) MAPs in the immunopeptidome. **d)** Proportion of canonical (dark gray) and non-canonical (light gray) MAPs with/without post-translational modifications. **e)** Fraction of binders versus non-binders for both canonical and non-canonical MAPs using NetMHCpan 4.1.

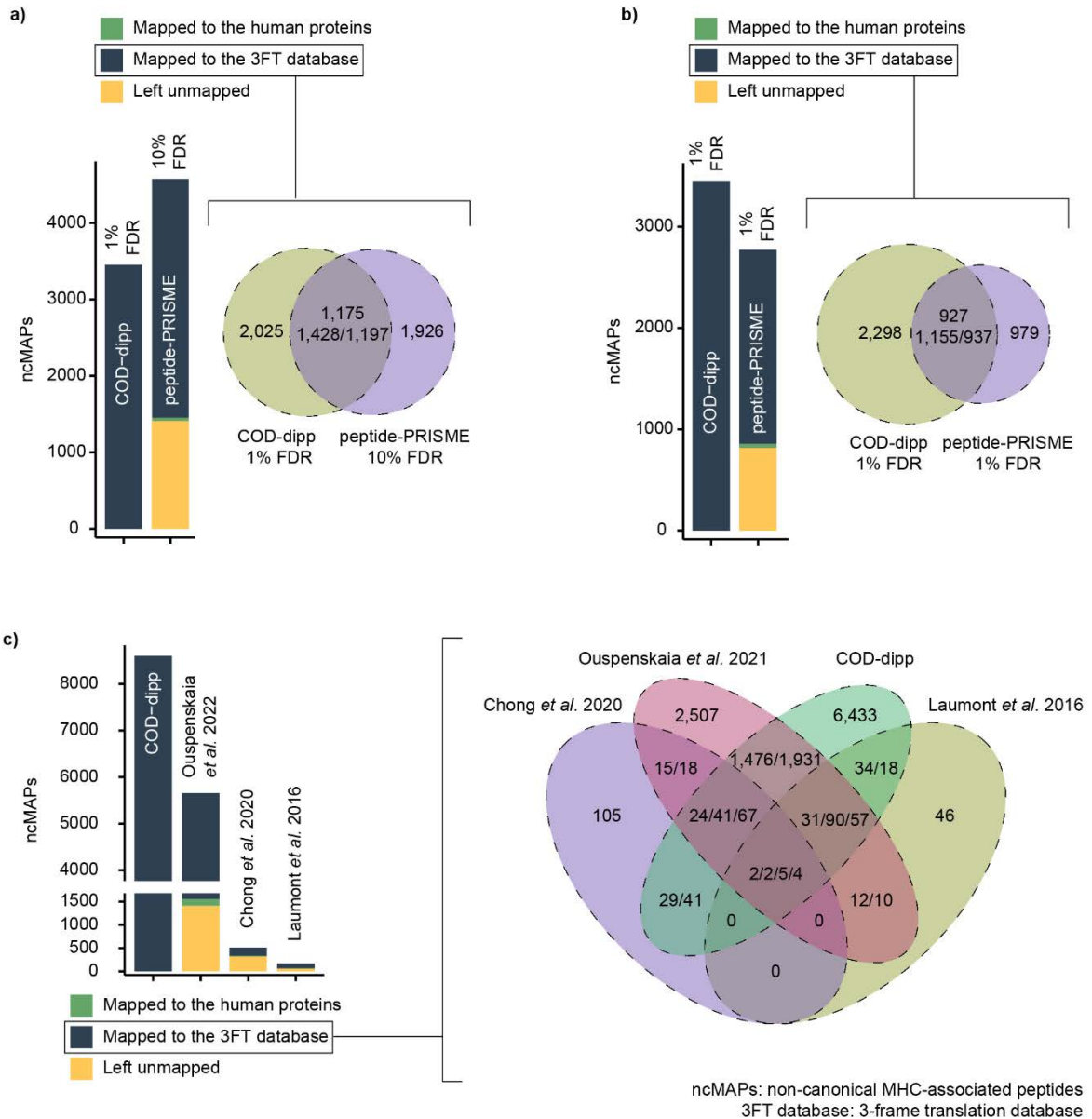


Figure 4: Comparison of COD-dipp non-canonical MHC class I-associated peptides (ncMAPs) with other studies. Since the COD-dipp ncMAPs are restricted to the 3-frame translation (3FT) of protein-coding genes, sequences from the literature were aligned to the same 3FT database for comparison purposes. The intersection is based on genomic coordinates to deal with sequences that partially match (i.e., longer, shorter, or partially overlapping). Since the Venn is generated by overlapping genomic coordinates of the ncMAPs, the original counts for each study are listed from left to right (i.e., on the right-hand side of panel c, the notation 29/41 refers to 29 instances for Chong *et al.* 2020 and 41 for COD-dipp). **a)** Comparison with peptide-PRISM published ncMAPs at a 10% FDR. COD-dipp ncMAPs were restricted to 3 studies in common with Erhard *et al.* 2020. **b)** Comparison with Peptide-PRISM published ncMAPs at a 1% FDR. COD-dipp ncMAPs were restricted to 3 studies in common with Erhard *et al.* 2020. **c)** Comparison of the atlas of ncMAPs revealed by COD-dipp to 3 previous studies.

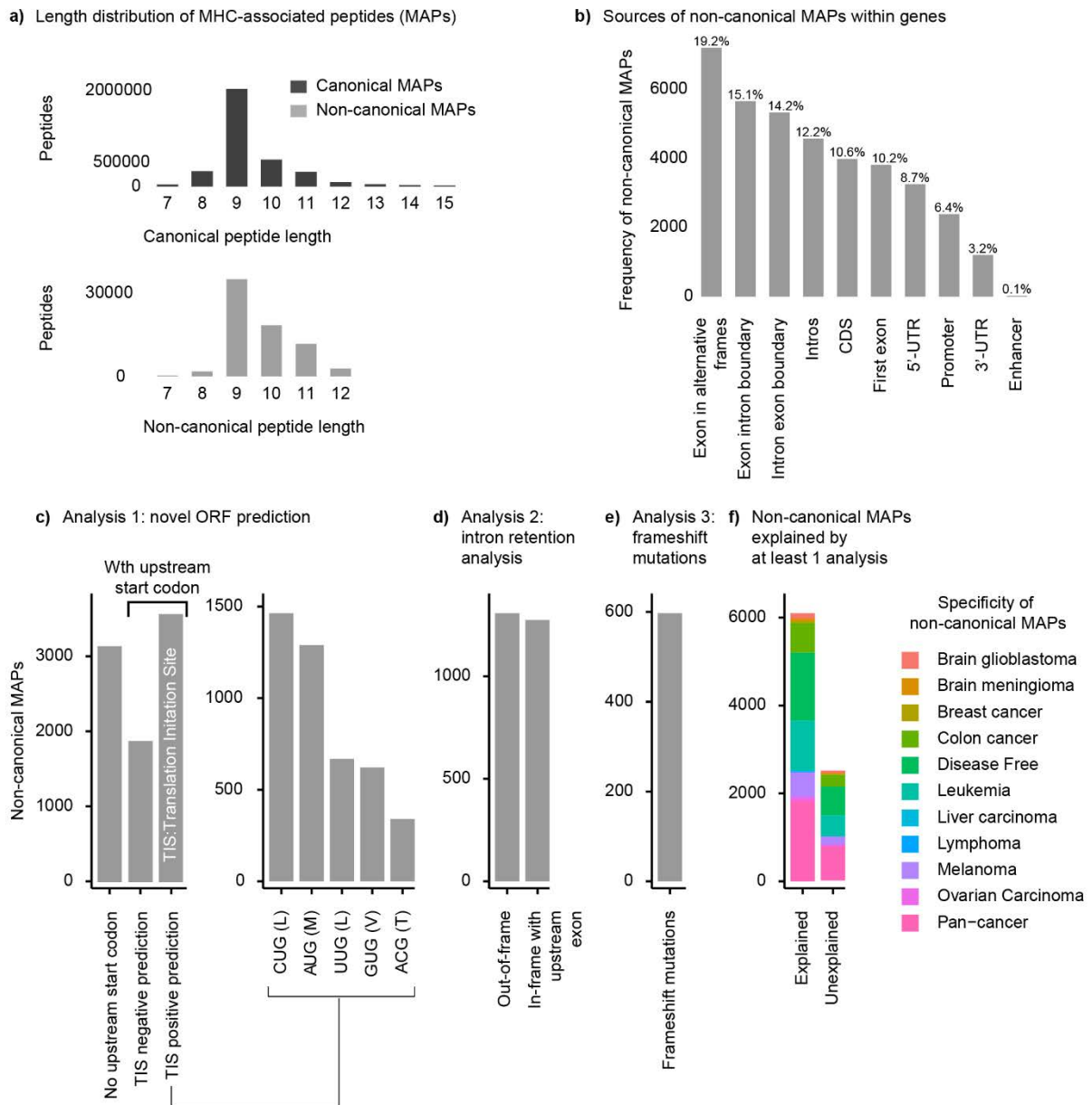


Figure 5: Origins of non-canonical MHC class I-associated peptides (ncMAPs). **a)** Peptide length distribution of canonical (dark gray) and non-canonical (light gray) MAPs. **b)** Annotation of ncMAPs across gene features. **c)** Analysis of ncMAPs that could originate from novel open reading frames (ORF). Upstream start codons of non-canonical MAPs are analyzed for their potential to initiate translation and produce ORFs (left-hand side) as a source of ncMAPs. The frequencies of different start codons for positively predicted translation initiation sites (TIS) are shown on the right-hand side. **d)** Analysis of ncMAPs from intronic regions that may originate from intron retention (IR) events. Translation of MAPs from IR sources should be in-frame with the corresponding upstream exons. **e)** Analysis of ncMAPs that could originate from frameshift mutations in cancer. ncMAPs are aligned to an in-silico translated protein database of COSMIC somatic frameshift mutations. **f)** Summary indicating whether the ncMAPs can be accounted for by any of the analyses conducted in panels c, d, or e.

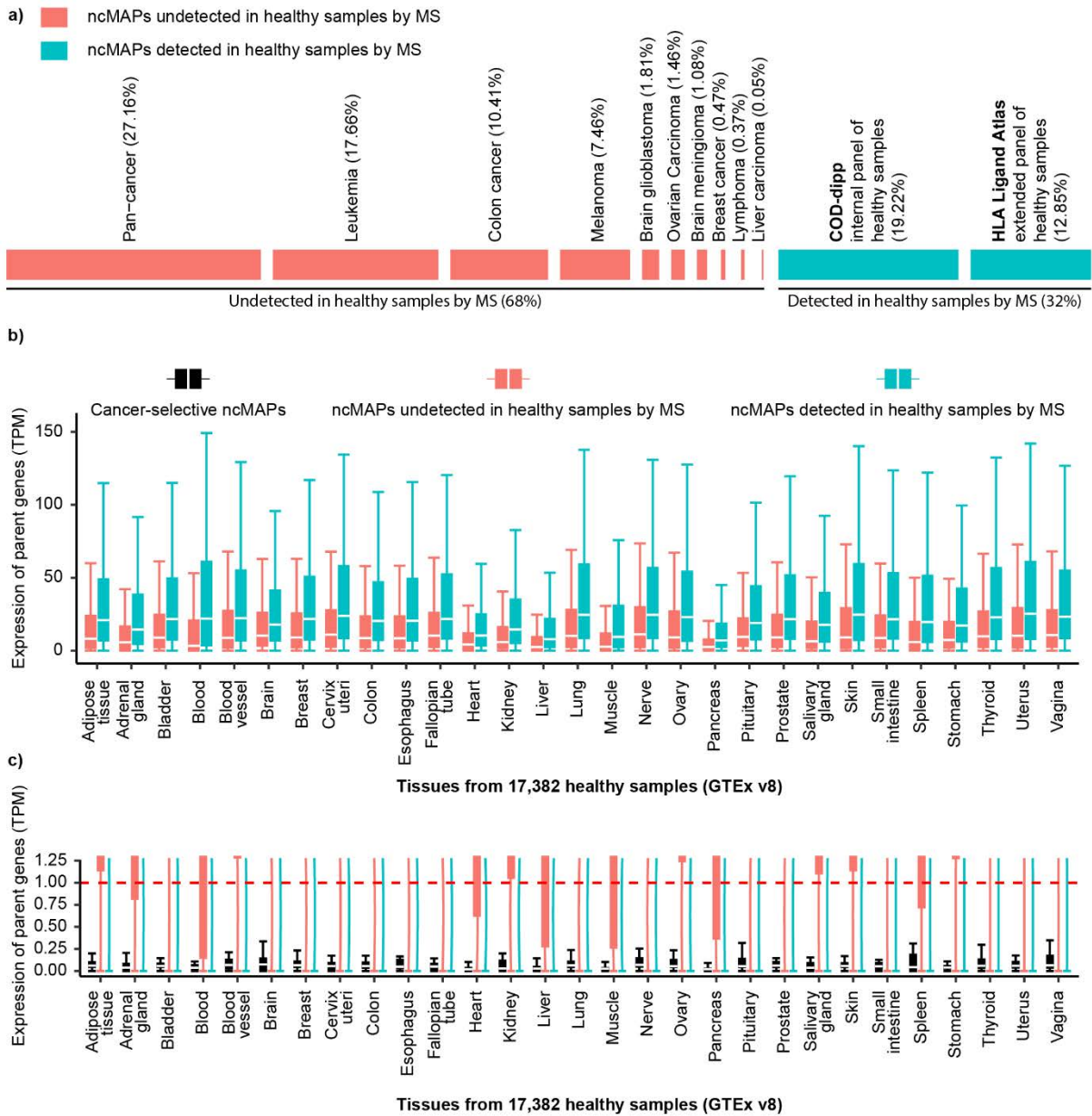


Figure 6: Cancer selectivity of non-canonical MHC class I-associated peptides (ncMAPs). (a) Percentage of ncMAPs that were solely in healthy and/or tumor samples by MS (blue) and ncMAPs undetected in healthy samples by MS (red). (b) Parent gene expression of ncMAPs in TPM in 29 healthy tissues from 17,382 samples (GTEx v8 dataset). ncMAPs are distributed into two groups: (I) ncMAPs detected in healthy samples by MS in blue, (II) ncMAPs undetected in healthy samples by MS in red. (c) Parent gene expression of ncMAPs in TPM in 29 healthy tissues from 17,382 samples (GTEx v8 dataset). A limit on the gene expression (y-axis) of 1.2 TPM was set to visualize cancer-selective ncMAPs in black.

Supplementary Figures S1-S7

Figure S1

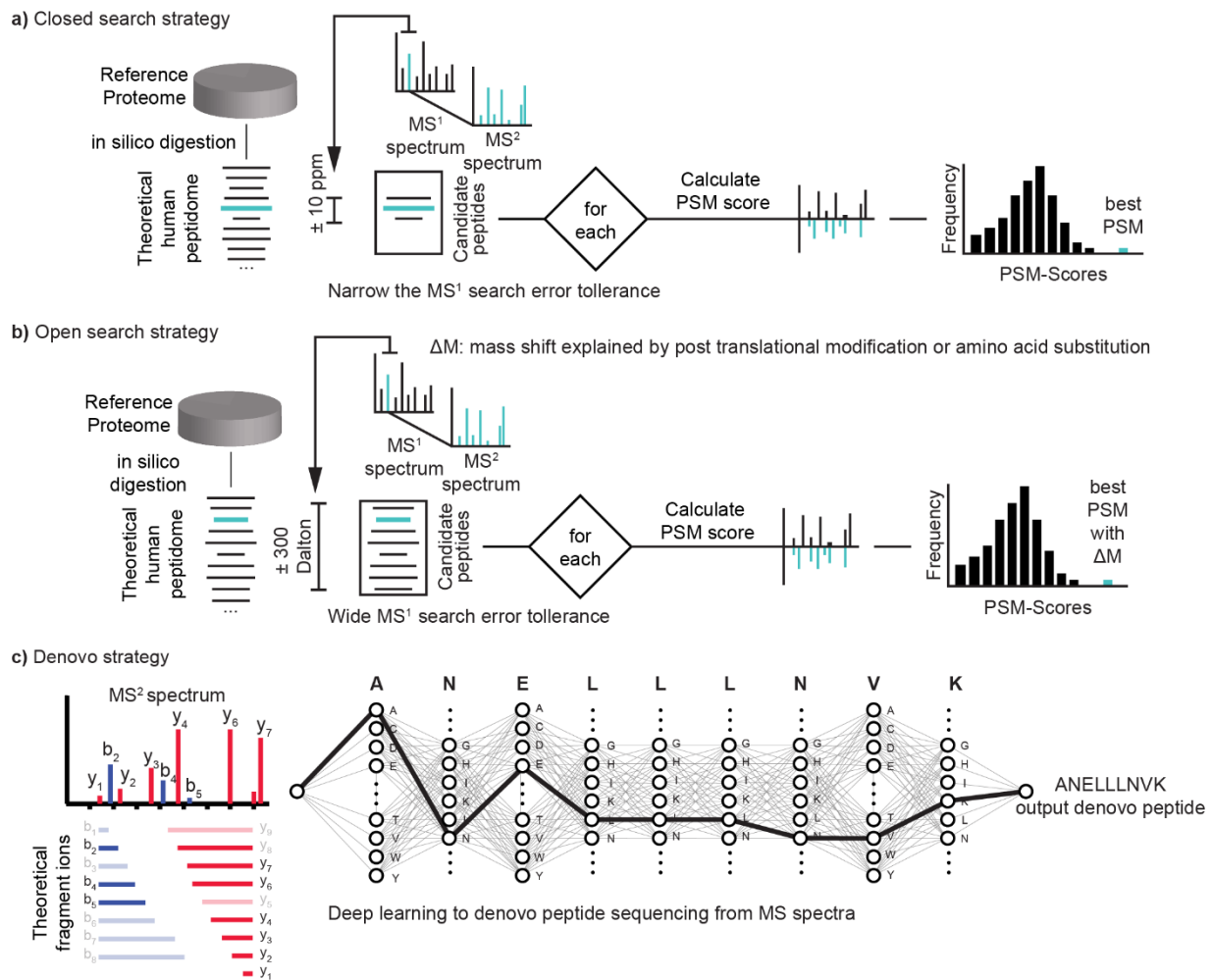
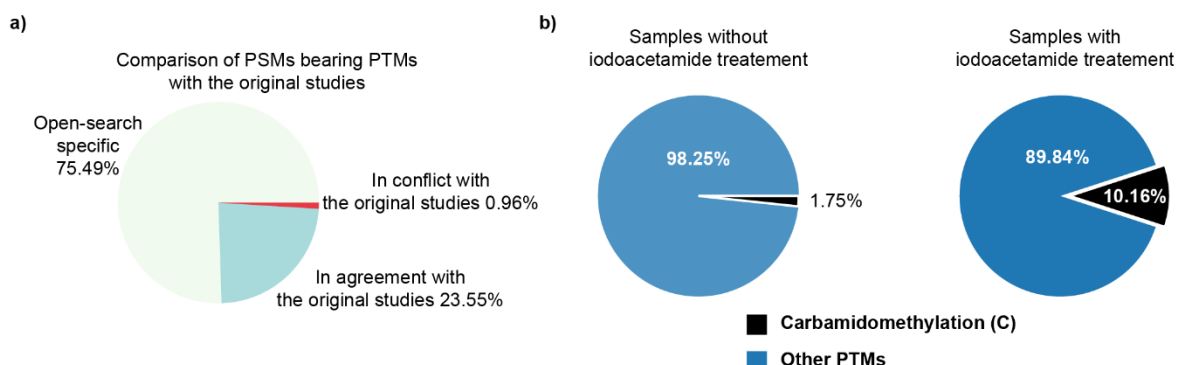


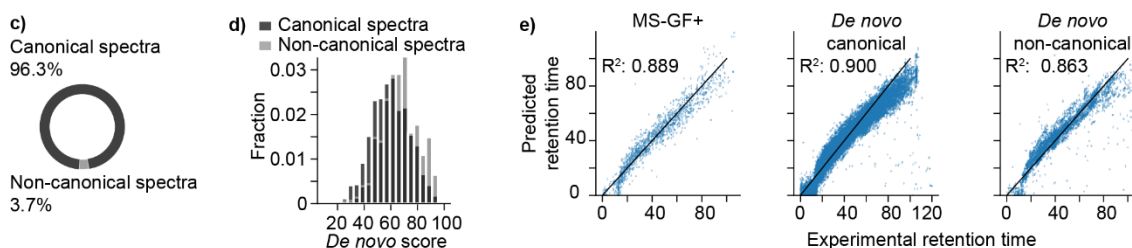
Fig. S1: The mass spectrometry strategies used in this study. a) A closed search approach (supervised approach) requires a reference protein sequence database that contains the expected proteins within the sample. These protein databases are in silico digested and peptides falling within a certain error tolerance are chosen as candidate peptide assignments. Each candidate peptide is then scored against the spectrum using an algorithm-specific methodology, and the highest scoring one is assigned as the sequence of the MS2 spectrum. **b)** Open search strategy (semi-supervised approach) widens the MS1 search error tolerance to identify peptides that would have been missed due to the mass shifts caused by mutations or post-translational modifications. **c)** *de novo* strategy (unsupervised approach) attempts to annotate spectra without a reference proteome by predicting a peptide sequence by directly reading the MS2 spectra.

Figure S2

Open search quality control



De novo peptides quality control



Second round search quality control

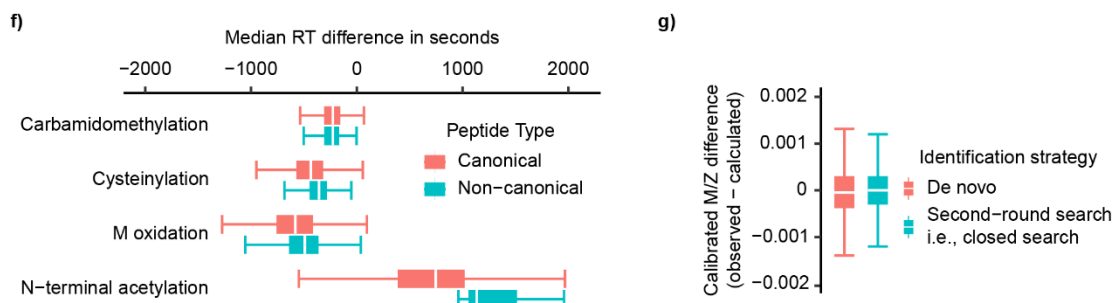


Fig. S2: Additional quality control for the mass spectrometry peptide-spectrum matches (PSMs) of non-canonical and post-translationally modified MHC-associated peptides. a) Comparison of PSMs identified by our open search with PSMs of the original studies. b) Percentage of carbamidomethylation within the subset of post-translationally modified peptide-spectrum matches for iodoacetamide-treated and untreated samples. c) *De novo* identified spectra from canonical (dark gray) and non-canonical (light gray) sources. d) *De novo* score distributions of canonical and non-canonical spectra. e) Correlation between predicted and experimental retention times for MS-GF+ and *de novo* peptides. f) Median retention time (RT) difference between peptides with and without a specific post-translational modification (PTM). Deviations between PTM-modified and unmodified canonical peptides are shown in red, and deviations between non-canonical peptides are shown in blue. Median RT refers to the retention time median value of multiple PSMs of the same peptides in a specific mass spectrometry run. The median RT difference refers to the difference between the median RT of the unmodified peptide and its modified counterpart. g) Boxplot of the mass differences between the observed mass over charge (M/Z) and the calculated M/Z of the non-canonical peptides identified using *de novo* (red) and second-round search (blue).

Figure S3

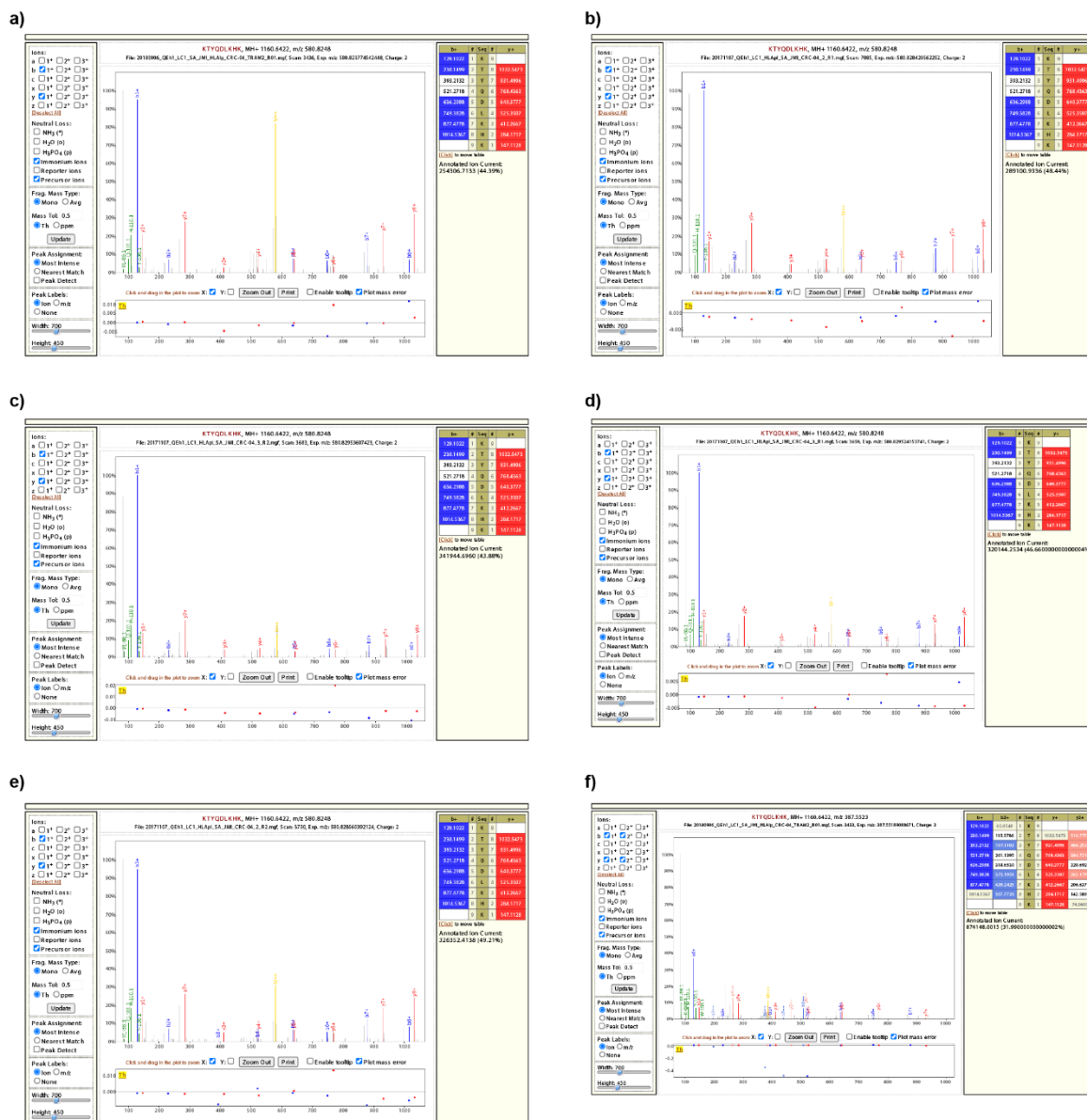


Fig.S3: Spectra of the non-canonical MHC-associated peptide KTYQDLKHK from the PXD014017 dataset of a colorectal cancer patient (CRC-4). Panels a and e show spectra from a replicate treated with trametinib. Panels b, c, d, and f show the spectra from four different replicates of the same patient (CRC-4) that were left untreated.

Figure S4



Fig. S4: Spectra for the non-canonical MHC-associated peptide HLLDNKTLFQQL from multiple datasets. Panel a shows a spectrum from the PXD012083 dataset of an acute myeloid leukemia patient. Panels b, c, and e show spectra from the PXD003790 dataset of a brain glioblastoma cell line (T98G). Panel d shows a spectrum from the PXD003790 dataset of a brain glioblastoma cell line (U87). Panel (f) shows the spectrum from the PXD007596 dataset of a breast cancer cell line (MCF-7).

Figure S5

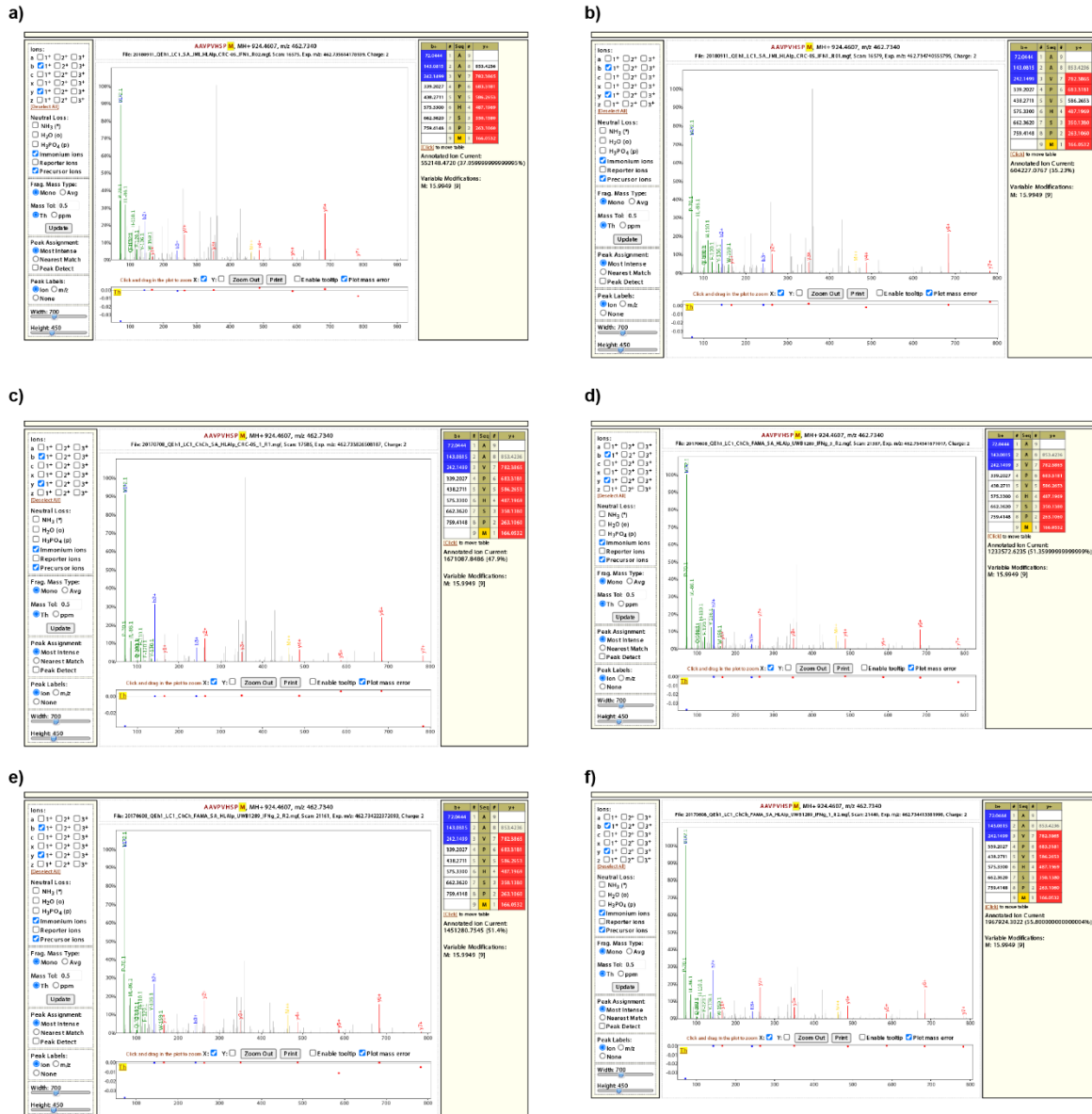


Fig. S5: Spectra for the non-canonical MHC-associated peptide AAVPVHSPM(oxidation) from multiple datasets. Panels a and b show the spectra from dataset PXD014017 of a colon cancer patient treated with IFN- γ . Panel c shows a spectrum from dataset PXD014017 of the same colon cancer patient who was left untreated. Panels d, e, and f show spectra from the PXD006939 dataset of an ovarian carcinoma cell line (UWB1289) for three different biological replicates.

Figure S6

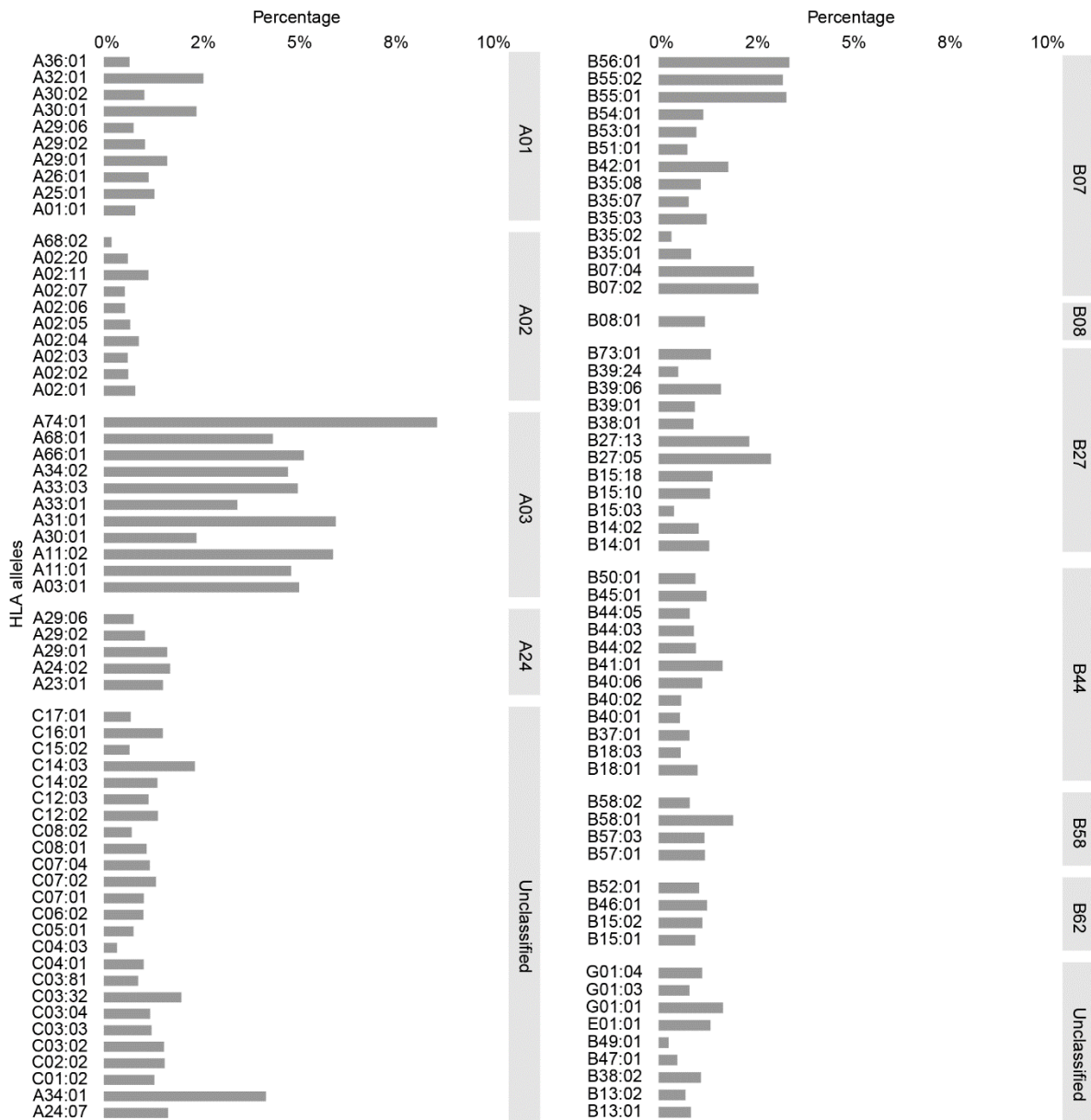


Figure S6: HLA supertypes and non-canonical MHC-associated peptides (ncMAPs) expression. The percentages of unique ncMAPs are shown for all 114 HLA alleles grouped into supertypes to reduce the complexity.

Figure S7

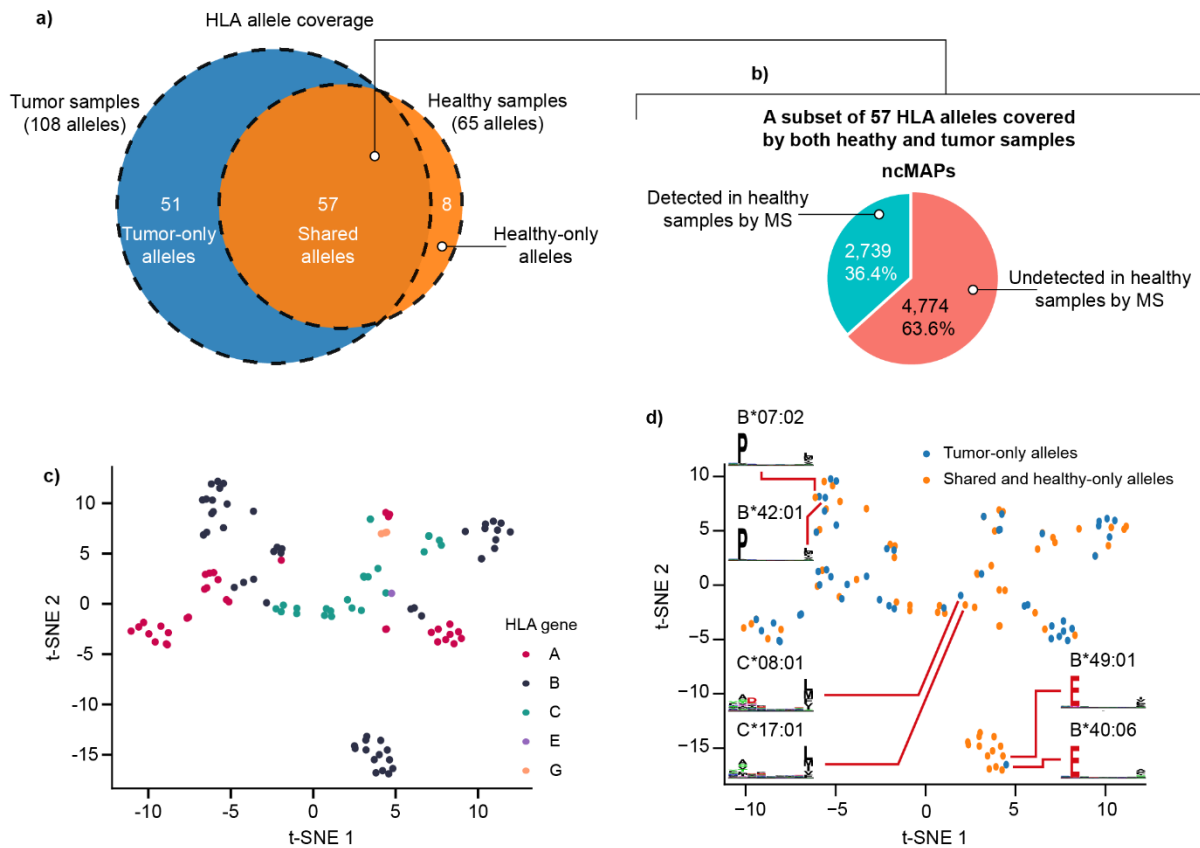


Fig.S7: Comprehensiveness of the panel of normals in term of HLA-binding motifs. **a)** HLA allele coverage intersection between tumor and healthy samples. **b)** Percentage of ncMAPs that are detected in healthy samples (red) versus those undetected in healthy samples (blue) for the subset of peptides presented by the shared allele *i.e.*, 57 HLA alleles common between tumor and healthy samples. **c)** HLA-binding landscape of all alleles colored by HLA gene type. **d)** HLA-binding landscape of all HLA alleles colored in blue for tumor-only alleles and in orange for shared or healthy-only alleles. Panels c and d show the similarity in HLA-binding motifs between all alleles in our dataset. As different HLA genes should present dissimilar binding motifs, panel c shows that different HLA genes map to distinct areas of the plot, supporting the idea that dissimilar HLA-binding motifs would appear in separate areas. Panel d shows a high similarity in the HLA-binding motifs between alleles covered by tumor-only samples (blue dots) and alleles covered by healthy samples (orange dots). In agreement with the findings of panel b, it is apparent that the 65 alleles in our panel of normals are representative of the tumor-only alleles in terms of HLA-binding motifs.

Supplementary Notes

Note 1: Dataset selection

List of keywords used for selecting datasets from PRIDE: Immunoprecipitation, Immunopeptidome, Peptidomics, Affinity purification, Mhc, Peptidome, Hla, Immunopeptidomics, Mhc class i, Ip, Hla peptidome, Hla-b*27, Hla class ii, Neoantigens, Immunoinformatics, Hla-c, Mhc class 1 ligands, Proteogenomic cryptic mhc lc-msms maps, Mhc class i antigen presentation pathway, Mhc-i peptides, Mhc i, Immunopeptidome; hla; lc-ms/ms; netmhcp; binding prediction, Mhc ii, Mhc-i peptide-loading complex, Mhc affinity prediction, Mhc-ii peptidomics, Mhc ligandome, Mhc i-associated peptides, Mhc-i, Mhc class ii, Antigen presentation/ mhc class ii/ immunopeptidome/ peptide editing/ polymorphism, Mhc-i peptidomics, Shotgun proteomics; immunoprecipitation; meiosis; conserved proteins; meioc; , Anti-hla immunopurification, Immunopeptidome; hla; lc-ms/ms; netmhcp; binding prediction, Personalized immunotherapy, Immunoprecipitation, Immunoprecipitation, Immunoaffinity purification, Immunoprecipitations, Immunopurification, Antigen presentation/ mhc class ii/ immunopeptidome/ peptide editing/ polymorphism, Hla-ii, Hla peptides, Hla-e, Hla-b*51, Hla class i peptides, Ducaf; hla-drb1*03:01, Hla typing, Hla-g, 'Hla class i ligandome; hla class i peptide ligands; high ph reversed phase; strong cation exchange; pre-fractionation', Hla-b40, Hla binding motifs, Hla-dm, Hla-b27, Immunopeptidome; hla; lc-ms/ms; netmhcp; binding prediction, Hla-b*58:01, Hla-b*40:02 peptidome, Hla-dr peptides, Hla-dr, Hla-a, Hla-b57, Hla class i, Hla-i, Hla-a2, Hla-b, Interferon gamma; proteomic analysis; hla class i; apm, Hla-i peptides, Hla-ligand, Hla-b*57:03, Hla-ligandomics, Hla-a*29:02, Hla-dr15, Hla-class i, Hla-restricted peptide

Note 2: Correctness of the identified peptides

The most definite validation metric of correctness is shown by a high similarity between the MS/MS spectra of the endogenous and synthetic peptides, as well as the co-elution of the light and heavy peptide pairs. Considering the impossibility of performing such an analysis due to the reliance of this study on publicly available data. We have assessed the correctness of the identified peptides in a series of quality control experiments.

Open-search quality control

Validation 1: We compared the identifications obtained with open search in this study with the identifications in the original studies at the peptide-spectrum match (PSM) level (*i.e.*, for each MS/MS spectrum). We successfully collected PSM information from 19 of the 25 analyzed datasets. The remaining 6 datasets presented some challenges. Three of these datasets (PXD004233, PXD008937, and PXD009531) reported PSM-level data but without FDR control, and three (PXD012083, PXD004746, and PXD010808) reported PSM in a format that prevented us from recovering the MS/MS scan numbers from the raw files. We compared the PSMs identified by our open search to those reported in the original studies and considered an agreement when the same mass spectrometry scan showed (I) an identical peptide sequence and (II) an identical mass shift introduced by the PTM. We found that 96.1% of the modified PSMs were identical in both sources (49,918 out of 51,945). To expand on the remaining 3.9% of the PSMs that were in conflict, we inspected the discrepancies and determined that they consisted of PTMs with monoisotopic masses close to those of some amino acids. For example, N-term glycidamide (87.03203 Da) can be misinterpreted as serine (87.0782 Da), carbamidomethyl (57.02146 Da) as glycine (57.02146 Da), N-term Propionamide (71.03711 Da) as alanine (71.03711 Da), phenethyl isothiocyanate (163.04557 Da) as tyrosine (163.1760 Da), N-term dicarbamidomethyl (114.04293 Da) as asparagine (114.04293 Da), and 4-hydroxynonenal (156.115030 Da) as arginine (156.10111 Da). In these cases, the open-search identified peptides were one amino acid shorter on the N-termini while bearing a PTM with a monoisotopic mass close to the missing amino acid. We believe that this conflict stems from an incomplete fragmentation pattern, in which the missing b1 and/or y(n) ions in the MS/MS spectrum leave the search engine with an equally fit decision to match it with the PTM- or non-PTM-bearing sequence.

To further validate our results and assess error rates, we employed a confirmatory procedure by evaluating the identification rate of cysteine carbamidomethylation in samples that were not treated with iodoacetamide. It is important to note that carbamidomethylation is a

deliberate PTM introduced to cysteine residues through a reaction with iodoacetamide; thus, samples that have not undergone iodoacetamide treatment should not exhibit cysteine carbamidomethylation. To minimize false identifications, we applied stringent filters, including a global false discovery rate (FDR) and false localization rate (FLR) of 1%. This means that one would expect a 1% false identification rate, and approximately 1% for each group of PTMs. Our findings revealed that the samples lacking iodoacetamide treatment incorporated 1.75% of peptide-spectrum matches with cysteine carbamidomethylation, which we consider reasonable. It is important to note that this percentage represents a group FDR rather than a global FDR. As such, each PTM group would theoretically have a group FDR of approximately 1%, which would balance out to a global FDR value of 1% when considering all PTM groups together.

De novo quality control

Our sequential *de novo* strategy showed that 96.3% of the MS spectra were canonical (*i.e.*, within known proteins) and a minority (3.7%) were non-canonical (*i.e.*, mapping to the 3-frame translation database).

Validation 2: We assessed the quality of the *de novo* sequences by examining their DeepNovoV2 scores. Canonical and non-canonical peptides had similar *de novo* score distributions, with a slight shift toward higher scores for non-canonical peptides.

Validation 3: We assessed the quality of the *de novo* sequences by examining the correlation between their experimental and theoretical liquid chromatography retention times. Canonical and non-canonical *de novo* sequences had a high correlation, with an R^2 score of 0.9 for *de novo* canonical and 0.863 for *de novo* non-canonical peptides in a melanoma sample (mel-15 from PXD004894), and an overall *de novo* non-canonical R^2 score of 0.88 among all samples.

Second-round search quality control

Validation 4: The results of the second and third validations showed strong evidence that the *de novo* non-canonical peptides were of high quality (*i.e.*, correctly predicted complete peptide sequences). Even with this strong evidence, it is possible that chromatic behavior remains unchanged in certain instances where neighboring amino acids are in flipped positions, or that a 90% accuracy rate still leads to an uncertain false discovery rate percentage. Hence, we confirmed the identified 10,413 *de novo*-based ncMAPs by performing a second-round search for additional validation and controlling the FDR at 1%. The second-round search recovered 7,029 of the 10,413 *de novo*-based ncMAPs, with 76.52% (5,379) recovered from the same

spectra (at least one spectrum per peptide). Overall, the second-round search identified 8,601 ncMAPs with a subset of 1,572 ladder sequences (subsequences) after cleavage of the 10,413 *de novo*-based ncMAPs by the search engine. As for post-translationally modified peptides, the second-round search recovered 51.85% of N-terminal acetylated peptides, 27.96% of peptides with cysteine carbamidomethylation, 74.75% of peptides with cysteinylolation, and 71.02% of peptides with oxidized methionine from the same spectra (at least one spectrum per peptide). The low recovery of carbamidomethylation was mostly due to incorrect open-search assignments in iodoacetamide-untreated samples, considering that 81.01% was recovered by the second-round search in iodoacetamide-treated samples.

Validation 5: We confirmed that post-translationally modified peptides from the second-round search exhibited a shift in retention time that was consistent with that of their unmodified counterparts. Furthermore, we observed that for a specific PTM, there was a similar shift in retention time between non-canonical and canonical MHC-associated peptides. In each case, the modification caused the retention times of PTM-bearing non-canonical MHC-associated peptides to deviate in the same direction relative to the unmodified peptides. We found a high degree of agreement in retention time shifts between canonical and non-canonical peptides for three PTMs: carbamidomethylation, cysteinylolation, and methionine oxidation. For N-terminal acetylation, the quantile ranges (Q1-Q3) were shifted between the two categories. However, it is important to note that the non-canonical category still fell within the standard range of the canonical category, which was mostly due to the low number of identified non-canonical N-terminal acetylated peptides with unmodified counterparts (9) compared to the large number in the canonical group (426).

Validation 6: We checked the mass difference between the observed and calculated masses (i.e., theoretical mass) of the peptide-spectrum matches (PSMs). We isolated the PSMs identified by the *de novo* strategy as well as those validated by second-round search. A similar distribution of mass differences between the *de novo* identified peptides (from -0.0014 to 0.0013 mass (M) / charge (Z)) and the second-round search validated from -0.0012 to 0.0012 M/Z) was observed.

Validation 7: We performed a comprehensive comparison between the PSMs obtained from our second-round search and those reported in the original studies. We hypothesized that if the non-canonical peptides were correctly identified, they would not have been recognized by the original studies that focused on detecting only canonical peptides originating from the proteome. Our analysis showed a remarkable correlation with our hypothesis, as 98.87% (9,495,747 of 9,508,165) of non-canonical PSMs were not detected in the original studies.

HLA-Glyco: A large-scale interrogation of the glycosylated immunopeptidome

Georges Bedran^{1,2,+}, Daniel A. Polasky^{2,+}, Yi Hsiao³, Fengchao Yu², Felipe da Veiga Leprevost², Javier A. Alfaro^{1,4}, Marcin Cieslik^{2,3,5}, Alexey I. Nesvizhskii^{2,3,5*}

¹ International Centre for Cancer Vaccine Science, University of Gdansk, Gdansk, Poland

² Department of Pathology, University of Michigan, Ann Arbor, MI, USA

³ Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, USA

⁴ Department of Biochemistry and Microbiology, University of Victoria, Victoria, Canada

⁵ Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI, USA

+ Co-first authors contributed equally

* Correspondence should be addressed to: nesvi@med.umich.edu

The authors declare no potential conflicts of interest.

Abstract

MHC-associated peptides (MAPs) bearing post-translational modifications (PTMs) have raised intriguing questions regarding their attractiveness for targeted therapies. Here, we developed a novel computational glyco-immunopeptidomics workflow that integrates the ultrafast glycopeptide search of MSFragger with a glycopeptide-focused false discovery rate (FDR) control. We performed a harmonized analysis of 8 large-scale publicly available studies and found that glycosylated MAPs are predominantly presented by the MHC class II. We created HLA-Glyco, a resource containing over 3,400 human leukocyte antigen (HLA) class II N-glycopeptides from 1,049 distinct protein glycosylation sites. Our comprehensive resource reveals high levels of truncated glycans, conserved HLA-binding cores, and differences in glycosylation positional specificity between classical HLA class II allele groups. To support the nascent field of glyco-immunopeptidomics, we include the optimized workflow in the FragPipe suite and provide HLA-Glyco as a free web resource.

Introduction

Protein glycosylation has been extensively studied and found to play a variety of biological roles, including antigen recognition, host-pathogen interactions, and immune modulation¹. Glycosylation causes dramatic alterations in response to cancer and has been suggested as a potential biomarker²⁻⁵. Moreover, glycosylation could be an attractive source of tumor-specific antigens, considering the viability of post-translational modifications (PTMs) on MHC-associated peptides⁶⁻⁹ (MAPs). Critically, glycosylation has been reported to have a significant impact on the immunogenic properties of MAPs in terms of T-cell recognition¹⁰⁻¹² and epitope generation due to interference with the proteolytic cleavage¹³.

High-throughput identification of glycosylated MAPs from mass spectrometry (MS) data involves combining two notoriously challenging problems in computational proteomics. First, the proteolytic processing of MAPs requires non-enzymatic searches (*i.e.*, non-specific cleavage of proteins at every peptide bond). Considering all possible cleavages of reference proteins results in an enormous search space of candidate sequences. Second, the non-templated nature of the glycosylation process results in hundreds of distinct glycans that can be detected across the proteome¹⁴. A combinatorial explosion thus takes place when considering all possible non-enzymatic peptide sequences with many possible glycans. As a result, a non-specific glycopeptide search is not feasible with many search engines due to prohibitively long run times and/or insufficient sensitivity. To the best of our knowledge, very few glycosylation analyses of MAPs have been performed. One of the earliest successful identifications of glycosylated class II MAPs was made in 2005¹⁵ with 2 N-linked glycopeptides found in an EBV-transformed human B-lymphoblastoid cell line. In 2017, Malaker *et al.* successfully identified 26 glycosites in 3 different melanoma cell lines⁹. Both studies required identification of glycopeptides by manual annotation of the spectra. More recently, a third effort from 2021 captured 209 unique human leukocyte antigen (HLA) II-bound peptide sequences from the SARS-CoV-2 virus¹⁶ using an automated glycopeptide search method assisted with a manual verification of all glycopeptide spectra.

Large-scale analysis of glycosylated MAPs requires automated methods with exceptional speed and accuracy to handle the enormous search space of glycosylated non-specific peptides. The above-mentioned challenges have been tackled by our recent developments to improve the search speed¹⁷ (MSFragger) and address the complexity of glycosylation¹⁸ (MSFragger-Glyco). Building on these advances, we developed an optimized workflow for HLA glyco searches with a focus on optimizing the false discovery rate (FDR) control of glycosylated MAPs. We assembled, carefully annotated, and analyzed 8 publicly available immunopeptidomic datasets for N-glycosylation using our workflow and investigated the glycosylated MAPs binding properties. From nearly 2,000 LC-MS/MS runs, we found 3409 class II N-glycosylated MAPs on 1049 distinct protein glycosylation sites of 677 unique proteins. We revealed characteristics of HLA glycopeptides, including high levels of truncated glycans, conserved HLA-binding cores across the 72 studied HLA class II alleles, and a different glycosylation positional specificity between the classical allele groups.

Induced expression and antigen-presentation by the MHC class II on tumor cells is increasingly being recognized as a mediator of anti-tumor immunity and neoantigen efficacy^{19–24}. Our results, made readily accessible as a free web resource, significantly expand our understanding of glyco-MAPs in cancer; and together with our novel optimized workflow, are expected to further the development and discoveries in the nascent field of glyco-immunopeptidomics.

Results

Computational glyco-immunopeptidomics workflow

The computational workflow developed in this work for the analysis of glycosylated MAPs is illustrated in **Fig. 1**. While O-glycosylated MAPs are also of potential interest²⁵, O-glycopeptide analysis typically requires electron-based activation to locate the glycosite(s) within the

peptide. As the vast majority of available immunopeptidomics data lacks such activation, we focused exclusively on N-glycosylated MAPs for this analysis. Briefly, MSFragger-Glyco performs N-glycosylation motif checks for the N-X-S/T consensus sequence, which serves as the attachment site for polysaccharides (*i.e.*, sequon). Simultaneously, spectra are checked for the presence of fragmentation products of peptide-conjugated glycans (*i.e.*, oxonium ions). The glycan search is only performed for peptides with a sequon and for spectra containing oxonium ions above a relative intensity threshold (10% in this case). A regular search is performed for all other spectra. Next, we use PeptideProphet²⁶ and ProteinProphet²⁷ within the Philosopher²⁸ toolkit to model and filter false discovery rates (FDR) to 1% for peptide-spectrum matches (PSMs), peptides, and proteins, respectively. As in previous glycopeptide analyses, we applied the extended mass model of PeptideProphet to simultaneously model the score and mass-shift distributions of the database search¹⁷. This provides a separate probability model for different glycan masses (*i.e.*, mass shifts) to account for the varying frequencies of the different glycans¹⁸.

Initially, we assessed the standard FDR procedures used for enzymatically digested and enriched glycopeptides on non-enzymatic unenriched immuno-glycopeptides. We observed that while 91% of the glycoPSMs corresponded to known glycosylation sites, less than half of the observed glycosites (46%) were previously known (**Supplementary Fig. 1a**). Thus, known sites tended to have many supporting spectra, while unknown sites had few and notably lower scores, likely indicating an unacceptable increase in false discoveries. Since glycoPSMs represent a small fraction of the identified spectra, the score thresholds used in our initial FDR filtering were mostly influenced by non-glycosylated peptides. As glycopeptides have a much larger search space, this results in an enrichment of false discoveries in the glycopeptide fraction when all PSMs are filtered together. To counter this, we applied a separate PeptideProphet probability (*i.e.*, score) filter for glycosylated and non-glycosylated PSMs to control FDR in each category despite the differences in search space, using a modified version of Philosopher (see **Methods** and **Supplementary Fig. 1b**). We further filtered glyco-PSMs

by glycan q-value ($q \leq 0.05$) to remove glycopeptides lacking sufficient evidence supporting the glycan composition assignment²⁹ by PTM-Shepherd³⁰. With this improved filtering method, the proportion of PSMs corresponding to known glycosites increased to 96%, and the proportion of identified glycosites corresponding to known glycoproteins increased to 95%, with 79% of sites previously identified in other glycoproteomic analyses (**Supplementary Fig. 1c**). These stringent glycopeptide-specific filters provide effective FDR control in a challenging search, allowing for confident construction of the HLA glycopeptide resources.

Large multi-tissue MHC immunopeptidome dataset

We selected 8 immunopeptidomic studies^{31–38}, prioritizing studies with a large amount of high-resolution mass spectrometry data and included a variety of instruments as a means to reduce instrumental bias (see **Methods**). Based on our careful curation and annotation of these data, our collection of 732 different HLA class II mass spectrometry samples incorporated 90.8% of HLA-typed data (**Fig. 2a**), 80.3% of patient tissues, 16.7% of cell lines, and 2.9% of tumor-infiltrating lymphocytes (**Fig. 2b**). The previously mentioned sample types covered up to 6 different cancers (**Fig. 2c**) located in the brain (meningioma and glioblastoma), skin (melanoma), colon (colorectal), and lung (adenocarcinoma and squamous carcinoma). In addition, 59% of the samples are non-cancerous and come from disease-free individuals. In terms of HLA diversity, up to 72 HLA class II alleles of the 3 classic genes (DP, DQ, and DR) are covered by varying numbers of mass spectrometry samples (**Fig. 2d**).

Leveraging the wealth of proteomic data, we queried the glycosites identified in our study against previously reported glycosylation sites in GlyGen³⁹. PSM level information showed 96.4% of previously reported glycosylation sites (**Fig. 2e**), 1.8% of glycosylation sites within previously reported glycosylated proteins, and 1.8% of new glycosylation sites. On the other hand, at the peptide level, 90% of glycopeptides mapped to previously reported glycosylation sites, 6.7% of glycopeptides were within previously reported glycosylated proteins, and 3.3% contained new glycosylation sites. A similar trend was observed at the glycosylation site level,

with 78.8% of previously reported glycosylation sites, 15.6% of glycosylation sites within previously reported glycosylated proteins, and 5.5% of new glycosylation sites. It appears that peptides containing previously reported glycosylation sites are abundant species, considering the high spectral count (**Fig. 2f** in gray) in comparison with the previously unreported ones (**Fig. 2f** in blue and black). We then benchmarked our findings against previous work by Malaker *et al.* 2017⁹ on glycosylated MAPs in 3 melanoma and 1 EBV-transformed B-cell lines. The original manuscript reported 93 glycosylated peptides corresponding to 26 glycosylation sites, split between N-glycosylation (23) and O-glycosylation (3). Our workflow recovered 20 of the 23 identified N-glycosylation sites, of which 4 did not pass the FDR filter. With a 45-fold increase in glycosylation sites, we identified 1033 new sites (see **Fig. 2g**).

Enrichment of N-glycosylation in the class II immunopeptidome

Several of the datasets we searched contained both HLA class I and II peptides from the same samples and, in one case, whole proteome data, allowing us to compare the frequency and characteristics of glycosylation across these categories. Fragmentation of glycopeptides by tandem MS (MS/MS) produces highly abundant oxonium ions resulting from the fragmentation of conjugated glycan(s), which can provide an estimate of the fraction of glycopeptides in a sample prior to a database search. To understand the abundance of glycosylation at different molecular levels, we compared the percentage of oxonium-containing MS/MS scans for the 4 datasets containing multiple HLA classes (**Fig. 3a**). Interestingly, datasets A³¹ (Bassani-Sternberg *et al.* 2016), B³⁴ (Chong *et al.* 2020), and D³⁷ (Forlani *et al.* 2021) showed, on average, an approximate 5-fold enrichment in potential HLA class II glycosylation events compared with HLA class I data. In dataset C³² (Marcu *et al.* 2021), the only dataset containing samples derived from healthy tissue, a similar proportion of oxonium-containing scans was observed in the HLA class II data as in the other datasets, but there were essentially no oxonium-containing scans in the HLA class I data. As expected, the percentage of glycosylated PSMs obtained from database searches of these datasets followed a similar trend, with 0.5 to 3% of observed PSMs glycosylated in HLA class II data versus less than

0.1% glycosylated in HLA class I data (datasets A, B, and C). Strikingly, glycosylated PSMs were also enriched approximately 7-fold in HLA class II compared with the whole proteome data in dataset D (**Fig. 3b**), a dramatic increase given the abundance of glycosylation in the proteome.

We also noticed that the composition of glycans observed in the immunopeptidomic datasets was different from that of their proteome counterparts. (**Fig. 3c**). The average glycan mass detected in the immunopeptidome was approximately 1000 Da, which was significantly lower than that observed in the proteome (1400 Da average). To further explore the nature of this compositional discrepancy, we compared glycan types between the two groups (**Fig. 3d**). A higher percentage of truncated glycans (68%) was observed in the HLA class II immunopeptidome compared to the more typical high-mannose and complex/hybrid categories in the proteome, as noted in a previous analysis⁹. This trend of truncated glycans on HLA peptides was preserved when only glycans from the same protein were considered. For example, LRP1, a highly glycosylated protein, was observed with a mix of high-mannose and complex glycans in the proteome sample, but with a mix of truncated and high-mannose glycans in the HLA-II sample with almost no mature complex glycans detected (**Fig. 3e**). There was very little overlap between the glycosylated proteins and sites in each category, with only 22.8% of HLA-II glycoproteins observed in the whole proteome data and even lower overlap (16.3%) when considering the specific glycosylation sites within proteins. (**Fig. 3f**). The whole proteome glyco search likely captures glycopeptides from the most abundant glycoproteins, as the experiment was performed without any glycopeptide enrichment, whereas the immunopeptide datasets presumably capture MAPs with much less dependence on overall protein abundance.

Overall, the data showed a remarkable enrichment of glycosylation in HLA class II-associated peptides relative to HLA class I and the whole proteome, leading us to focus the remainder of our efforts on HLA class II-associated and glycosylated peptides.

Glycosylation of MAPs does not influence the HLA binding motif

To explore glycosylation in the context of HLA class II presentation, we focused on the HLA-binding core, a 9-mer sequence that interacts with the HLA molecule. In most mass spectrometry experiments, samples express multiple HLA alleles, leading to an ambiguous association between the identified peptides and the pool of available HLA molecules. Hence, a deconvolution step to find the HLA motifs and the corresponding binding core offsets of each peptide was deemed necessary for further experimentation (see **Methods**).

Deconvolution of peptides using a semi-supervised approach

We first chose to use MoDec³⁸ for deconvolution, a fully probabilistic framework that learns both the motifs and preferred binding core position offsets from the sequences themselves. The fact that MoDec does not rely on a pre-trained model is crucial when exploring HLA-bound peptides with post-translational modifications (*i.e.*, glycosylation) to avoid the removal of all peptides that were not well modeled. Such a deconvolution strategy requires manual intervention to choose the number of HLA motifs (*i.e.*, number of clusters) and assign each discovered motif to one of the expressed HLA alleles of a given sample. We carefully selected a case study on a human B lymphoblastoid cell line (C1R) from Ramarathinam *et al.* 2021³⁶. The purification protocol of the HLA-bound peptides in this study was performed sequentially with pan anti-class I, followed by class II anti-DP (**Fig. 4a**), class II anti-DQ (**Fig. 4b**), and class II anti-DR antibodies (**Fig. 4c and d**). Hence, the resulting mass spectrometry samples were mono-allelic (*i.e.*, presenting one allele at a time), except for the DR samples with the DRB1*12:01 and DRB3*02:02 alleles eluting together. **Fig. 4** presents 4 sections **a**, **b**, **c**, and **d** standing for the HLA class II alleles DPA1*02:01/02-DPB1*04:01, DQA1*05:05-DQB1*03:01, DRB1*12:01, and DRB3*02:02, respectively. All alleles showed a similar percentage of glycosylated and non-glycosylated peptides with the corresponding HLA motifs after deconvolution (**Fig. 4, panel I**). All 25 replicates showed an unaltered HLA-binding core with glycosylation (two-sided Fisher's exact test, 25 P-values > 0.05).

Considering the concordance of glycopeptide sequences with the HLA-binding cores, we checked the absolute glycosylation position per peptide length (*i.e.*, glycosylation offset within the peptide). **Fig. 4** panel **II** shows a glycosylation tendency towards the N- and C-termini for both DQ and DR alleles (**Fig. 4** sections **b**, **c**, and **d** at panel **II**) and only the C-terminal tendency for the DP allele (**Fig. 4** section **a** at panel **II**). To further decipher glycosylation in the context of the HLA-binding cores, we looked at the relative position shown in **Fig. 4** panel **III** (*i.e.*, glycosylation offset from the HLA-binding core start). Negative values indicate sites upstream of the HLA-binding motif start, 0 to 8 values reference positions within the HLA-binding core, and values greater than 8 denote glycosylation sites downstream of the HLA-binding core. For the DPA1*02:01/02-DPB1*04:01 allele, glycosylation occurred 91% of the time within the HLA motif at position 8 (**Fig. 4** section **a** at panel **III**). In contrast, for the other 3 alleles, glycosylation was more likely (86% of the time) to take place up- or downstream of the HLA-binding core.

Deconvolution of peptides using a fully unsupervised approach

Despite the usefulness of MoDec for a previously unexplored category of peptides, such a tool suffers from several limitations^{40,41}: (I) the need for manual intervention to associate the identified motifs with known allele specificities present in the sample; (II) the difficulty of assigning peptides to MHC molecules when alleles with overlapping motifs are co-expressed; (III) low sensitivity with low expression of MHC molecules; and (IV) the complexity of HLA class II specificities due to the involvement of the variable alpha and beta chains for the HLA-DQ and HLA-DP groups. All these, render motif-allele assignment a daunting task, especially with up to 87 subjects in our dataset. Thus, we used the state-of-the-art binding model NetMHCIIpan 4.1^{41,42} to perform MHC motif deconvolution and assign glycopeptide sequences to their most likely HLA alleles without the need for manual intervention (see **Methods**). Consistently, glycosylated and non-glycosylated peptides from Ramarathinam *et al.* 2021 showed similar binding properties, indicating that the detected glycosylation fit within the known HLA-binding cores (two-tailed Fisher's exact test, P-value: 0.48). Interestingly,

NetMHCIIpan 4.1 confirmed most peptides with glycosylation located at P8 within the HLA-binding core (97% for DPA1*0201 and 100% for DPA1*0202) for the C1R DP allele (**Fig. 5a**). Overall, 95%, 83%, 76%, and 87% of glycopeptides were found to bind to C1R DP (**Fig. 5a**), DQ (**Fig. 5b**), DRB1*12:01 (**Fig. 5c**), and DRB3*02:02 (**Fig.5d**), respectively. Hence, we carried out the NetMHCIIpan 4.1 deconvolution for the 83 remaining subjects in our dataset.

The HLA class II N-glycosylation characteristics

We noticed a high tendency of glycosylation within the HLA-binding core for HLA DP alleles, followed by a lower tendency for HLA DQ, and even lower one for HLA DR alleles. Hence, we checked for the occurrence of such events for each of the 3 HLA groups (DP, DQ, and DR). **Fig. 6a** shows that up to 57% of HLA DP associated peptides have glycosylation inside the HLA-binding core, 30% for HLA DP, and 13% for HLA DR. In terms of glycan types, **Fig. 6b** shows that HLA DP associated peptides showed the highest fraction (0.67) of truncated glycans compared to DQ (0.55) and DR (0.41). High-mannose glycans showed a reverse trend for DR, DQ, and DP alleles, with fractions of 0.37, 0.27, 0.21, respectively. All DP, DQ, and DR associated peptides showed a depletion in complex/hybrid glycans in accordance with previous findings^{9,16}.

Discussion

Post-translational modifications increase the diversity of the immunopeptidome and may provide new targets for the immune system to recognize tumor cells or respond to pathogens. With PTM-driven antigenicity being continuously highlighted^{9,31,43,44}, glycosylation is a key PTM that, despite its long history of research, remains understudied in the context of MHC presentation due to computational related challenges. In this work, we have developed a workflow for glyco-immunopeptidomics that combines the speed and sensitivity of MSFragger-Glyco, with the inclusion of glycopeptide-specific FDR control in Philosopher, which is critical for filtering out low-confidence identifications. We used this workflow to produce a resource of HLA class II N-glycosylated MAPs arising from a harmonized analysis of 8 publicly available studies. Overall, we identified 1049 glycosylation sites from 3409 different glycopeptides, an order of magnitude greater than any previous effort in this area. Leveraging this large-scale resource, we explored the properties of glycosylated MAPs, including the types of glycans conjugated, MHC binding affinity predictions, and the positioning of glycosylation relative to the HLA binding core. Interestingly, we observed no difference in binding motif predictions with glycopeptides compared to non-glycopeptides, despite some peptides containing glycans within the binding core. HLA DP alleles presented a majority of glycans within the binding core (57%) compared with HLA DQ alleles (30%) and HLA DR alleles (13%). Moreover, we found a difference in the glycan types between HLA groups (DP, DR, and DQ), with truncated glycans enriched for DP alleles and a higher mannose content for DR alleles.

A study by Malaker *et al.*⁹ on HLA class II N-glycosylation covered 5 DR alleles (DRB1*0101, DRB1*0401, DRB1*0404, DRB1*1502, DRB4*0103) and showed that 3 out of 23 peptides had glycosylated residues within the binding core. In combination with molecular modeling, this allowed the authors to postulate that glycan residues are most likely to protrude out of the HLA-binding pocket and interact with the complementary determinant region of the T-cell receptor. Our findings expand the coverage to 28 DR alleles, along with multiple DP and DQ alleles, adding up to 87 HLA molecules overall, when considering the combination of alpha

and beta chains. In addition to the preference of terminal glycosylation for peptides associated with DR and DQ alleles, we observed an HLA-binding core glycosylation tendency for peptides associated with DP alleles. Future studies should explore whether the correlation between smaller glycans and presence within the HLA-binding core is related simply to size restrictions preventing larger glycans from occupying the core or is a reflection of other processing of MAPs for presentation.

The enrichment of glycosylated peptides on the MHC-II, while preserving canonical binding motifs, offers the tantalizing possibility of designing and developing glycosylated neoantigen vaccines with improved affinity over wild-type peptides^{22,23}. Which is further notable, in light that most of the known anti-tumor CD4+ T cells are specific for highly immunogenic self-derived MHC-II antigens, demonstrating that self-antigen CD4+ T cells can mount anti-tumor responses. Cancer-specific glycosylation of MAPs may further contribute to the restriction of those mechanisms to the tumor microenvironment. We made our findings readily available as a web resource to query pertinent information about the identified glycosylated MAPs. Users can search for a specific glycan and/or MAP sequence, protein, or glycosylation site associated with a specific HLA allele. In addition, we included deconvolution information allowing further interpretation of the data within the HLA haplotype context. We are planning to grow this initiative, introduce more studies, and increase the HLA allele coverage. Moreover, by providing the optimized computational workflow file, which can be loaded directly into FragPipe to reproduce the method described here, we make it easy for others to carry out challenging glyco-immunopeptidomics analyses on new datasets. It is our hope that the method and findings presented here will expand the field of tumor-specific antigen discovery, broaden the scope of possible antigens to target, and improve strategies for vaccine design. O-glycosylated MAPs, for example, represent another potential class of antigens that can, in principle, be studied by our method for further exploration⁴⁵. Finally, given the promising nature of glycosylated MAPs, we anticipate the attraction of glycosylation-oriented research towards the immunopeptidomics field.

Methods

Dataset selection

Studies from the PRIDE⁴⁶ database were first screened based on a list of keywords related to immunopeptidomics. Next, low-resolution analyses were eliminated, and MHC-related datasets conducted with at least one of the following instruments were kept: Orbitrap Lumos/Fusion, Q Exactive, LTQ Orbitrap, Orbitrap Exploris 480, TripleTOF, impact II, and maXis. Then, manual curation of the resulting 312 studies was performed to filter non-relevant datasets, resulting in 140 HLA Class I, II, or I & II datasets. The number of identified proteins per study was retrieved from gpmDB⁴⁷ and datasets with a high number of protein groups were prioritized. A final manual curation step resulted in the selection of the 8 datasets included in this study.

Mass spectrometry N-glycan search

Raw and wiff files were first downloaded from PRIDE and converted to mzML format using msconvert⁴⁸ with peak picking, FragPipe (TPP) compatibility, and removal of zero values filters. The analysis was executed within the FragPipe suite v18.1-build5 using headless mode. Glyco-searches were performed using MSFragger v3.5 with methionine oxidation, N-terminal acetylation, and cysteinylolation as variable modifications, and a list of 198 glycans. A list of contaminants was added to the UniProt Swiss-Prot (UP000005640) proteins⁴⁹, along with their corresponding reversed decoy sequences. Enzymatic cleavage was set to non-specific with peptide lengths from 7 to 25 amino acids for the 8 HLA class II datasets, from 7 to 12 amino acids for HLA class I datasets (A, B, C, D), and fully enzymatic cleavage with peptide lengths from 7 to 50 amino acids for the whole proteome dataset D. Peptides containing the consensus sequon (N-X-S/T) and decoy (reversed) peptides containing the reversed sequon were considered as potential glycopeptides to ensure the that same number of potential glycopeptides was searched in both target and decoy databases. Only spectra containing oxonium ion peaks with summed intensity of at least 10% of the base peak were

considered for glycan searches, while all others were searched without considering glycosylation. Data were deisotoped⁵⁰ and decharged in MSFragger-Glyco, calibrated, and searched with 20 ppm mass tolerances for precursors and 15 ppm for products with MSFragger's built-in parameter optimization performed for each study⁵¹. Errors in monoisotopic peak detection by the instrument were allowed (+1 and +2 Da).

FDR control

Filtering was performed using Philosopher²⁸ (v4.5.1-RC10), including PeptideProphet modeling of peptide probabilities, ProteinProphet protein inference, and Philosopher's internal filter for FDR control. The semi-parametric modeling of PeptideProphet was used with the expectation value as the only contributor to the f-value. The number of tolerable termini (ntt) model was disabled, as it is not applicable to non-enzymatic searches. Filtering was performed in Philosopher using a modified, group-specific FDR procedure. Non-glycosylated and glycosylated PSMs were filtered separately, using a delta mass cutoff of 145 Da (the size of the smallest glycan considered in the search) to distinguish glycosylated PSMs from non-glycosylated PSMs. This allowed different score thresholds to be used to filter glycosylated and non-glycosylated PSMs to 1% FDR. This is essential as the large search space for glycosylated PSMs results in higher scoring false matches, requiring a higher score threshold for effective filtering than for non-glycosylated PSMs. Since non-glycosylated PSMs make up the majority of the results, filtering all PSMs together would yield an insufficiently low score threshold for glycosylated PSMs. After the group-specific 1% FDR filter was applied to glycosylated and non-glycosylated PSMs, 1% peptide- and protein-level FDR filters were applied. A sequential filtering step was then applied to remove any PSMs matched to proteins that did not pass the 1% protein-level FDR. Glycan assignment was subsequently performed in PTM-Shepherd using the default N-glycan database²⁹ and parameters along with a 0.05 glycan q-value threshold.

Deconvolution of the MHC associated peptides

Motif deconvolution is the process of finding HLA-binding motifs and their corresponding binding core offsets for a set of peptides. A first deconvolution that required manual inspection was performed using MoDec³⁸. The peptides were grouped by subject (*i.e.*, instances of the same replicates). A maximum of 10 clusters, 20 runs, and a minimum peptide length of 12 amino acids were considered. Since HLA-II ligands from the same subject come from different alleles, MoDec provides a direct interpretation and assigns peptides with similar binding cores to clusters (*i.e.*, HLA motifs). However, manual inspection is still required to (I) the number HLA motifs MoDec detected per subject and (II) annotate these motifs (*i.e.*, clusters) to their respective HLA II alleles. Hence, the MoDec-identified HLA motifs were assigned to the correct HLA class II alleles by manual inspection for each analyzed subject. A second deconvolution that didn't require manual inspection, inspired from Kaabinejadian *et al.* 2022⁴¹, was performed using NetMHCIIpan 4.1⁴². Briefly, all unique peptides were predicted for MHC presentation towards all the MHC alleles expressed in the given subject. The likelihood of peptides being presented by a given MHC molecule is given by the percentile rank score, which ranges from 0 to 100, with 0 being the strongest binding score. Peptides with a percentile rank score > 20 were considered MS co-immunoprecipitated contaminants and labeled as trash. Peptides with a percentile rank score ≤ 20 were assigned to the lowest scoring allele of a given subject. We applied the second deconvolution method using NetMHCIIpan 4.1 to the entirety of the subjects in this study, considering the similarity of the results to the first deconvolution method (*i.e.*, MoDec).

Figure generation

Motif plots were generated using the Python library Logomaker⁵², heatmaps using seaborn⁵³ and other plots using matplotlib⁵⁴.

Authorship contribution

G.B. collected and curated the data, generated the figures and supplementary materials, and drafted and coordinated the manuscript. D.P. performed the immunopeptidomics analysis, supported figure generation, interpretation of results, and drafting and coordination of the manuscript. Y.H. produced the web portal and helped to revise the manuscript. F.Y. supported the study with software development related tasks. F.L. supported the study by adding a group-specific FDR feature to Philosopher. J.A. supported with the writing of the manuscript. M.C. helped with the study design and manuscript revision. A.I.N. conceived the project, helped with the study design and revision of the manuscript, and provided overall supervision.

Acknowledgments

This work was funded in part by NIH grants R01-GM-094231, U24-CA210967, U24-CA271037, and by the International Centre for Cancer Vaccine Science, a project carried out within the International Research Agendas programme of the Foundation for Polish Science and co-financed by the European Union under the European Regional Development Fund.

References

1. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).
2. Costa, A. F., Campos, D., Reis, C. A. & Gomes, C. Targeting Glycosylation: A New Road for Cancer Drug Discovery. *Trends Cancer* **6**, 757–766 (2020).
3. Thomas, D., Rathinavel, A. K. & Radhakrishnan, P. Altered glycosylation in cancer: A promising target for biomarkers and therapeutics. *Biochim. Biophys. Acta BBA - Rev. Cancer* **1875**, 188464 (2021).
4. Wang, M., Zhu, J., Lubman, D. M. & Gao, C. Aberrant glycosylation and cancer biomarker discovery: a promising and thorny journey. *Clin. Chem. Lab. Med. CCLM* **57**, 407–416 (2019).

5. Mereiter, S., Balmaña, M., Campos, D., Gomes, J. & Reis, C. A. Glycosylation in the Era of Cancer-Targeted Therapy: Where Are We Heading? *Cancer Cell* **36**, 6–16 (2019).
6. Mangalaparthy, K. K. *et al.* Digging deeper into the immunopeptidome: characterization of post-translationally modified peptides presented by MHC I. *J. Proteins Proteomics* **12**, 151–160 (2021).
7. Mei, S. *et al.* Immunopeptidomic Analysis Reveals That Deamidated HLA-bound Peptides Arise Predominantly from Deglycosylated Precursors. *Mol. Cell. Proteomics* **19**, 1236–1247 (2020).
8. Carra, G. Selective association of a 22–38 kDa glycoprotein with MHC class II DP antigen on activated human lymphocytes at the plasma membrane. *Mol. Immunol.* **33**, 269–278 (1996).
9. Malaker, S. A. *et al.* Identification and Characterization of Complex Glycosylated Peptides Presented by the MHC Class II Processing Pathway in Melanoma. *J. Proteome Res.* **16**, 228–237 (2017).
10. Olvera, A. *et al.* Does Antigen Glycosylation Impact the HIV-Specific T Cell Immunity? *Front. Immunol.* **11**, 573928 (2021).
11. Xu, Y., Sette, A., Sidney, J., Gendler, S. J. & Franco, A. Tumor-associated carbohydrate antigens: A possible avenue for cancer prevention. *Immunol. Cell Biol.* **83**, 440–448 (2005).
12. Housseau, F. *et al.* N-linked carbohydrates in tyrosinase are required for its recognition by human MHC class II-restricted CD4+ T cells. *Eur. J. Immunol.* **31**, 2690–2701 (2001).
13. Li, H. *et al.* Identification of an N-Linked Glycosylation in the C4 Region of HIV-1 Envelope gp120 That Is Critical for Recognition of Neighboring CD4 T Cell Epitopes. *J. Immunol.* **180**, 4011–4021 (2008).

14. Thaysen-Andersen, M., Packer, N. H. & Schulz, B. L. Maturing Glycoproteomics Technologies Provide Unique Structural Insights into the N-glycoproteome and Its Regulation in Health and Disease. *Mol. Cell. Proteomics* **15**, 1773–1790 (2016).
15. Dengjel, J., Rammensee, H.-G. & Stevanovic, S. Glycan side chains on naturally presented MHC class II ligands. *J. Mass Spectrom.* **40**, 100–104 (2005).
16. Parker, R. *et al.* Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. *Cell Rep.* **35**, 109179 (2021).
17. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513 (2017).
18. Polasky, D. A., Yu, F., Teo, G. C. & Nesvizhskii, A. I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **17**, 1125–1132 (2020).
19. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
20. Oliveira, G. *et al.* Landscape of helper and regulatory antitumour CD4+ T cells in melanoma. *Nature* **605**, 532–538 (2022).
21. Axelrod, M. L., Cook, R. S., Johnson, D. B. & Balko, J. M. Biological Consequences of MHC-II Expression by Tumor Cells in Cancer. *Clin. Cancer Res.* **25**, 2392–2402 (2019).
22. Alspach, E. *et al.* MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* **574**, 696–701 (2019).
23. Tay, R. E., Richardson, E. K. & Toh, H. C. Revisiting the role of CD4+ T cells in cancer immunotherapy—new insights into old paradigms. *Cancer Gene Ther.* **28**, 5–17 (2021).

24. Johnson, A. M. *et al.* Cancer Cell–Intrinsic Expression of MHC Class II Regulates the Immune Microenvironment and Response to Anti–PD-1 Therapy in Lung Adenocarcinoma. *J. Immunol.* **204**, 2295–2307 (2020).
25. Mukherjee, S., Sanchez-Bernabeu, A., Demmers, L. C., Wu, W. & Heck, A. J. R. The HLA Ligandome Comprises a Limited Repertoire of O-GlcNAcylated Antigens Preferentially Associated With HLA-B*07:02. *Front. Immunol.* **12**, 796584 (2021).
26. 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).
27. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
28. da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870 (2020).
29. Polasky, D. A., Geiszler, D. J., Yu, F. & Nesvizhskii, A. I. Multiattribute Glycan Identification and FDR Control for Glycoproteomics. *Mol. Cell. Proteomics* **21**, 100205 (2022).
30. Geiszler, D. J. *et al.* PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results. *Mol. Cell. Proteomics* **20**, 100018 (2021).
31. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404 (2016).
32. Marcu, A. *et al.* HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J. Immunother. Cancer* **9**, e002071 (2021).

33. Newey, A. *et al.* Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. *J. Immunother. Cancer* **7**, 309 (2019).
34. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
35. Chong, C. *et al.* High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferon-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. *Mol. Cell. Proteomics* **17**, 533–548 (2018).
36. Ramarathinam, S. H., Ho, B. K., Dudek, N. L. & Purcell, A. W. HLA class II immunopeptidomics reveals that co-inherited HLA-allotypes within an extended haplotype can improve proteome coverage for immunosurveillance. *PROTEOMICS* **21**, 2000160 (2021).
37. Forlani, G. *et al.* CIITA-Transduced Glioblastoma Cells Uncover a Rich Repertoire of Clinically Relevant Tumor-Associated HLA-II Antigens. *Mol. Cell. Proteomics* **20**, 100032 (2021).
38. Racle, J. *et al.* Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).
39. York, W. S. *et al.* GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* **30**, 72–73 (2020).
40. Gfeller, D. & Bassani-Sternberg, M. Predicting Antigen Presentation—What Could We Learn From a Million Peptides? *Front. Immunol.* **9**, 1716 (2018).

41. Kaabinejadian, S. *et al.* Accurate MHC Motif Deconvolution of Immunopeptidomics Data Reveals a Significant Contribution of DRB3, 4 and 5 to the Total DR Immunopeptidome. *Front. Immunol.* **13**, 835454 (2022).
42. 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 Res.* **48**, W449–W454 (2020).
43. Kacen, A. *et al.* Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01464-2.
44. Symonds, P. *et al.* Citrullinated Epitopes Identified on Tumour MHC Class II by Peptide Elution Stimulate Both Regulatory and Th1 Responses and Require Careful Selection for Optimal Anti-Tumour Responses. *Front. Immunol.* **12**, 764462 (2021).
45. Marino, F. *et al.* Extended O-GlcNAc on HLA Class-I-Bound Peptides. *J. Am. Chem. Soc.* **137**, 10922–10925 (2015).
46. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
47. Craig, R., Cortens, J. P. & Beavis, R. C. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *J. Proteome Res.* **3**, 1234–1242 (2004).
48. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
49. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

50. Teo, G. C., Polasky, D. A., Yu, F. & Nesvizhskii, A. I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *J. Proteome Res.* **20**, 498–505 (2021).
51. Yu, F. *et al.* Identification of modified peptides using localization-aware open search. *Nat. Commun.* **11**, 4065 (2020).
52. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
53. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
54. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Figures

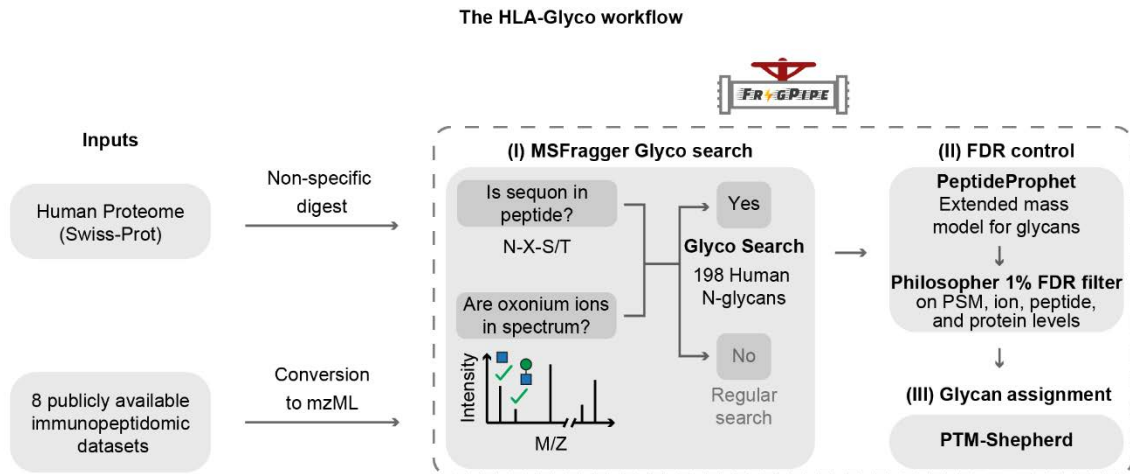


Figure 1: The HLA-Glyco workflow for the detection of glycosylated MHC associated peptides. The FragPipe suite was used to (I) perform a search for glycosylated peptides (glyco search) with the MSFragger search engine; (II) control the FDR with PeptideProphet in combination with a modified version of Philosopher; and (III) assign a glycan composition for each glycopeptide-spectrum match using PTM-shepherd.

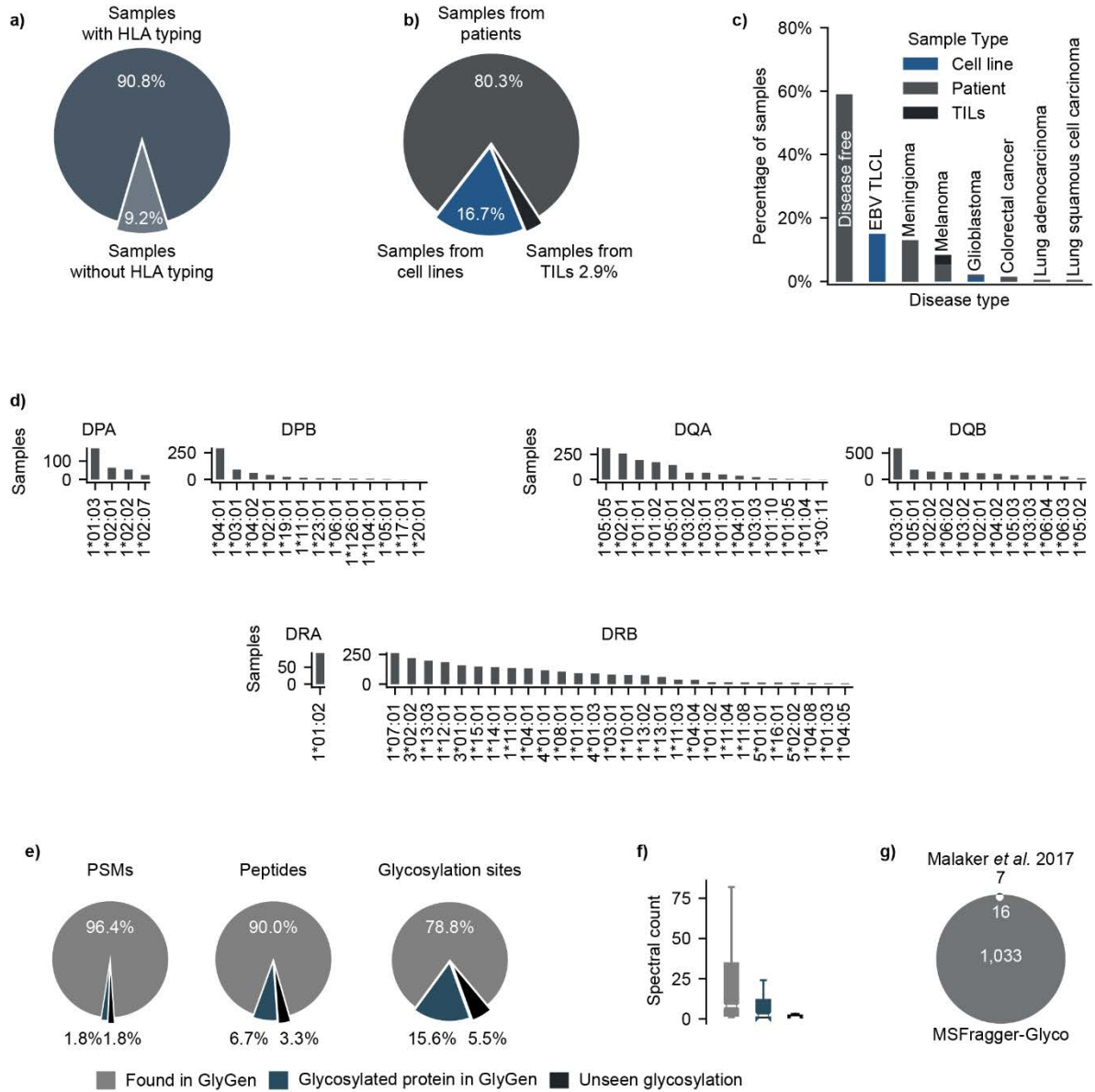


Figure 2: HLA class II infographics of the 8 collected datasets in this study. a) Percentage of samples with HLA class II typing information. **b)** Sample types of the collected mass spectrometry samples (i.e., patient tissues, cell lines, and tumor-infiltrating lymphocytes/TILs). **c)** Cancer types across the collected mass spectrometry samples. **d)** HLA class II alleles (DR, DB, and DQ) across the collected mass spectrometry samples. **e)** Percentage of glyco-PSMs, glycopeptides, and glycosylation sites found in GlyGen. **f)** Abundance of the 3 categories from panel (a) by spectral count. **g)** Comparison of the identified glycosylation sites with Malaker *et al.* 2017 findings.

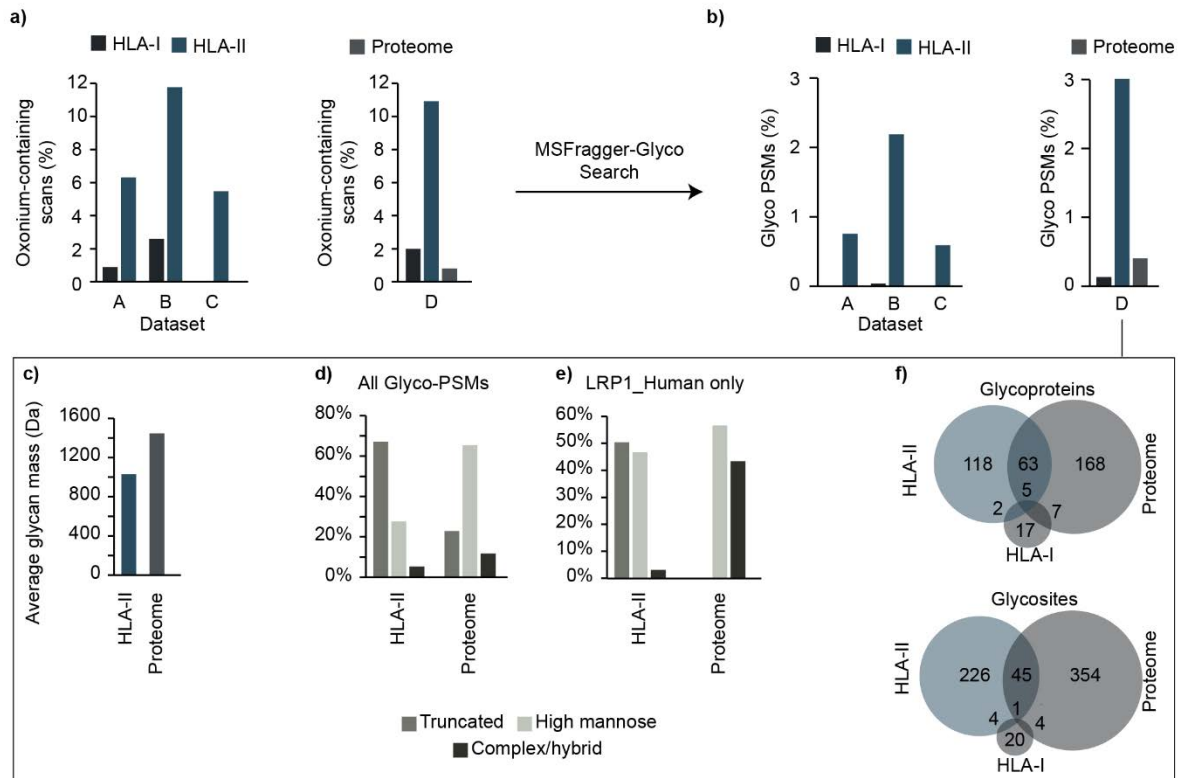


Figure 3: A comparison of the glycosylation on the proteome, HLA I, and HLA II peptidome levels. a) Levels of oxonium ions for HLA class I and II in 3 datasets (A: Bassani-Sternberg *et al.* 2016, B: Chong *et al.* 2020, C: Marcu *et al.* 2021), along with the whole proteome in dataset D: Forlani *et al.* 2021. **b)** Percentage of Glycosylated PSMs for the HLA class I and II immunopeptidome in 3 datasets (A, B, C), along with the whole proteome in dataset D. **c)** Average glycan mass in Dalton (Da) for the HLA class II immunopeptidome versus the whole proteome in dataset D. **d)** Glycan types for the class II immunopeptidome versus whole proteome in dataset D. **e)** Glycan types found in the low-density lipoprotein receptor-related protein 1 (LRP1) for the class II immunopeptidome versus the whole proteome in dataset D. **f)** Comparison of glycoproteins (top) and glycosites (bottom) found in the HLA class I, II immunopeptidome, and whole proteome of dataset D.

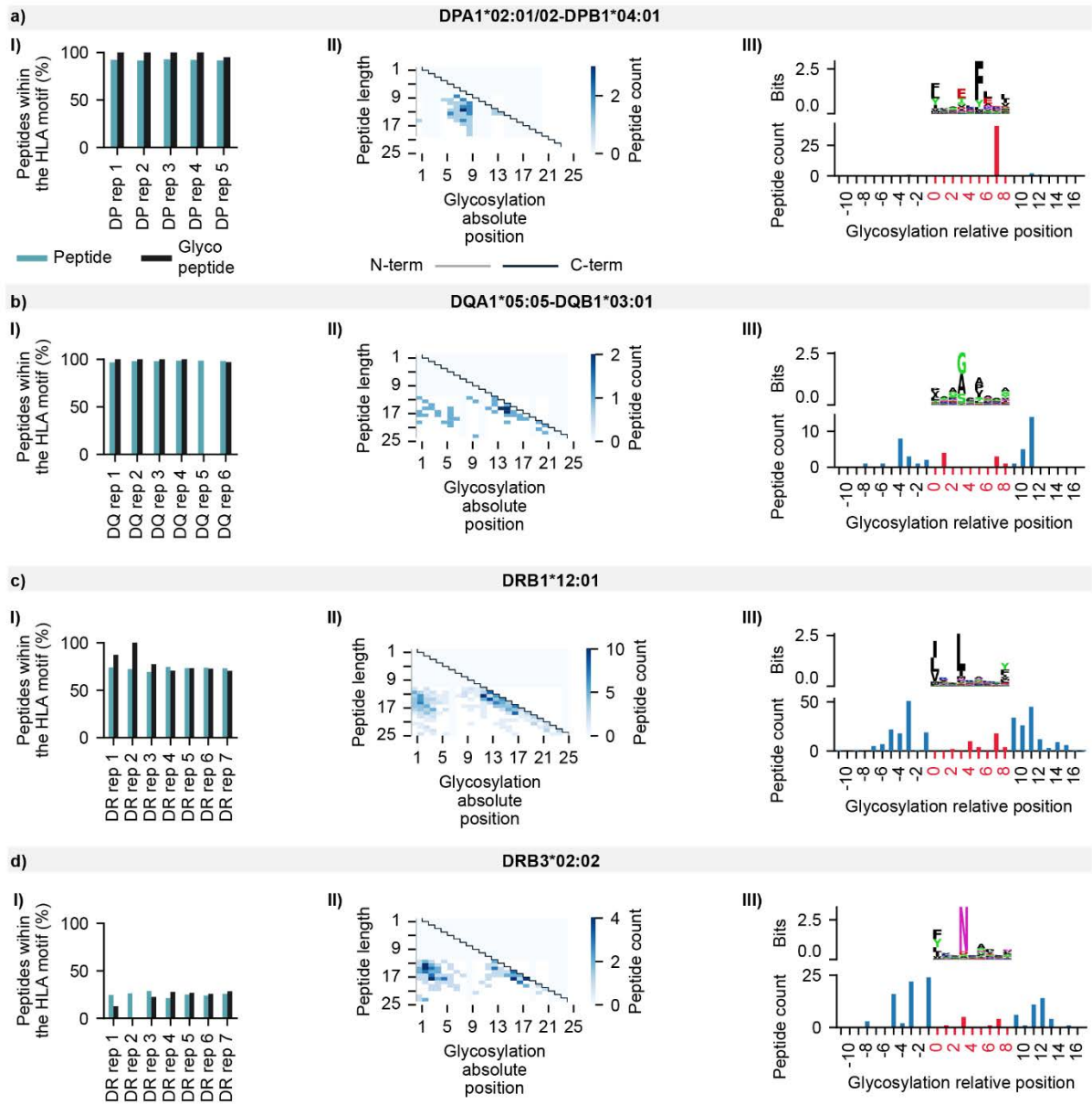


Figure 4: Semi-supervised deconvolution of glycosylated HLA peptides from Ramarathinam *et al.* 2021 using MoDec. Panels I show the percentage of peptides and glycopeptides presenting the HLA binding motif. Panels II display the glycosylation absolute position within the peptidic sequence (x-axis) and the peptide length (y-axis). Gray and black lines indicate the N-term and C-term respectively while the white to blue gradient represents the number of peptides with a specific glycosylation position at a specific peptide length. Panels III present the HLA binding motif after deconvolution with MODEC (top) and the number of glycopeptides per relative glycosylation position (bottom). Negative values refer to glycosylation position upstream the HLA-binding core, values between 0 and 8 represent positions within the HLA-binding core, and values ≥ 9 refer to positions downstream the HLA-binding core. **a) Peptides associated with the HLA allele DPA1*02:01/02-DPB1*04:01 of the C1R cell line. **b)** Peptides associated with the HLA allele DQA1*05:05-DQB1*03:01 of the C1R cell line. **c)** Peptides associated with the HLA allele DRB1*12:01 of the C1R cell line. **d)** Peptides associated with the HLA allele DRB3*02:02 of the C1R cell line.**

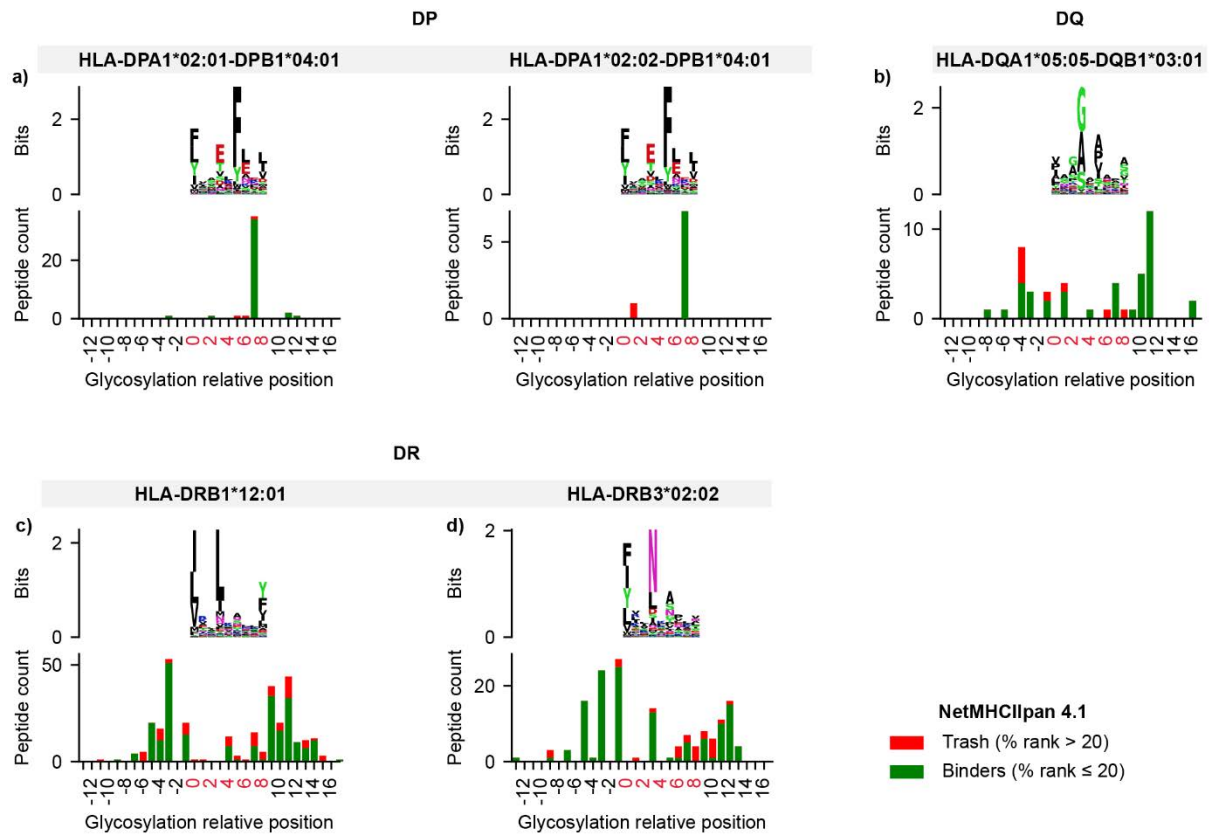


Figure 5: Fully unsupervised deconvolution of glycosylated HLA peptides from Ramarathinam *et al.* 2021 with NetMHCIIpan 4.1. Each panel illustrates 2 levels of information: the top level shows the HLA-binding motif of peptides passing a NetMHCIIpan 4.1 percentile rank threshold of 20 after binding affinity prediction. The bottom level shows glycopeptides that are predicted to bind to a given allele in green (%rank \leq 20), otherwise non-binder peptides (*i.e.*, trash) are shown in red (%rank $>$ 20). Positions are shown relatively to the HLA binding core with negative values referring to glycosylation position upstream the HLA-binding core, values between 0 and 8 represent positions within the HLA-binding core, and values \geq 9 refer to positions downstream the HLA-binding core. **a)** Deconvolution of glycosylated peptides associated with the HLA-DPA1*02:01/02-DPB1*04:01 alleles. **b)** Deconvolution of glycosylated peptides associated with the HLA-DQA1*05:05-DQB1*03:01 alleles. **c)** Deconvolution of glycosylated peptides associated with the HLA-DRB1*12:01 allele. **d)** Deconvolution of glycosylated peptides associated with the HLA-DRB3*02:02 allele.

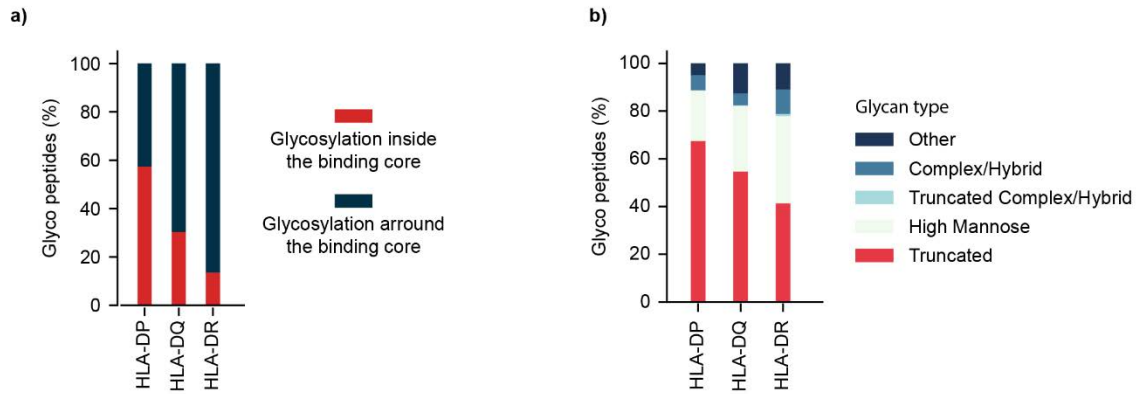
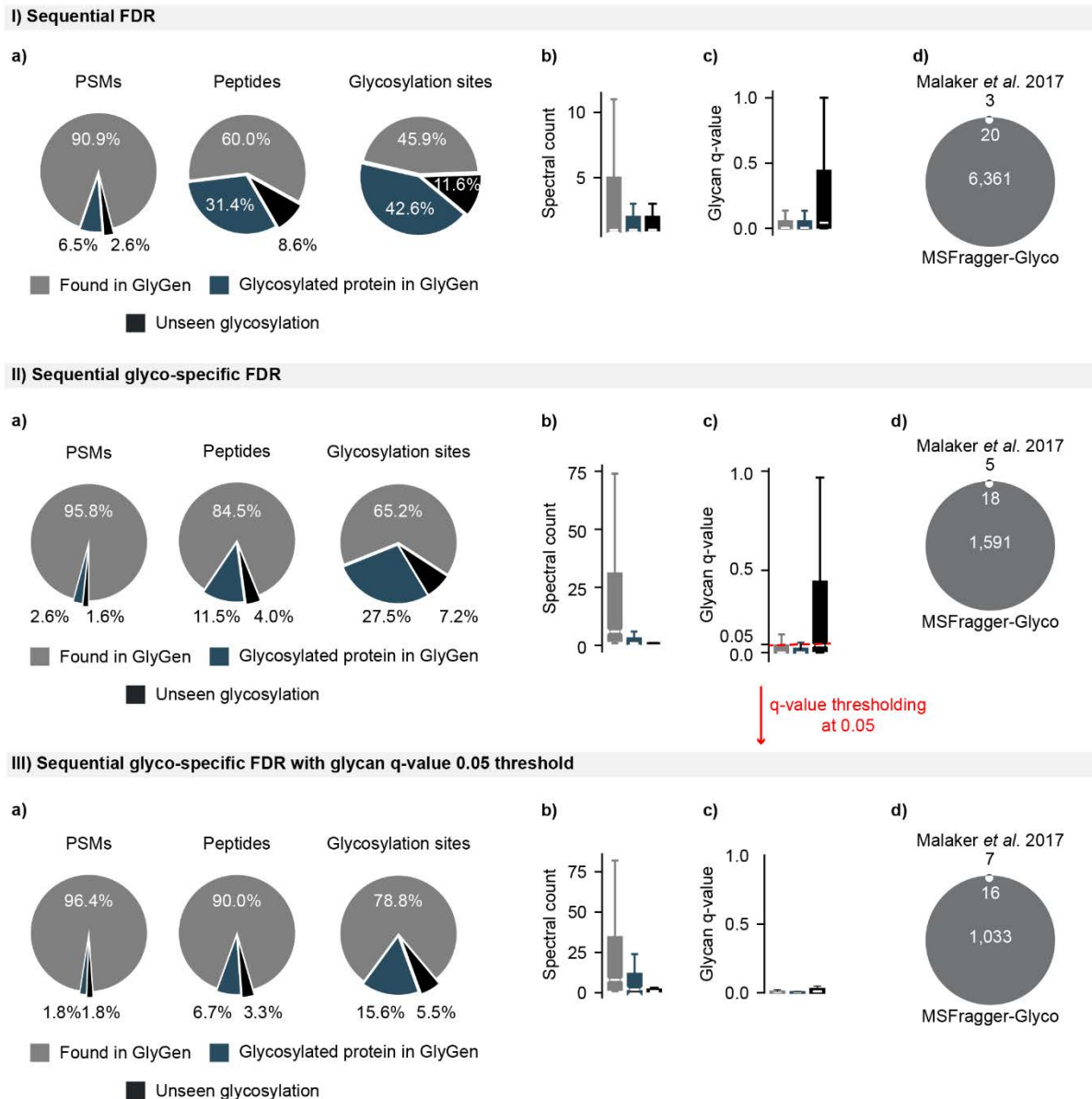


Figure 6: Glycan characteristics of the glycosylated HLA class II associated peptides. a) Percentage of glycosylation inside (red) and outside (blue) the HLA binding motif per HLA group (DP, DQ, and DR). **b)** Distribution of glycan types among the studied HLA class II group (DP, DQ, and DR).

Supplementary materials

Supplementary Figure 1



Supplementary Figure 1: Comparison of 3 different FDR control strategies for HLA glycosylated peptides. **Strategy I** referred to as “sequential FDR” is typically used with enzymatic (*i.e.*, trypsin) glycoproteomic searches. **Strategy II** referred to as “sequential glyco-specific FDR” has been developed in this study to handle non-specific (*i.e.*, non-specific cleavage of proteins at every peptide bond) glyco searches. **Strategy III** is the one being used in this study and consists of applying the sequential glyco-specific FDR with an additional glycan q-value threshold of 0.05. **a)** Percentage of glyco-PSMs, glycopeptides and glycosylation sites found in GlyGen. Peptides with glycosylation sites reported in GlyGen are shown in gray, within glycosylated protein are shown in blue, and unreported are shown in black. **b)** Abundance of the 3 categories from panel (a) by spectral count. **c)** The glycan q-value range of the 3 categories from panel (a). **d)** Comparison of the identified glycosylation sites identified in this study with Malaker *et al.* 2017 findings.

Chapter 4: Summary, milestones, and future directions

Previous chapters have paved the way for exploring nonconventional sources of antigens. In this chapter, I go beyond antigen presentation to discuss the underdeveloped aspects of antigen recognition. The first section provides a summary of my findings and highlights the novelty of the research. The second section outlines ongoing work that moves beyond MHC presentation to improve neoantigen prioritization.

Summary and highlights of the presented work

The new age of T-cell therapeutics is ushered in by the characterization of cancer neoantigens using both mass spectrometry and in silico approaches. The work presented in this thesis paves the way for exploring alternative sources of cancer antigens through two computational pipelines: COD-dipp and HLA-Glyco.

The development of COD-dipp allowed us to study the landscape of non-canonical MHC class I-associated peptides (ncMAPs), that is, peptides from non-coding regions of the genome. We designed a workflow for a large-scale analysis to explore the intricacies of ncMAPs. COD-dipp is completely free, open source, and does not require any paid or licensed software. The 772 collected immunopeptidomics samples spanned 11 cancer types, provided a pan-healthy panel of normals, and covered 114 HLA class I alleles. The analysis revealed a repertoire of 8,601 ncMAPs that proved to be shared not only between patients with the same cancer but also between different cancer types. Moreover, the panel of normals served to detect and filter ncMAPs expressed in healthy tissues. This is particularly important for clinical applications where off-target toxicities can pose an issue. To ensure minimal levels of toxicity, we evaluated the repertoire in the context of parental gene expression from 29 healthy tissues of 17,382 individuals, to stringently shortlist 17 cancer-selective ncMAPs according to our definition of 'cancer selectivity'.

The development of HLA-Glyco has allowed us to study the landscape of glycosylated MHC-associated peptides. The ultrafast search engine MSFragger combined with several layers of stringent False Discovery Rate (FDR) control enabled the large-scale study of the glyco-immunopeptidome for the first time. Glyco-searches on their own are not new; however, none of the existing free tools offer immunopeptidomic-oriented analysis owing to limitations in speed, sensitivity and lack of glycosylation-enrichment. We optimized the workflow using an iterative approach that included comparing the detected glycosylation sites with the proteome, and assessing the consistency of HLA motifs between glycosylated and non-glycosylated peptides. As we explored these two features, we achieved the best sensitivity by applying (I) a 1% global FDR, (II) a 1% group-specific FDR, and (III) a 5% cutoff for the glycan FDR (i.e., glycan q-value, which is a specific feature of MSFragger-Glyco). We created a library of over 3,400 HLA class II glycopeptides from 1,049 different protein-glycosylation sites from eight publicly available studies. The analysis revealed high levels of truncated glycans, conserved HLA-binding cores among the 72 HLA class II alleles under study, and distinct glycosylation positional specificity across classical allele groups. To assist further development in the field of glyco-immunopeptidomics, we (I) added the HLA-Glyco pipeline to the fragpipe suite, a tool used by thousands of scientists, and (II) provided the library as an online website for ease of access.

Over the past two to three decades, numerous studies have proposed various potential candidates for cancer vaccines, but some have failed, at least in part, due to a lack of tumor specificity. Laumont *et al.*^{1,2} addressed this issue using an elegant approach based on the elimination of genes expressed in the medullary thymic epithelial cells (mTECs). mTECs are found in the thymus and should represent the full antigenic repertoire of normal tissues in the body. The process of central tolerance depends on the presentation of self-antigens by mTECs to eliminate self-reactive T cells before entering circulation. Our work interrogates publicly available studies to explore ncMAPs in a large number of patients, cell lines, and

cancer types. However, most of these datasets do not contain mTECs gene expression data; hence, we addressed this issue by labeling non-tumor-selective non-canonical MHC class I-associated peptides (ncMAPs) when detected in a panel of normals (see Chapter 2). Although mass spectrometry made a long way in terms of improved accuracy and throughput, it still lags behind next-generation sequencing technologies in terms of sensitivity. Hence, the lack of MHC class I-associated peptides detection by mass-spectrometry does not guarantee their absence from a particular sample. In other words, the lack of detection in normal samples does not inherently qualify an ncMAP as tumor specific. Therefore, we introduced a filter based on parental gene expression in healthy tissues, referred to as cancer selectivity, to shortlist 17 non-canonical peptides with minimum healthy tissue toxicity for further clinical applications. From a post-translational perspective, assessing the tumor specificity or association of glycosylated MHC class I-associated peptides (see Chapter 3) is less straightforward. This is due to the inadequacy of next-generation sequencing technologies for measuring these events. All sources of antigens could benefit from a deeper understanding of the recognition process that I expand on in the next section.

Beyond antigen presentation, towards antigen recognition

Effective neoantigen selection begins with an accurate direct measurement (MS) or prediction of MHC presentation for a set of genomic, transcriptomic, or proteomic aberrations. MHC binding predictors have reached a decent level of accuracy for MHC class I. While early prediction tools relied solely on affinity data³, recent advances in MS immunopeptidomics have provided extensive ligand elution training data and helped boost accuracy. However, most predicted neoantigens do not end up being presented by the MHC system because available models partly model the processing and presentation of MHC-associated peptides. A plethora of conditions within the cell can influence the presentation of a particular peptide via the MHC system. For instance, recent developments have shown that performance can be improved

by incorporating antigen abundance data from RNA-Seq experiments⁴⁻⁸. Moreover, even when presented most identified neoantigens do not activate the immune system. It is clear that, in addition to MHC presentation, neoantigens must be recognized by T cell receptors (TCR) to illicit T cell activation. The surface receptor, known as the TCR, is a unique feature of T cells that mediates epitope identification through interactions with the peptide-MHC (pMHC) complex. TCRs are produced through a genomic rearrangement process that results in an astounding level of diversity. It is now widely acknowledged that TCRs exhibit high levels of cross-reactivity, that is, the ability to identify more than one pMHC complex⁹. According to certain theories¹⁰, a single TCR may be able to distinguish between 10^4 and 10^7 distinct MHC-associated epitopes. However, it has also been demonstrated that the likelihood of a TCR interacting with a different randomly chosen peptide drops to 10^{-4} once it interacts with a particular pMHC complex¹¹. Thus, TCR recognition is both cross-reactive and highly specific at the same time.

T-cell assays offer the most accurate assessment of immunogenicity when selecting antigens. For instance, the enzyme-linked immunosorbent spot (ELISpot) assay can be used to assess T cell reactivity by priming them with neoantigens and measuring activation markers like IFN- γ ^{12,13}. Although these assays are clinically robust predictors, they are time-consuming, expensive, and have a low throughput. Instead, *in silico* pipelines routinely associate strong binding with immunogenic potential¹⁴, however this practice is debatable since Ebrahimi-Nik et al.^{15,16} showed CD8+ T cell activation by low affinity pMHC complexes.

Modeling T-cell recognition is much more complex than MHC binding for many reasons, including the scarcity of training data, low binding affinity between the pMHC complexes and the TCRs (pMHC:TCR), and the large diversity of TCRs^{17,18}. Many attempts to predict TCR binding based on various hypotheses have been proposed and elegantly summarized by Gfeller *et al.* 2023¹⁹, Xie *et al.* 2023²⁰, Lee *et al.*²¹, Szeto *et al.*²², and Sim et al.²³. Here, I recapitulate the general knowledge around the pMHC:TCR recognition organized into three

discipline-based groups. Group I — *approaches based on biochemical characteristics of MHC-I ligands*; group II — *approaches based on structural information*; and group III — *approaches based on machine learning or deep learning*.

Group I — *Approaches based on biochemical characteristics of MHC-I ligands* — Calis *et al.* 2013²⁴ were the first to show that large and aromatic amino acid residues increase the likelihood of MAPs being immunogenic, and positions 4–6 have a significant impact on immunogenicity. Two years later, Chowel *et al.*²⁵ observed hydrophobic amino acid residues are enriched in immunogenic epitopes. Other features that predict peptide immunogenicity and hence neoantigen quality have been discovered. Quality metrics, such as the (I) differential agretopicity index^{26,27} (DAI), that is, the ratio of MHC affinity of the mutant peptide to that of its non-mutated counterpart; (II) dissimilarity to self^{28,29} (non-mutated proteome), have been shown to have some predictive power for immunogenicity; and (III) relative and absolute binding affinities with respect to the position of the mutation within the presented peptide³⁰. Several tools, such as PRIME^{31,32}, NeoScore³³, INeo-Epp³⁴, and pTuneos³⁵, rely on these ligand characteristics. Gfeller *et al.*^{31,32} suggested PRIME as an immunogenicity predictor and produced results consistent with the aforementioned characteristics. Likewise, NeoScore³³ predicts immunotherapy outcomes in melanoma patients, and INeo-Epp³⁴ incorporates the position information of the mutation along with amino acid-related characteristics. pTuneos³⁵ uses multiple features, including similarity between normal and mutant peptides, similarity with known immunogenic peptides, and hydrophobicity. However, these characteristics and metrics are sought as tendencies rather than rules. For instance, the impact of hydrophobicity on immunogenicity is suspected to be HLA allele dependent³³.

Group II — *approaches based on structural information* — TCRs are composed of two distinct protein chains qualifying them as heterodimers. Most Human T cells are composed of alpha (α) and beta (β) chains encoded by TRA and TRB loci, respectively. Each TCR chain is composed of two extracellular domains, a variable region (V) and a constant region (C). The

variable regions of each chain have three hypervariable or complementarity-determining regions (CDR1, CDR2, and CDR3). These six flexible CDR loops (3 α and 3 β) are generated through VDJ recombination, a process by which T and B cells randomly combine various gene segments, that is variable (V), diversity (D), and joining (J) genes, to create unique receptors.

Structural analysis of TCR:pMHC complexes revealed certain general principles. The hypervariable CDR3 loops are the primary drivers of peptide recognition, whereas germline-encoded CDR1 and CDR2 loops are primarily focused on the recognition of MHC molecules. The co-contribution from both α and β TCR chains is a common occurrence, with a roughly shared and balanced contribution. All currently available TCR:pMHC-I structures demonstrate that the TCR contacts both the peptide antigen and MHC. Despite the small size of the peptides relative to the MHC molecule, they might nonetheless contribute significantly to the pMHC:TCR interaction. This feature is not shared with lipid- or metabolite-derived specific TCRs, for which the recognition of both MHC/MHC-like molecules and the bound antigen is not required³⁶. Despite the wide range of docking orientations, the structures that have been solved thus far demonstrate that MHC-I-restricted TCRs must sit on top of the cleft to contact both the peptide and MHC-I helices. This is a specific feature of peptide-MHC-I recognition, for which no exception has been observed, even with a large number of solved structures. Moreover, Peptide length is associated with successful TCR engagement according to recent findings. Ekeruche-Makindeet *et al.*³⁷ demonstrated that TCR cross-reactivity was dependent on the length of the presented peptide, and that TCRs were unable to react to peptides of different lengths. On the same note, structural dissimilarity from the self, that is, structural and dynamic changes induced by point mutations at non-anchor sites, can influence TCR recognition and transmit effective T-cell activation³⁸.

There have been a few instances of reversed docking topology, in which the TCR chain is docked over the MHC-I α 1-helix and the TCR chain is in contact with the MHC-I α 2-helix³⁹. Interestingly, these TCRs were very weakly activated upon pMHC-I identification, while being

able to bind to pMHC-I with moderate affinities compared to the range in other TCR:pMHC-I complexes. This demonstrates that T cell activation is not only determined by the affinity of TCR:pMHC-I. Furthermore, it is possible that conventional docking topology is a prerequisite for T-cell activation. In addition to binding in a reversed orientation, some pMHC complexes interact with TCR with a C-terminal shift where the CDR3 Loop does not interact with the peptide^{39,40}.

In terms of computational development, recent studies have combined structural information to build a generalized TCR scoring system. Riley *et al.*⁴¹ designed an approach to capture both peptide-MHC and TCR-pMHC binding, based on six structural and physicochemical features. Aranha *et al.*⁴² showed that adding three-dimensional modeling to NetMHCpan increases specificity and precision and reduces the number of false positives when predicting neoantigens. Borrman *et al.*⁴³ suggested a scoring method and modeling approach that uses the structural characteristics of TCR-pMHC complexes to predict the binding of cross-reactive peptides. It should be noted that these models began to reveal the TCR:pMHC complex and are not yet able to execute ab initio prediction based on biophysical and structural data.

Group III — *approaches based on machine learning or deep learning* — Despite the potential variability of T-cell TCRs, there is evidence that they recognize the same pMHC epitopes frequently and possess similar sequence characteristics. For instance, DeWitt *et al.*⁴⁴ showed the existence of common patterns in the TCR repertoire across individuals exposed to the same disease. These findings suggest that the TCR epitope specificity can be predicted.

Identifying TCRs specificity to given antigens requires sorting, sequencing, and clustering of both naïve and antigen-experienced T-cell repertoires. Recent advances in bulk- and single-cell sequencing technologies have enabled the generation of high-throughput datasets. Consequently, software development has allowed computational biologists to further examine and profile TCR repertoires using specialized algorithms^{45,46}. Furthermore, efforts to catalog

such information have resulted in the creation of multiple databases such as McPAS-TCR⁴⁷ and VDJdb⁴⁸. McPAS-TCR is a manually curated database TCR sequences identified in human and mouse T cells linked to diverse clinical disorders. VDJdb is a database of TCR sequences with known antigen specificities.

Researchers have attempted to identify common features among antigen-specific TCRs by studying a collection of sequences. Based on the known interaction of the CDR3 loop with MHC-associated peptides, methods for clustering the recurrent short stretches of amino acids of these loops (i.e., CDR3 motifs) have emerged. Several techniques use distance metrics to assign previously unobserved TCRs to characterized repertoires or rely on clustering TCRs with comparable levels of specificity, such as pMTnet⁴⁹, GLIPH⁵⁰, TCRDist⁵¹, TCRnet⁵², ERGO II⁵³, and NetTCR-2.0⁵⁴.

An investigation was conducted by Grazioli *et al.*⁵⁵ to determine how well state-of-the-art deep learning models^{53,54,56-63} can predict TCR:pMHC binding and generalize to unknown peptides by evaluating ERGO II⁵³ and NetTCR-2.0⁵⁴. ERGO II relies on Long short-term memory (LSTM) networks and autoencoders to compute representations of peptides and CDR3s. NetTCR-2.0 uses a straightforward 1D convolutional neural network (CNN) model that integrates information from CDR3 and peptide sequences to predict TCR peptide specificity. The models did not generalize well to unseen peptides when using a hard split, a simple heuristic for training/test splits, which ensures that test samples exclusively present peptides that do not belong to the training set. The authors showed that this is largely due to suboptimal training/testing splits causing models to simply memorize the CDR3 sequences and ignore the peptides. To better predict the interactions between T-cell receptors (TCRs) and peptides, Grazioli *et al.*⁶⁴ proposed a new model called Attentive Variational Information Bottleneck (AVIB). The authors' benchmark shows that AVIB significantly outperforms cutting-edge techniques in predicting TCR-peptide interactions. However, the authors stated that

generalization to unseen sequences remains difficult because of the sparsity of the available training data.

Despite these efforts, it is still not possible to predict the set of TCRs that recognizes a certain antigen or the set of antigens that are recognizable by a certain TCR. This is due to several factors, including sparsity of training data relative to the size of the TRC repertoire, lack of alpha and beta pairing when performing TCR sequencing, excessive focus on the CDR3 β loops, lack of training data, and lack of structural modeling integration in the architecture of machine/deep-learning models.

When naïve T cells are activated by a pMHC complex, extensive proliferation and differentiation events occur. Qualitative differences arise when responding to antigens, including a less stringent requirement for activation with the ability to respond to lower concentrations of antigens than naïve T cells. This interplay between the low and high avidities of naïve and trained T cells, along with the large number of recognition patterns, poses challenges when it comes to the comprehensiveness of training data.

While TCR sequencing provides a high-throughput way to characterize repertoires, the majority of studies have focused on the β chain of the CDR3 loop due to its known interaction with the peptide and high combinatorial potential. However, both the α and β chains of CDR3 loops, and occasionally CDR 1 and 2 loops, contribute to antigen recognition.

Another key challenge in training accurate machine/deep-learning models is the cross-reactive nature of the TCRs. This implies that a comprehensive training dataset would require screening of the cross-reactivity spectrum for each unique group of TCRs. Regardless of the technical feasibility of this goal using wet-lab techniques, it is evident that such requirements are both labor- and cost-inefficient. Hence, there is an urgent need to better understand cross-reactivity for the rational simulation of data to supplement the true training data.

Finally, sequence-based deep-learning models could benefit from estimating the chemical interactions between TCR:pMHC complexes as well as the 3D structures in subsequent iterations.

References

1. Laumont, C. M. *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238 (2016).
2. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
3. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
4. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).
5. Garcia Alvarez, H. M., Koşaloğlu-Yalçın, Z., Peters, B. & Nielsen, M. The role of antigen expression in shaping the repertoire of HLA presented ligands. *iScience* **25**, 104975 (2022).
6. Chen, B. *et al.* Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* **37**, 1332–1343 (2019).
7. Koşaloğlu-Yalçın, Z. *et al.* Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions. *iScience* **25**, 103850 (2022).
8. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).

9. Matzinger, P. & Bevan, M. J. Why do so many lymphocytes respond to major histocompatibility antigens? *Cell. Immunol.* **29**, 1–5 (1977).
10. Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol. Today* **19**, 395–404 (1998).
11. Frank, S. A. *Immunology and Evolution of Infectious Disease*. (Princeton University Press, 2002).
12. Roudko, V. *et al.* Shared Immunogenic Poly-Epitope Frameshift Mutations in Microsatellite Unstable Tumors. *Cell* **183**, 1634-1649.e17 (2020).
13. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 1–16 (2016).
14. Koşaloğlu-Yalçın, Z. *et al.* Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncoimmunology* **7**, e1492508–e1492508 (2018).
15. Ebrahimi-Nik, H. *et al.* Mass spectrometry–driven exploration reveals nuances of neoepitope-driven tumor rejection. *JCI Insight* **4**, e129152 (2019).
16. Ebrahimi-Nik, H. *et al.* Reversion analysis reveals the in vivo immunogenicity of a poorly MHC I-binding cancer neoepitope. *Nat. Commun.* **12**, 6423 (2021).
17. Mora, T. & Walczak, A. M. *Quantifying lymphocyte receptor diversity*. <http://biorxiv.org/lookup/doi/10.1101/046870> (2016) doi:10.1101/046870.
18. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).

19. Gfeller, D., Liu, Y. & Racle, J. Contemplating immunopeptidomes to better predict them. *Semin. Immunol.* **66**, 101708 (2023).
20. Xie, N. *et al.* Neoantigens: promising targets for cancer therapy. *Signal Transduct. Target. Ther.* **8**, 9 (2023).
21. Lee, C. H. *et al.* Predicting Cross-Reactivity and Antigen Specificity of T Cell Receptors. *Front. Immunol.* **11**, 565096 (2020).
22. Szeto, C., Lobos, C. A., Nguyen, A. T. & Gras, S. TCR Recognition of Peptide–MHC-I: Rule Makers and Breakers. *Int. J. Mol. Sci.* **22**, 68 (2020).
23. Sim, M. J. W. & Sun, P. D. T Cell Recognition of Tumor Neoantigens and Insights Into T Cell Immunotherapy. *Front. Immunol.* **13**, 833017 (2022).
24. Calis, J. J. A. *et al.* Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* **9**, e1003266–e1003266 (2013).
25. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci.* **112**, E1754–E1762 (2015).
26. Sercarz, E. E. *et al.* Dominance and Crypticity of T Cell Antigenic Determinants. *Annu. Rev. Immunol.* **11**, 729–766 (1993).
27. Duan, F. *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* **211**, 2231–2248 (2014).
28. Richman, L. P., Vonderheide, R. H. & Rech, A. J. Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst.* **9**, 375–382 (2019).

29. Coelho, A. C. M. F. *et al.* neoANT-HILL: an integrated tool for identification of potential neoantigens. *BMC Med. Genomics* **13**, 30 (2020).
30. Capietto, A.-H. *et al.* Mutation position is an important determinant for predicting cancer neoantigens. *J. Exp. Med.* **217**, e20190179 (2020).
31. Schmidt, J. *et al.* Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep. Med.* **2**, 100194 (2021).
32. Gfeller, D. *et al.* Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *Cell Syst.* S2405471222004707 (2023) doi:10.1016/j.cels.2022.12.002.
33. Borden, E. S. *et al.* NeoScore Integrates Characteristics of the Neoantigen:MHC Class I Interaction and Expression to Accurately Prioritize Immunogenic Neoantigens. *J. Immunol.* **208**, 1813–1827 (2022).
34. Wang, G. *et al.* INeo-Epp: A Novel T-Cell HLA Class-I Immunogenicity or Neoantigenic Epitope Prediction Method Based on Sequence-Related Amino Acid Features. *BioMed Res. Int.* **2020**, 1–12 (2020).
35. Zhou, C. *et al.* pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. *Genome Med.* **11**, 67 (2019).
36. Rossjohn, J. *et al.* T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
37. Ekeruche-Makinde, J. *et al.* Peptide length determines the outcome of TCR/peptide-MHCI engagement. *Blood* **121**, 1112–1123 (2013).
38. Devlin, J. R. *et al.* Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nat. Chem. Biol.* **16**, 1269–1276 (2020).

39. Gras, S. *et al.* Reversed T Cell Receptor Docking on a Major Histocompatibility Class I Complex Limits Involvement in the Immune Response. *Immunity* **45**, 749–760 (2016).
40. Wu, D., Gallagher, D. T., Gowthaman, R., Pierce, B. G. & Mariuzza, R. A. Structural basis for oligoclonal T cell recognition of a shared p53 cancer neoantigen. *Nat. Commun.* **11**, 2908 (2020).
41. Riley, T. P. *et al.* Structure Based Prediction of Neoantigen Immunogenicity. *Front. Immunol.* **10**, 2047 (2019).
42. Aranha, M. P. *et al.* Combining Three-Dimensional Modeling with Artificial Intelligence to Increase Specificity and Precision in Peptide–MHC Binding Predictions. *J. Immunol.* **205**, 1962–1977 (2020).
43. Borrman, T., Pierce, B. G., Vreven, T., Baker, B. M. & Weng, Z. High-throughput modeling and scoring of TCR-pMHC complexes to predict cross-reactive peptides. *Bioinforma. Oxf. Engl.* **36**, 5377–5385 (2020).
44. DeWitt III, W. S. *et al.* Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* **7**, e38358 (2018).
45. Bolotin, D. A. *et al.* Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* **35**, 908–911 (2017).
46. Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Comput. Biol.* **11**, e1004503 (2015).
47. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).

48. Bagaev, D. V. *et al.* VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* **48**, D1057–D1062 (2020).
49. Lu, T. *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
50. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
51. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
52. Pogorelyy, M. V. & Shugay, M. A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies. *Front. Immunol.* **10**, 2159 (2019).
53. Springer, I., Tickotsky, N. & Louzoun, Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Front. Immunol.* **12**, 664514 (2021).
54. Montemurro, A. *et al.* NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR α and β sequence data. *Commun. Biol.* **4**, 1060 (2021).
55. Grazioli, F. *et al.* On TCR binding predictors failing to generalize to unseen peptides. *Front. Immunol.* **13**, 1014256 (2022).
56. Tong, Y. *et al.* SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction. *Comput. Biol. Chem.* **87**, 107281 (2020).
57. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs. *Front. Immunol.* **11**, 1803 (2020).

58. Fischer, D. S., Wu, Y., Schubert, B. & Theis, F. J. Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol. Syst. Biol.* **16**, e9416 (2020).
59. Weber, A., Born, J. & Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinforma. Oxf. Engl.* **37**, i237–i244 (2021).
60. Gielis, S. *et al.* Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Front. Immunol.* **10**, 2820 (2019).
61. Moris, P. *et al.* Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief. Bioinform.* **22**, bbaa318 (2021).
62. Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M. & Lähdesmäki, H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput. Biol.* **17**, e1008814 (2021).
63. De Neuter, N. *et al.* On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* **70**, 159–168 (2018).
64. Grazioli, F. *et al.* Attentive Variational Information Bottleneck for TCR–peptide interaction prediction. *Bioinformatics* **39**, btac820 (2023).

Appendix 1: A technical guide for the COD-dipp pipeline

General description

The COD-dipp workflow is based on Snakemake, a workflow manager responsible for bundling and integrating multiple tools. One major advantage of this setup is the breakdown of complex bioinformatic workflows into several small jobs that can be run in parallel. In addition, it automatically identifies the already completed tasks and avoids re-running them in the case of a re-launch. On top of that, the execution is modular, allowing the user to choose specific parts of the analysis. The tight integration with conda, a package and environment management system, simplifies the first deployment on new clusters. This means that users do not have to spend any time coordinating the installation of the dependencies. Another major strength of COD-dipp is the intelligent use of high-performance computing (HPC) resources. The setup relies on a configuration file in YAML format to specify the resource allocation in terms of CPU, GPU, memory, and time requirements. This allows the analysis of a large number of samples in a short amount of time using just one command.

COD-dipp integrates two orthogonal mass spectrometry DDA data analysis strategies. The first strategy is called open search and utilizes MSFragger, one of the fastest search engines, to identify peptides with or without post-translational modifications (*i.e.*, chemical modifications). *de novo* is the second strategy and is key for finding peptides from unannotated proteins holding great promise for the identification of neoantigens. COD-dipp uses DeepNovoV2 as the *de novo* engine, which leverages a deep learning architecture to extract features from the mass spectrometry spectra themselves and uses natural language processing. These two aspects of deep learning help in interpreting the noisy nature of mass spectrometry data and imputing missing values by learning the amino acid sequences from proteins themselves. Since deep learning models require a training step based on previously

available data examples, we used spectral matching results from the MS-GF+ search engine to train on-the-go *de novo* models in a personalized manner to each sample.

COD-dipp uses a multitude of quality control measures to ensure that the reported immunopeptidomes are not the result of computational errors or are simply false positives. To begin, the MS-GF+ results go through a rigorous post processing validation implemented by scavenger, a versatile post-search validation algorithm. Scavenger relies on gradient boosting, a machine learning technique that leverages up to 31 mass spectrometry features to differentiate between target (correct) and decoy (incorrect) identifications. This procedure is well established under the name of False Discovery Rate (FDR) control. *De novo* derived peptides are required to go through a stringent accuracy filter (90%) along with a first of a kind approach to map these sequences to the proteome as a first step then to the 3-frame translated transcriptome as a second step. This step is responsible for identifying non-canonical MHC class I-associated peptides (*i.e.*, peptides from non-coding regions). First, *de novo* peptides are aligned to a set of known proteins (*i.e.*, proteome). Peptides with at most one mismatch are labeled canonical peptides, and all other sequences are mapped to a 3-frame translation (3FT) database provided by the COD-dipp suite. A peptide labeled non-canonical would have at most one mismatch with the 3FT database and at least 3 amino acid differences from any known protein sequence. When it comes to open search, extra care needs to be taken due to the wide error tolerance of the strategy. The concept here is to allow a certain error tolerance when attributing peptides to mass spectrometry spectra as a first step, followed by an attempt to identify a chemical modification on one of the constituent amino acids explaining the mass shift. Thus, an extra control step is required to quality control the mass shift and its localization. For this, COD-dipp relies on PTMiner to control both the FDR and False Localization Rate (FLR) using a robust Bayesian method.

As a Final step COD-dipp tracks back all immuno-peptides to the genome by using PoGo for canonical peptides and pysam-based scripts for the non-canonical peptides.

Applications of the method

COD-dip was originally developed to deal with mass spectrometry-based immunopeptidomics in human samples. However, with flexibility in mind, all integrated tools are fully compatible with proteomics, making this pipeline easily applicable to standard proteomic mass spectrometry studies. This modification requires minor edits to the search engine parameter files (`FileMSGFPlus_Params.txt` and `Fragger_Params.txt`). Similarly, it can be adapted to other species with minor modifications to the below scripts in order to change the organism:

1. `scripts/prepare_annotation/generate_annotation.py`
2. `scripts/prepare_annotation/genes3FT_generator.R`

Experimental design

The analysis setup is relatively simple and requires pooling all the mass spectrometry files in a folder. First, each sample requires its own folder with the naming convention 'sample_*', where * is any chosen string. The sample folder must contain the MS files in mzML format. In addition, the sample folder must contain a sub folder named 'denovo' containing the MS files in MGF format. COD-dipp automatically detects all samples along with their corresponding mass spectrometry files.

Expertise needed to implement the protocol

COD-dipp requires basic knowledge of the bash syntax to execute commands over the command line. Familiarity with the workload manager SLURM is appreciated in very specific cases where the HPC has an unusual setup or lacks GPU availability.

Hardware requirements

COD-dipp was designed to run on HPC clusters to leverage parallel computation. The minimum requirements include 64 GB of RAM and 12 cores. The recommended requirements are 120 GB of ram, 24 cores, and 1 GPGPU (General-Purpose Graphics Processing Unit).

Software requirements

COD-dipp can be found at <https://github.com/immuno-informatics/COD-dipp> and has been tested on a Centos 7 linux system. It requires SLURM to be installed, Python 3, Snakemake v5.4.5, Anaconda, Singularity, and MSFragger.

PROCEDURE

Annotation generation step

1. For a human immunopeptidomics analysis, it is sufficient to download the pre-generated data on this figshare link (<https://doi.org/10.6084/m9.figshare.16538097>) under the file name “pipeline_annotation_files.zip” and skip the next step.
2. For non-human immunopeptidomics analysis, additional steps are required:
 - a. Download the pre-generated data on this figshare link (<https://doi.org/10.6084/m9.figshare.16538097>) under the file name “pipeline_annotation_files.zip”.
 - b. Download the protein database for the organism in question from ENSEMBL BioMart.
 - i. Go to www.ensembl.org and click on biomart in the tools section. Choose the ‘Ensembl Genes’ database. Then choose the desired organism dataset. Click on ‘Sequences’, select ‘Peptides’ in the ‘sequences’ section. Expand the ‘header information’ section and unselect everything, then select the following attributes in the exact order: Protein stable ID, Transcript stable ID, Gene stable ID. Gene name, Gene description. then click on ‘Results’ on the top left corner. At this point click on ‘results’ in the top left corner and retrieve the fasta file. Then use Philosopher (<https://philosopher.nesvilab.org>) to add a list of contaminants and the decoys.

- ii. Edit the script `scripts/prepare_annotation/1_generate_annotation.py` at `g = Genome(db="hg38")` to the genome of the organism of interest. Then execute the script to generate a 'UCSC_knownGene_hg38_features.tsv' like file.
- iii. Go to <https://genome.ucsc.edu> head to 'tools' then table browser. Choose the organism of interest in 'genome' and select 'ENSEMBL genes' in track and click 'get output' to download the file. Feed this file to the `scripts/prepare_annotation/genes3FT_generator.R` to generate the equivalent of 'df_features_inframes.tsv'.
- iv. Edit the script `scripts/prepare_annotation/genes3FT_generator.R` at `ens94_human_dna` and `txdbENS` to your own organism of interest and execute the script to generate the 3FT database like file '3FTgenes_coding.fasta'.

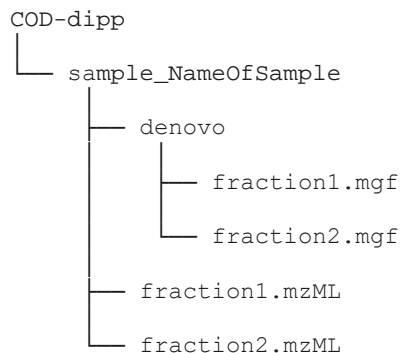
Raw data conversion

3. Ensure that your raw files are correctly converted to mzML and MGF formats. The `msconvert` gui or command-line tool can be used. Please ensure that the peak picking filter is the first filter and that TPP compatibility option is checked. Please check the example below for command line conversion:

```
msconvert fraction1.raw --mgf --filter "peakPicking true 1-" --
filter 'titleMaker <RunId>.<ScanNumber>.<ScanNumber>.<ChargeState>
File:"<SourcePath>", NativeID:"<Id>"'
msconvert fraction1.raw --mzML --filter "peakPicking true 1-" --
filter 'titleMaker <RunId>.<ScanNumber>.<ScanNumber>.<ChargeState>
File:"<SourcePath>", NativeID:"<Id>"'
```

Environment setup

4. Please ensure that MS files follow the below organization:



5. Create the conda environments required for the workflow to run:
- Edit `conda_prefix` path in `prepare_envs.sbatch` to a desired location on the cluster.
 - Edit the `prepare_envs.sbatch` SLURM `-A` parameter to specify an active SLURM account.
 - Launch this command `sbatch prepare_envs.sbatch` to create the conda environments at the specified `conda_prefix`.
6. Downloading COD-dipp can be performed using a few simple lines of code:

```
# create a directory for your study
study_id="Example_Study"
# go to the directory for your study
mkdir $study_id && cd $study_id

# clone the git environment (workflow code)
git clone https://github.com/immuno-informatics/COD-dipp.git
cd cod-dipp

# clean up test data
rm -rf sample_test database.fasta
mv $(ls -A) ../
cd ..
rmdir cod-dipp
```

```
# download link: https://doi.org/10.6084/m9.figshare.16538097
unzip pipeline_annotation_files.zip
resource_files_dir="/PATH/TO/pipeline_annotation_files/folder"

# copy your database to your study working dir, if you are
analysis human samples
cp $resource_files_dir/2019-04-30-td-
Homo_sapiens_GRCh38_biomart.fasta ./database.fasta
```

7. Edit the HPC cluster job settings:

a. Edit the `integrated-pipeline-profile/config.yml` configuration file.

- i. In the `PATH` section, add the full path to each of the required files.
- ii. In the `SEARCH ENGINE params` section edit the amount of memory to be allocated in MB in case the existing values exceed your HPC capacity. Setting the memory requirement too low will raise an `OutOfMemory` error for either of the search engines.

b. Edit `integrated-pipeline-profile/cluster-config.json` to adapt it to your own setup. Here, we describe the list of parameters and how to tune them for the best compatibility:

i. **Cluster-specific parameters that requires tuning:**

1. "Account": assigns resources used by the pipeline to a specified account on the HPC. Equivalent of `--account` when using SLURM. Reverts to 'normal' if not specified.
2. "partition": Requests a specific partition for the resource allocation. Reverts to 'Long' if not specified. Equivalent of `--partition` when using SLURM.

ii. **Pipeline-specific parameters** that in most cases do not need to be modified:

1. "cpus": number of cores to allocate. Equivalent of `--account` when using SLURM.

2. "memory": amount of RAM to allocate in MB. Equivalent of `--mem` when using SLURM.
3. "time": "HH:MM:SS" Sets a limit on the total run time of the job allocation. Equivalent of `--time` when using SLURM.
4. "name": the default assigned name, will automatically get modified by the `launch.sh` script to `'dir_name.job_name'`.
5. "nodes": the minimum requested number of nodes for resource allocation. Equivalent of `--nodes` when using SLURM.
6. "ntasks": in all cases should have a value of 1. Except for the rule `'denovo_annotation'` where the Message Passing Interface (MPI) is used. Equivalent of `--ntasks` when using SLURM.
7. "gres": Specifies a comma-delimited list of generic consumable resources. Should have a value of 0 in all cases except for the `denovo` rule where 1 GPU is requested with the following value `"gpu:1"`. Equivalent of `--gres` when using SLURM.

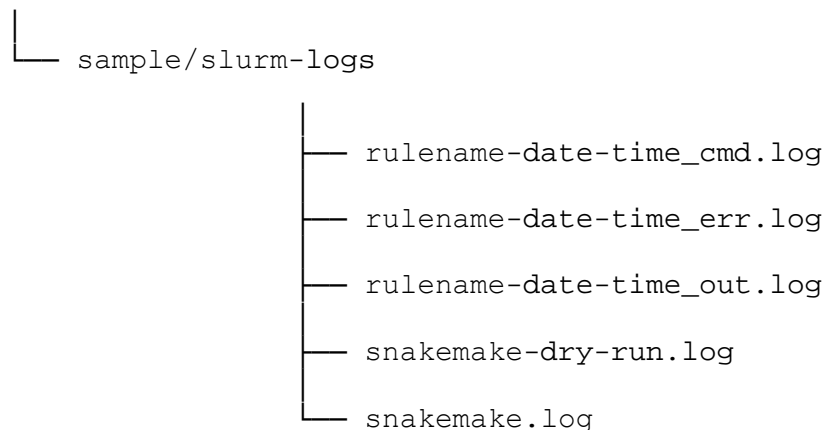
Launching the analysis

We made it easy to execute the pipeline with a bash wrapper script under the name `launch_pipeline.sh`. This script assumes that the sample folder names start with `'sample_'`. The value of the variable `'type'` in the `launch_pipeline.sh` script could take the values `"cluster"`, `"local"`, or `"dry-run"`. If `'type'` is given the value `'cluster'` the pipeline will execute the step as SLURM jobs on the HPC. However, before running the actual jobs, it is always helpful to launch a `'dry-run'` to ensure all the requirements are satisfied. `'dry-run'` will only display what would be done without executing the commands. The `'local'` option is particularly useful for debugging since it will launch the commands directly on the machine without the use of SLURM.

TROUBLESHOOTING

- The pipeline generates the following logs when launched:
 - slurm-logs directory containing all the launched jobs:
 - Rulename-date-time_cmd.log containing the command used to launch the job in question.
 - Rulename-date-time_err.log as the standard error output when something goes wrong.
 - Rulename-date-time_out.log as the standard output, where all printed information from the executed commands within the job in question are printed.
 - Snakemake-dry-run.log when a dry-run launch is executed by the user. This file contains the Snakemake log of all planned jobs, along with the expected output files.
 - Snakemake.log when a 'dry-run' launch is executed by the user. This file contains the Snakemake log of the executed jobs. In the case of a communication error with the clusters, this file contains useful information on how to fix it.

working_dir



Troubleshooting advice can be obtained from the COD-dipp help forum, which can be found at <https://groups.google.com/g/cod-dipp>. In case you encounter a bug please raise an issue at <https://github.com/immuno-informatics/COD-dipp/issues>.

TIMING

A study with multiple patients would still take 10 to 12 hours to complete on a cluster owing to the parallel computations. For instance, the analysis of the pride dataset PXD004894 (i.e., 25 patients) comprising 140 MS files took over 12 hours (real time) and approximately 28892

computational hours (~5000 GPU hours for DeepNovoV2, ~7000 CPU hours for MS-GF+, ~16800 CPU hours for MSFragger, and ~92 CPU hours for Scavager).

ANTICIPATED RESULTS

File name	Description
Folder: reports/denovo_annotation	
3ft_coords_3m.tsv	<i>Non-canonical immunopeptides genomic coordinates.</i>
3ft_coords_4m.tsv	This table reports the alignment of the non-canonical peptides on the genome including the chromosome, star end and number of mismatches.

3ft_coords_annotation_3m_framecheck.tsv	<i>Non-canonical immunopeptides frame analysis.</i>
	This table reports if the non-canonical peptides from introns follow the upstream exon frame.

3ft_coords_annotation_3m.tsv	<i>Non-canonical immunopeptides annotation.</i>
3ft_coords_annotation_4m.tsv	<i>This table reports the type of feature the non-canonical peptides align to: Exon, intron, Exon out of frame, 5UTR, 3UTR.</i>

3ft_features_3m.tsv	<i>Non-canonical immunopeptides annotation.</i>
3ft_features_4m.tsv	<i>Simplified table that reports the type of feature the non-canonical peptides align to: Exon, intron, Exon out of frame, 5UTR, 3UTR.</i>

Denovo_exon_spectra_3m.tsv	<i>De novo peptides that map to known proteins.</i>
denovo_exon_spectra_4m.tsv	<i>De novo peptide spectrum matches for sequences coming from protein (i.e., exons)</i>

Denovo_nonexons_spectra_3m.tsv	<i>De novo peptides that map to non-canonical sequences.</i>
denovo_nonexons_spectra_4m.tsv	<i>De novo peptide spectrum matches for sequences coming from ltrons, out of frame exons, 5 and 3' UTRs.</i>

Stats_3mismatches.pdf	<i>Descriptive analysis of the results.</i>
Stats_4mismatches.pdf	PDF files with bar plots and pie charts of comparison between canonical and non-canonical <i>de novo</i> peptides.
denovo_plots.pdf	

Folder: reports/denovo_annotation/TITER

TIS_analysis.pdf *Descriptive analysis of the results.*
Bar plots to describe the intronic peptides coming from upstream Translation Initiation Sites (TIS).

df_titer_pos.tsv *De novo intronic peptides resulting from an upstream Translation Initiation Site.*
This Table reports *de novo* peptides coming from introns and that TITER predicts an upstream Translation Initiation Site (TIS) for.

Folder: Reports

denovo_data_prep.tsv *De novo peptide spectrum matches with 90% accuracy.*
The output of DeepNovoV2 after applying a 90% accuracy filter. These results have not been mapped to a gene source and must be used with caution.

Folder: reports/PTMiner

filtered_result_processed.tsv *Open search validation results.*
This table contains the open search PSM results after applying a 1% False Discovery Rate by PTMiner.

loc_result_pocessed.tsv *Open search Localization results.*
This table contains the open search PSMs results after applying a 1% False Localization Rate for spectra identified with a mass shift.

anno_result_processed.tsv *Open search annotation results.*
This table contains the open search results for PSMs that passed the 1% FLR filter and went through mass shift annotation with UNIMOD.

Folder: reports/frs_msgfplus

Model.pdf *Descriptive analysis of MS-GF+ results validation model.*
This pdf offers a peek at the model that used to validate MS-GF+ results.

scavager_PSMs_full.tsv

Peptide spectrum matches of MS-GF+ results.

scavager_PSMs.tsv

These 2 tables are the output of Scavager after False Discovery Rate control to 1% of MS-Gf+ PSMs.

scavager_peptides.tsv

Peptide summary level of MS-GF+ results.

This table is the output of Scavager after False Discovery Rate control to 1% of MS-GF+ peptides.

scavager_proteins.tsv

Protein summary level of MS-GF+ results.

This table is the output of Scavager after False Discovery Rate control to 1% of MS-GF+ proteins. This table might sometimes be omitted due to the nature of HLA associated peptides hindering the protein inference.

Supporting document 1: Contribution statement from the co-authors of chapter 2**RE:** A statement of authorship contribution.March 16th, 2023

Dear dr hab. Joanna N. Izdebska, prof. UG,

We, the contributing authors, are writing to confirm our participation in the manuscript titled **“The immunopeptidome from a genomic perspective: Establishing the non-canonical landscape of MHC class I-associated peptides.”**, which is being included as chapter 2 in Georges Bedran’s PhD doctoral thesis. We would like to emphasize that Georges Bedran’s contribution to this work was utterly significant, as he was primarily responsible for conducting the research, performing data analysis and interpretation, and drafting the manuscript.

The authors’ contribution can be found below:

Georges Bedran:

1. Conceived and initiated the project.
2. Wrote the first draft of the manuscript.
3. Collected online studies.
4. Developed the computational approach and software.
5. Processed the data.
6. Created and revised figures.
7. Coordinated the manuscript.

Hans-Christof Gasser:

1. Created a supplementary figure.

Tongjie Wang:

1. Revised the manuscript.

Dominika Bedran:

1. Revised the manuscript.

Kenneth Weke:

1. Revised the manuscript.

Alexander Laird:

1. Revised the manuscript.

Christophe Battail:

1. Revised the manuscript.

Fabio Massimo Zanzotto:

1. Revised the manuscript.

Catia Pesquita:

1. Revised the manuscript.

Håkan Axelsson:

1. Revised the manuscript.

Ajitha Rajan:

1. Revised the manuscript.

David J. Harrison:

1. Revised the manuscript.

Aleksander Palkowski:

1. Revised the manuscript.

Maciej Pawlik:

1. Revised the manuscript.

Maciej Parys:

1. Revised the manuscript.

Robert O’Neill:

1. Revised the manuscript.

Paul M. Brennan:

1. Revised the manuscript.

Stefan N. Symeonides:

1. Revised the manuscript.

David R. Goodlett:

1. Revised the manuscript.

Kevin Litchfield:

1. Revised the manuscript.

Robin Fahraeus:

1. Revised the manuscript.

Ted R. Hupp:

1. Revised the manuscript.

Sachin Kote:

1. Coordinated and supervised the project.
2. revised the manuscript.

Javier A. Alfaro:

1. Conceived and initiated the project.
2. Coordinated and supervised the project.
3. wrote the first draft of the manuscript.
4. Created and revised figures.
5. Revised the manuscript.

Supporting document 2: Acceptance letter from Cancer Immunology Research

From: Javier Alfaro
Sent: Thursday, March 16, 2023 10:03 PM
To: Georges Bedran; Georges BEDRAN
Subject: Fwd: Decision Rendered: CIR-22-0621R3

----- Forwarded message -----

From: **Cancer Immunology Research** <cancerimmunolres@msubmit.net>
Date: Thu, Mar 16, 2023, 9:02 PM
Subject: Decision Rendered: CIR-22-0621R3
To: <javier.alfaro@proteogenomics.ca>
Cc: <Javier.alfaro@ug.edu.pl>

Re: CIR-22-0621R3

"The immunopeptidome from a genomic perspective: Establishing the non-canonical landscape of class I MHC-associated peptides."

Dear Dr. Alfaro:

I am pleased to inform you that your above-referenced manuscript has been accepted for publication in *Cancer Immunology Research*. Thank you very much for this interesting contribution to the journal; we have appreciated working with you throughout this process. Please ensure that you read this letter in its entirety for important details surrounding production and publication.

Proofs

Please note that you will receive page proofs at this email address in 2 to 4 weeks' time. If you expect your email address to change within this period, please notify us immediately.

Detailed editing instructions will be sent to you along with the proofs. We ask that you please read, correct, and return the proofs within 2 business days. **Please note that, if proofs are not returned within this time frame, final publication of the article may be delayed.**

Press Releases

If your institution's public relations office is planning a press release or other press-related activity for this paper, please send an email to richard.lobb@aacr.org immediately and cc the journal office at cancerimmunolres@aacr.org to alert us to hold online publication of your manuscript. A member of the AACR Communications Department will then liaise with your institution to ensure that embargo policies are followed.

OnlineFirst Publication

If a press release is not planned for your paper, please note that the accepted manuscript will be posted on our website as an OnlineFirst article in about 48 to 72 hours. OnlineFirst publication entails publishing online the author manuscript in its current form, which is not yet copyedited or typeset. After proof corrections have been returned, the final edited version of your article will replace the author manuscript on our website.

Funding Mandates

If during submission you requested that the AACR deposit the accepted version of this manuscript on your behalf to PubMed Central (PMC) or Europe PMC to satisfy public access requirements of certain US and European funders, this deposit will now be initiated. Please note this is not required if you chose an open access license, as AACR will automatically deposit the final typeset article in PMC and Europe PMC with no embargo period. If you did not select this option (or are otherwise unsure of your selection) and you prefer that the AACR deposit the paper on your behalf, please notify us immediately. Please be aware that deposition will not be complete until you respond to a request from NIHMS or Europe PMC to verify and approve the deposit.

Billing

All article processing charges (e.g., the flat publication fee, any display item fees, and any open access fee) are administered by the Copyright Clearance Center (CCC) through their online billing platform RightsLink. If any fees are associated with this manuscript, you will soon receive an email from the RightsLink Author system at copyright.com. The email notification will contain a link to view and pay these charges. Please confirm your order promptly upon receipt to begin the payment process and avoid publication delays.

Twitter

Cancer Immunology Research is on Twitter at @CIR_AACR. The journal may tweet about your article upon its publication. If you have not done so already and would like to give the journal the option to tag you in any such tweets, please click the link below to log into your SmartSubmit account profile and add your Twitter handle.

Thank you very much again for submitting your outstanding manuscript to *Cancer Immunology Research*. We sincerely appreciate your support of this important journal, and we are looking forward to publishing your work shortly.

Sincerely,


Karen Honey, PhD
Senior Associate Editor
For the *Cancer Immunology Research* Editorial Board

Supporting document 3: Contribution statement from the co-authors of chapter 3**RE:** A statement of authorship contribution.March 16th, 2023


Dear dr hab. Joanna N. Izdebska, prof. UG,

We, the contributing authors, are writing to confirm our participation in the manuscript titled “**HLA-Glyco: A large-scale interrogation of the glycosylated immunopeptidome.**”, which is being included as chapter 3 in Georges Bedran’s PhD doctoral thesis. We would like to emphasize that Georges Bedran’s contribution to this work was utterly significant, as he was primarily responsible for conducting the research, performing data analysis and interpretation, and drafting the manuscript.

The authors’ contribution can be found below:

Georges Bedran:  _____

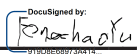
1. Collected and curated the data.
2. Generated the figures.
3. Generated the supplementary materials.
4. Drafted and coordinated the manuscript.

Daniel A. Polasky:  _____


1. Performed the immunopeptidomics analysis.
2. Supported figure generation, interpretation of results, and drafting and coordination of the manuscript.

Yi Hsiao:  _____

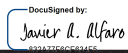
1. Produced the web portal.
2. Revised the manuscript.

Fengchao Yu:  _____

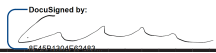
1. Supported the study with software development related tasks.

Felipe V. Leprevost:  _____


1. Supported the study by adding a group-specific FDR feature to Philosopher.

Javier A. Alfaro:  _____

1. Supported with the writing of the manuscript.

Marcin Cieslik:  _____

1. Helped with the study design.
2. Revised the manuscript.

Alexey I. Nesvizhskii:  _____

1. Conceived the project.
2. Helped with the study design.
3. Revised the manuscript.
4. Provided overall supervision.