Year : 2020

Deciphering the Landscape of HLA class-I and class-II Phosphopeptidomes leads to Robust Predictions of Phosphorylated HLA ligands

Solleder Marthe

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'Oncologie Fondamentale

Deciphering the Landscape of HLA class-I and class-II Phosphopeptidomes leads to Robust Predictions of Phosphorylated HLA ligands

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Marthe SOLLEDER

M.Sc. in Bioinformatics
Freie Universität Berlin, Germany

Jury

Prof. Pedro Romero	Président
Prof. David Gfeller	Directeur de thèse
Prof. Jacques Fellay	Expert
Dr. Pedro Beltrao	Expert

Lausanne, 2020

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'Oncologie Fondamentale

Deciphering the Landscape of HLA class-I and class-II Phosphopeptidomes leads to Robust Predictions of Phosphorylated HLA ligands

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Marthe SOLLEDER

M.Sc. in Bioinformatics
Freie Universität Berlin, Germany

Jury

Prof. Pedro Romero	Président
Prof. David Gfeller	Directeur de thèse
Prof. Jacques Fellay	Expert
Dr. Pedro Beltrao	Expert

Lausanne, 2020

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Pedro	**Romero**
Directeur·trice de thèse	Monsieur	Prof.	David	**Gfeller**
Expert·e·s	Monsieur	Prof.	Jacques	**Fellay**
	Monsieur	Dr	Pedro	**Beltrao**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Marthe Solleder

Master of Science, Freie Universität Berlin, Allemagne

intitulée

Deciphering the landscape of HLA class-I and class-II phosphopeptidomes leads to robust predictions of phosphorylated HLA ligands

Lausanne, le 30 octobre 2020

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Niko GELDNER
Directeur de l'Ecole Doctorale

i stand
on the sacrifices
of a million women before me
thinking
what can I do
to make this mountain taller
so the women after me
can see further

legacy – rupi kaur

Table of Content

Acknowledgements ... 9

Summary ... 11

Resumé ... 13

List of Figures ... 15

Abbreviations ... 17

Chapter 1 Introduction ... 21
1.1 Immune System ... 21
1.2 Antigen Presentation ... 22
1.2.1 Human Leukocyte Antigen System ... 22
1.2.2 Cancer Antigens ... 27
1.3 *In silico* Models to Predict HLA – Ligand Interactions ... 29
1.3.1 Predictors for HLA Ligand Interactions ... 30
1.3.2 Identification of Binding Motifs in HLA Peptidomics Data ... 32
1.4 Phosphorylation ... 33
1.4.1 Phosphorylation in HLA Ligands ... 35
1.5 Aim and Objective of this Thesis ... 38

Chapter 2 Manuscript "Mass Spectrometry – Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands" ... 41
2.1 Summary of Results ... 42

Chapter 3 Deciphering the landscape of phosphorylated HLA-II ligands ... 45
3.1 Introduction ... 45
3.2 Results ... 47
3.3 Discussion ... 56
3.4 Methods ... 58
3.5 Supplemental Figures ... 62

Chapter 4 Conclusions & Discussion ... 65

References ... 71

Appendix A – Manuscript "Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands" ... 89

Acknowledgements

First and foremost, I would like to acknowledge **Prof. David Gfeller** for all his support and guidance during my PhD. I am forever grateful for every opportunity during the last four years, your expertise shaping my scientific mind, and the support I received for my work and career. I would also like to express my deepest gratitude to all members of the lab, **Julien Racle, Giancarlo Croce, Loc Tran, Mariia Bilous,** and **Simon Eggenschwiler**. Thank you to **Julien** for all your support and help on my projects! I am grateful to my fellow PhDs, **Mariia** and **Simon**, I wouldn't have made it without you in the office – thank you for sharing those scientific and non-scientific moments together! I would also like to thank all former members of the lab for their help and support on my PhD journey. In particular my gratitude goes out to **Santiago Carmona** and **Rachel Marty Pyke**. You helped me feel like home in Lausanne in the beginning, always engaged me in capturing scientific (and non-scientific) conversations, and you were the best travel companions. I am very grateful for your friendship! I would also like to acknowledge **Dr. Michal Bassani-Sternberg** and her lab and thank everyone for all their contribution, expertise, and support to my thesis work. Furthermore, I would like to thank my PhD jury for all the valuable advice and engaging discussions contributing to my research: thesis president **Prof. Pedro Romero** as well as my experts **Prof. Daniel Speiser, Dr. Immanuel Luescher**, **Prof. Jacques Fellay, and Dr. Petro Beltrao** who were present during my first-year, mid-thesis as well as final thesis evaluations.

I am thankful for all my friends in Lausanne who made the last four years special, who were always there for me, and who supported me in these last months. A big shout out to **Valentina, Justine, Mariia, Chloe,** and **Julie** – you made the years of my PhD to what they were and I am forever grateful for all your support and your friendship. So many memories will stay with me forever! Thank you to **Sadeq** for always believing in me, even when I couldn't. Words can't describe my gratitude for your endless support. I would also like to thank **Frauke**, **Johanna**, and **Lydia** for being there for me. Thank you for every shared memory from our time in Berlin and elsewhere, for all those long conversations, and for supporting me in everything professional and personal! **Rukeia**, I am more than thankful for you by my side in the same situation throughout these last four years. Thank you for your support and your friendship! **Alexandra**, **Diandra**, **Hannah**, **Lena**, and **Vicky**, I am so grateful for calling you my friends for so many years. You were my "non-science"–rock throughout the last 10 years, you always make *back home* more home, and you are there for me whenever I need you! **Charlotte** – thank you for your endless friendship and every shared memory. Thank you for allowing me to be a big part of Emilia's life. **Lena**, I can't even begin to find the words to thank you for your love and friendship. Thank you for checking in on me every day, for keeping me grounded, and for supporting my decisions no matter where they take me. You are the best friend anyone could wish for.

Finally, I would like to give my deepest gratitude to my family. **Julian**, you showed me that everything is possible, no matter which way we go, and I can always count on you. **Mama** and **Papa**, I am thankful for every opportunity you provided me and for all your support, particularly in the last four years. I wouldn't be where I am today without you and I cannot thank you enough for your love.

Summary

Activation of CD8+ and CD4+ T cells through recognition of antigens presented by class I and class II human leukocyte antigen (HLA-I/HLA-II) molecules is crucial for immune responses against infected or malignant cells. In cancer, neoantigens can arise from cancer-specific genomic or proteomic alterations, including mutations and aberrant post-translational modification, such as phosphorylation. Identifying HLA ligands remains a challenging task that requires either heavy experimental work for *in vivo* identification or optimized bioinformatics tools for accurate predictions. While much work has been done on unmodified HLA-I and HLA-II ligands, only little is known about the presentation of phosphorylated peptides, in particular by HLA-II molecules. Moreover, none of the existing *in silico* models for predictions of HLA – ligand interactions are specifically trained on phosphorylated ligands.

This thesis presents in-depth analyses of phosphorylated HLA-I and HLA-II ligands and introduces predictors for HLA – phosphorylated ligand interactions. The first part of this thesis comprises the curation of phosphorylated HLA-I ligands from several Mass Spectrometry – based peptidomics studies, identifying more than 2,000 unique phosphorylated peptides covering 72 HLA-I alleles. Furthermore, it was see that phosphorylated HLA-I ligands are shaped by a combination of HLA-I binding motifs, intrinsic HLA-I binding properties of phosphorylated ligands and kinase motifs. Combining phosphorylated HLA-I ligands with unmodified data for training a prediction model resulted in improved predictions of phosphorylated HLA-I ligands.

The second part addresses phosphorylated HLA-II ligands presented by professional antigen presenting cells for CD4+ T cell activation. MS – based HLA-II peptidomics data resulted in the identification of binding motifs for more than 30 HLA-II alleles, comprising 2,473 unique phosphorylated ligands. These were used to retrain a predictor for HLA-II - ligand interactions and showed improved accuracy for phosphorylated ligands. The analysis of the phosphorylated HLA-II peptidomes revealed a more diverse repertoire of kinases responsible for the phosphorylation of peptides presented on HLA-II compared to HLA-I.

In summary, the current work presents in-depth studies on phosphorylated HLA ligands as well as bioinformatics tools for the predictions of phosphorylated peptide interactions with HLA-I and HLA-II molecules.

Resumé

L'activation des cellules T CD8+ et CD4+ suite à la reconnaissance d'antigènes présentés par les antigènes des leucocytes humains de classe I et II (HLA-I/HLA-II) est cruciale pour les réponses immunitaires contre les cellules infectées ou cancéreuses. Dans le cancer, les néoantigènes peuvent provenir d'altérations génomiques ou protéomiques spécifiques au cancer, par exemple des mutations ou des modifications post-traductionnelles aberrantes, telles que la phosphorylation. L'identification des ligands HLA reste une tâche difficile qui nécessite soit un travail expérimental lourd pour l'identification *in vivo*, soit des outils bio-informatiques optimisés pour des prédictions précises. Si beaucoup de travail a été réalisé sur les ligands HLA-I et HLA-II non modifiés, on ne sait que peu de choses sur la présentation des peptides phosphorylés, en particulier par les molécules HLA-II. De plus, aucun des modèles *in silico* existants pour la prédiction des interactions HLA - ligands n'est spécifiquement entraîné sur les ligands phosphorylés.

Cette thèse présente des analyses détaillées sur les ligands HLA-I et HLA-II phosphorylés et introduit des prédicteurs pour les interactions HLA - ligands phosphorylés. La première partie de cette thèse comprend la curation des ligands HLA-I phosphorylés provenant de plusieurs études peptidiques de spectrométrie de masse, identifiant plus de 2'000 peptides phosphorylés uniques couvrant 72 allèles HLA-I. De plus, il a été constaté que les ligands HLA-I phosphorylés sont obtenus par une combinaison de motifs de liaison aux HLA-I, de propriétés intrinsèques de liaison entre les HLA-I et les ligands phosphorylés et de motifs de kinases. La combinaison de ces ligands HLA-I phosphorylés avec des données de ligands non modifiés pour l'entraînement du prédicteur a permis d'améliorer les prédictions des ligands HLA-I phosphorylés.

La deuxième partie de cette thèse porte sur les ligands HLA-II phosphorylés qui sont présentés par des cellules présentatrices d'antigènes professionnelles pour l'activation des lymphocytes T CD4+. Les données peptidiques de HLA-II basées sur la spectrométrie de masse ont permis d'identifier des motifs de liaison pour plus de 30 allèles HLA-II, comprenant 2'473 ligands phosphorylés uniques. Ces motifs ont été utilisés pour re-entraîner un prédicteur des interactions entre les ligands et HLA-II qui a montré une meilleure précision pour les ligands phosphorylés. En outre, l'analyse du peptidome HLA-II phosphorylé a révélé un répertoire plus diversifié de kinases responsables de la phosphorylation des peptides présentés par les HLA-II par rapport aux HLA-I.

En résumé, cette thèse présente des études détaillées sur les ligands HLA phosphorylés ainsi que des outils bio-informatiques pour la prédiction des interactions des peptides phosphorylés avec les molécules HLA-I et HLA-II.

List of Figures

Figure 1.1: Schematic representation of heterodimeric HLA molecules. 24

Figure 1.2: Antigen presentation pathway of HLA-I and HLA-II ligands. 25

Figure 1.3: Binding pockets of HLA molecules. 26

Figure 1.4: Overview cancer antigens in cancerous and normal cells. 29

Figure 1.5: MAPK pathway and disruption in cancer. 34

Figure 3.1: MS-based HLA-II peptidomics identifies multiple phosphorylated HLA-II ligands. 49

Figure 3.2: Phosphorylated peptides bind to HLA-II molecules with specific motifs. List of alleles with phosphorylated peptides. 50

Figure 3.3: Phosphorylated residues show positional specificity in HLA-II ligands. 53

Figure 3.4: HLA-II ligands are phosphorylated by a broader repertoire of kinases than HLA-I ligands. 55

Figure 3.5: The HLA-II phosphopeptidome improves prediction of phosphorylated HLA-II ligands. 56

Abbreviations

APC	Antigen presenting cells
ANN	Artificial neural network
ATP	Adenosine triphosphate
AUC	Area under the curve
BCR	B cell receptor
BRAF mut	Mutant melanoma antigen BRAF^{V600E}
CAR	Chimeric antigen receptor
CD	Cluster of differentiation
CLIP	Class II associated li peptide
CT	Cancer testis
DC	Dendritic cell
EGFR	Epidermal growth factor receptor
EM	Expectation-Maximization
ER	Endoplasmic reticulum
ERAD	ER-associated protein degradation system
FDR	False discovery rate
HLA	Human leukocyte antigen complex
HLA-I	HLA class I
HLA-II	HLA class II
HPV	Human papillomavirus
IEDB	Immune Epitope Database
GF	Growth factor
GTP	Guanosine-5′-triphosphate
Ii	Invariant chain

IP	Immunoaffinity purification
KLD	Kulback-Leibler divergance
LC-MS	Liquid chromatography – mass spectrometry
LC-MS/MS	Liquid chromatography – tandem mass spectrometry
MAPK	Mitogen-activated protein kinase
MHC	Major histocompatibility complex
MIIC	MHC class II compartment
MS	Mass spectrometry
NK cell	Natural killer cell
NSCLC	Non-small cell lung cancer
PLC	Peptide loading complex
pMelan-A/pMART-1	Phosphorylated Melan-A/MART-1
PRR	Pattern recognition receptor
PTM	Post-translational modification
PTPRD	Phosphatase receptor protein tyrosine phosphatase delta
ROC	Receiver operating characteristic
RTK	Receptor tyrosine kinase
TAA	Tumor-associated antigen
TAP	Transporter associated with antigen presentation
TCR	T cell receptor
TKI	Tyrosine kinase inhibitor
TSA	Tumor-specific antigen
PKA	Protein kinase A
PKB	Protein kinase B
PWM	Position weight matrix

Chapter 1 Introduction

The work presented in this thesis aims to contribute to the advancements of epitope predictions for vaccine and immunotherapy development by combining existing prediction methods with a new approach of including phosphorylated immunopeptidomics data. In this first chapter, the immune system is briefly introduced (Section 1.1), followed by a detailed discussion of antigen presentation and cancer antigens (Section 1.2). Thereafter, an overview over existing computational tools for epitope prediction is provided, including the most recent advances in the field (Section 1.3). Lastly, post-translational modification, in particular phosphorylation, and its connection to malignancies, as well as previous studies on antigen presentation and recognition of phosphorylated peptides are discussed (Section 1.4). The introduction concludes with a detailed description of the aims and objectives of this thesis (Section 1.5).

1.1 Immune System

The human body has developed efficient mechanisms to protect itself against intrinsic and extrinsic threats. Intruding pathogens or arising malignancies can be detected and eliminated by the immune system, which is composed of different cell types and molecules. Immune cells emerge from hematopoietic stem cells developed in the bone marrow and further differentiate into myeloid and lymphoid progenitors. Myeloid progenitor cells produce among others neutrophils and monocytes, the latter further differentiating into dendritic cells (DC) and macrophages. B cells, natural killer (NK) cells, and T cell progenitors are derived from the lymphoid lineage and T cell progenitors can further develop memory, cytotoxic, and helper T cells. Matured immune cells from the myeloid and lymphoid lineages are released into the blood or lymphatic system for immune surveillance and potential immune responses in periphery. The immune system can be divided into the innate and the adaptive immune system (extensively reviewed in (1)). Innate immunity provides a first line of defense for the human body against potential pathogens. Skin and other epithelial tissues, such as lung or gut epithelium with mucosal surfaces, act as a first physical barrier for pathogens (2). Pattern recognition receptors (PRRs) found in different subcellular compartments, including cellular and endosomal membranes, as well as in the bloodstream and interstitial fluids, can recognize pathogen-associated molecular patterns frequently conserved in pathogens. Recognition of PRRs causes activation of different immune cells, such as dendritic cells and granulocytes, which induce phagocytosis of the pathogen as well as inflammation (3, 4). The innate immune response is fast but unspecific. Contrarily, adaptive immunity is a slow, but highly pathogen-specific response. The adaptive immune response is initiated through a process called antigen presentation which activates a response cascade involving B and T cells targeting the infected cells. Recognition of presented antigens by B cell receptors (BCRs) initiates further differentiation of B cells into plasma cells, producing and releasing antigen-specific antibodies (5). T cell receptors (TCRs) on the surface of T cells can directly interact with antigens and produce antigen-specific TCRs. Two types of T cells are active in antigen recognition expressing different cluster of differentiation (CD) co-receptors alongside the TCRs. Cytotoxic T cells, characterized by their CD8 co-receptors (alias CD8+ T cell), are responsible for elimination of infected cells, and helper T cells with CD4 co-receptors (alias CD4+ T cell) play a role in the activation of other immune cells such as CD8+ T cells or macrophages (1, 6, 7). It was also observed that PRRs recognizing pathogen-associated molecular patterns can release signals controlling adaptive immunity (8). Furthermore, the adaptive immune system can develop immunological memory after first exposure to a pathogen through memory B and T cells, providing a fast response to re-infection with a previously seen pathogen (9).

1.2 Antigen Presentation

Antigen presentation is a crucial part in the fight against infected or malignant cells whereby short protein fragments are presented on the cell surface for T cell recognition (Section 1.2.1). The pool of presented peptides may be influence by the pathological state of a cell such as viral infections or malignancies like cancer (1.2.2).

1.2.1 Human Leukocyte Antigen System

Proteins of the major histocompatibility complex (MHC) are among the most important players in the protection against pathogens and malignancies. They present short peptide sequences, so-called antigens, on the cell surface to surrounding immune cells. T cells can recognize if a cell presents non-self or self but immunogenic antigens and subsequently initiate an immune response to effectively eliminate the infected or malignant cell. In contrast, healthy cells show a reflection of the proteome on the cell surface and in healthy tissue T cells are tolerant towards such self-peptides (10). In humans, MHC is encoded by genes of the human leukocyte antigen (HLA) system. There are two classes of HLA molecules responsible for antigen presentation, HLA class I (HLA-I) and class II (HLA-II). HLA-I molecules are encoded by three genes positioned on chromosome 6 (HLA-A, -B and -C) and cells are able to express up to six different HLA-I alleles. It was observed that HLA-C alleles are expressed at lower levels than HLA-A and HLA-B alleles, due to several reasons including (post-)transcriptional control (11, 12). Furthermore, genes coding for HLA-I are one of the most polymorphic genes in the human genome, resulting in a big variety of alleles for each gene and producing a huge variability within the human population. Up to now, approximately 20,000 alleles coding for more than 12,300 proteins were identified (13). HLA-II molecules are encoded by three pairs of genes on chromosome 6 (HLA-DPA/B, HLA-DQA/B, and HLA-DRA/B), which form heterodimers and can produce up to 12 different alleles. Polymorphism in all HLA-II genes except HLA-DRA have produced a pool of more than 4,800 known HLA-II proteins encoded by more than 7,400 different alleles (13). Additionally, the two classes of HLA molecules differ in the cells they are expressed by, in the antigen presentation pathway as well as in T cell recognition (14). Almost all nucleated cells express HLA-I, while HLA-II expression is restricted to so-called professional antigen presenting cells (APCs) including dendritic cells, macrophages, and B cells. The peptide repertoire of HLA-I constitutes mainly peptides resulting from intracellular proteins and the HLA-I – peptide complex is recognized by CD8+ T cells. In contrast, peptides that are presented by HLA-II are resulting from proteins of the extracellular compartment that are digested through endocytosis and are presented for recognition by CD4+ T cells. Noteworthy, studies have shown that through a mechanism called cross-presentation HLA-I molecules can also present peptides that underwent the endocytic pathway (15, 16). Additionally, it was observed that a subset of peptides of the HLA-II peptide repertoire can result from intracellular proteins, processed for instance by autophagy (17), reflecting the complexity of the antigen presentation machinery.

The HLA-I Antigen Presentation Pathway

HLA-I is a heterodimer synthesized in the endoplasmic reticulum (ER) from an α polypeptide chain (α1 and α2 chains build up the binding region of the HLA and a transmembrane α3 chain) and complexed with β_2-microglobulin (see Figure 1.1A). A peptide makes up the final component and provides the stability for the HLA-I. Prior to peptide binding, the HLA-I molecule is stabilized through the so-called peptide loading complex (PLC), consisting of the HLA-I molecule, the transporter associated with antigen presentation (TAP) complex, two chaperones ERp57 and calreticulin as well as an additional chaperone called tapasin, which is playing a direct role in the peptide loading to the HLA (18–20) (see Figure 1.2A). Peptides are acquired through the antigen presentation pathway. This is initiated by proteasomal degradation of cytosolic or nucleic proteins into short peptide fragments. These peptides are translocated into the ER through TAP and can be further trimmed by aminopeptidases ERAP1 and ERAP2 in the ER lumen (21, 22). Tapasin regulates the binding of high-affinity peptides with suitable length and peptide sequence to the binding pocket of the HLA-I molecule (23) and the HLA-I – peptides complex is transported out of the ER lumen through the Golgi apparatus to the cell surface for antigen presentation to CD8+ T cells. Misfolded HLA-I molecules, that failed peptide binding for antigen presentation, are transported back into the cytosol for degradation by the ER-associated protein degradation (ERAD) system (24). While the main cleavage enzyme for the HLA-I peptidome is the proteasome, other non-proteasomal degradation pathways are also known to play a role in HLA-I antigen presentation, such as antigen cleavage by the insulin-degrading enzyme in the cytosol (25) or proteases of the endocytic pathway (26).

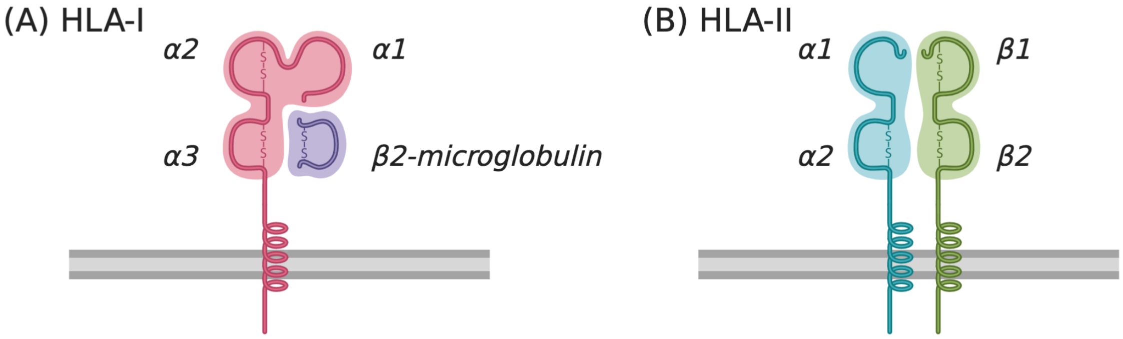


*Figure 1.1: **Schematic representation of heterodimeric HLA molecules.** (A) Membrane – bound HLA-I molecules are made up of α1, α2, and α3 chains in complex with β_2-microglobulin. α1 and α2 make up the binding region of the HLA-I molecule and α3 contains the transmembrane element. (B) Two chains (α and β) form the heterodimeric HLA-II molecules. Subdomains α1 and β1 contain the binding region for HLA-II peptides and α2 and β2 contain transmembrane regions of the HLA-II molecule. [Created with BioRender.com]*

The HLA-II Antigen Presentation Pathway

Heterodimeric HLA-II molecules consist of an α and β chain made up by two domains each (α_1/β_1 domains with the peptide binding pocket and transmembrane α_2/β_2 domains) (see Figure 1.1B). During synthesis in the ER, HLA-II molecules are paired with an invariant chain (Ii), preventing early peptide binding (27–29), and the complex is transported to the endosomal MHC class II compartment (MIIC) (see Figure 1.2B). In the MIIC, Ii is digested by cysteine proteases cathepsins S and L into a shorter peptide, the class II associated Ii peptide (CLIP) (30). Prior to peptide loading into the HLA-II binding pocket, proteins are processed by the HLA-II presentation pathway. Initially, extracellular proteins are taken up into the cell through endosomal ingestion, fragmented into shorter peptides by endosomal proteases, and transported to the MIIC for binding with HLA-II (31). With the help of HLA-DM, CLIP is removed from the HLA-II binding pocket and substituted with a higher affinity peptide (32). Thereafter, vesicles translocate the HLA-II – peptide complex into the plasma membrane for presentation on the cell surface to CD4+ T cells.

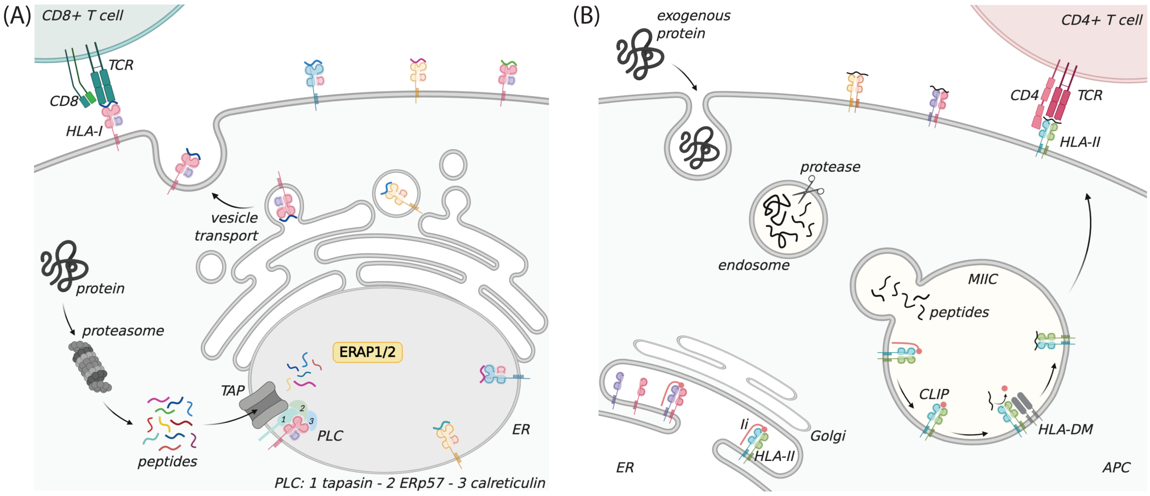


Figure 1.2: Antigen presentation pathway of HLA-I and HLA-II ligands. *(A) HLA-I antigen presentation pathway starts with proteasomal degradation of intracellular proteins into peptides. Peptides can enter the ER through TAP, can be further digested by ERAAP, and are complexed with HLA-I molecules with the help of tapasin, ERp57, and calreticulin. HLA-I – ligand complexes are transported to the cell surface for recognition by CD8+ T cells. (B) In HLA-II antigen presentation exogenous proteins enter APCs through endocytosis and are transported to the MIIC. HLA-II molecules are synthesized in the ER, complexed with Ii, and translocated to the MIIC. In the MIIC, Ii is reduced to CLIP and with the help of HLA-DM CLIP is substituted with a suitable peptide. HLA-II – ligand complexes are transported to the cell surface and can be recognized by CD4+ T cells. [Antigen presentation pathways are depicted according to Neefjes et al.* (14) *and created with BioRender.com.]*

Finally, the binding pockets of HLA-I and HLA-II molecules show different characteristics. While HLA-I is composed of a closed binding pocket, limiting the size of peptides bound to the allele to 8 to 15 amino acids (33, 34), HLA-II have open binding pockets (see Figure 1.3A, B, D, E). The average length of peptides bound by HLA-II was observed to be between 12 and 20 amino acids (35–37). Furthermore, binding of peptides to HLA-I is shaped by two anchor positions P2 and PΩ and these positions are highly conserved in the peptide repertoire of most HLA-I alleles (see Figure 1.3A-C). Though, the HLA-I binding pocket is known to be closed, it was suggested that longer peptides, known to bind with a bulge (38), can also show C′-terminal extensions (39). The open binding pocket of HLA-II facilitates the binding of longer peptides with a peptide binding core of 9 amino acids (see Figure 1.3D). Within this binding core, there are two main anchor positions (P1 and P9) and additional secondary anchor positions that depend on the allele (mainly P4, P6, and P7) (see Figure 1.3D-F). It was estimated that approximately two percent of HLA-II ligands bind with a shorter or longer binding core (8- and 10-mers) to the allele (35). This could also be confirmed in another HLA-II immunopeptidomics dataset, where binding cores were observed to either be reduced to an 8-mer core or extended N- or C-terminally, while keeping anchor positions of the standard 9-mer binding core (37).

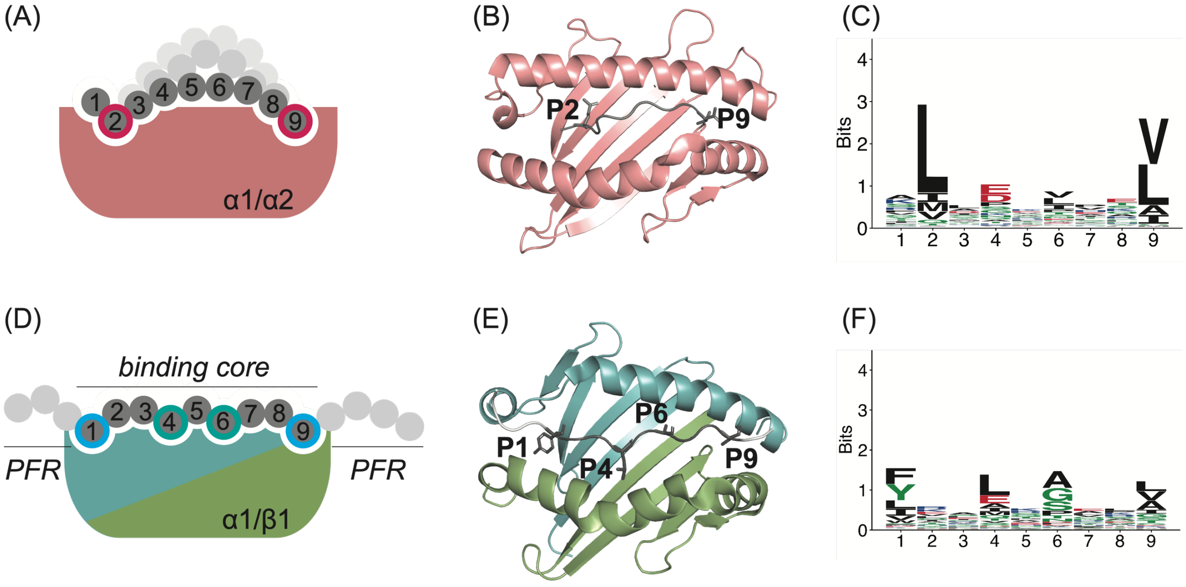


Figure 1.3: Binding pockets of HLA molecules. *(A) Schematic representation of HLA-I binding pockets with a 9-mer ligand in dark grey, anchor positions P2 and P9 (pink circle), and in light grey shades 10- and 11-mer peptides binding with a bulge. (B) Crystal structure of HLA-A*02:01 allele with a phosphorylated ligand (PDB accession code 4NNX). (C) Binding motif of HLA-A*02:01 represented by sequence logo of HLA-A*02:01 peptide repertoire. (D) Schematic representation of the binding pocket of HLA-II molecules, here depicted binding a 16-mer ligand. The 9-mer binding core (dark grey), peptide flanking regions (PFR, light grey), and main (blue circle) and secondary (turquoise circle) anchor positions are shown. (E) Crystal structure of HLA-DRB1*01:01 in complex with a phosphorylated ligand (PDB accession code 3L6F). (F) Binding motif of HLA-DRB1*01:01. Sequence logos, a graphical representation of aligned*

sequences with heights of letters corresponding to amino acid frequencies per position, were drawn using the R package ggseqlogo (40).

The importance of antigen presentation for immune responses also is one of the critical points for infected or malignant cells. Many studies have presented evidence of direct connection between development of autoimmune diseases, susceptibility to viral infections, or progression of cancer and HLA alleles, HLA expression, and antigen repertoire (41). For instance, CD8+ T cell killing of islet beta cells in type 1 diabetes was shown to be attributed to the recognition of an HLA-A*02:01 – presented glucose-sensitive preproinsulin peptide (42). Among others, HIV patients with high expression levels of HLA-C alleles resulting from a genetic variation 36 kilobase pairs upstream of the HLA-C genes, showed slower disease progression compared to patients with low-expressing HLA-C alleles, likely due to better antigen presentation (43, 44). HLA genotypes were also seen to correlate with cancer susceptibility connected to HLA – presentation of mutations (45), immune evasion through cancer – induced loss of heterozygosity (46), response to immunotherapies (47), or overall survival (48).

1.2.2 Cancer Antigens

Malignancies such as cancer can influence the antigen repertoire presented by HLA-I and HLA-II molecules on the cell surface. Antigens generated by cancer can be categorized into two groups: tumor-associated antigens (TAAs) which are antigens that have a low tumor specificity and tumor-specific antigens (TSAs) also known as antigens with high tumor specificity (see Figure 1.4) (49). The latter describes antigens that are only seen on cancer cells caused by viral infection or cancer-specific genomic or proteomic alterations while TAAs can also be found on other cells but show specific features in cancer. For instance, TAAs resulting from overexpression are caused by aberrant gene expression that results in higher levels of presented antigens compared to normal cells. Another group of TAAs are differentiation antigens derived from specific proteins expressed in the specific tissue of origin from which the cancer cell originates. TSAs are grouped into cancer testis (CT), viral, and mutated antigens and have not been seen by the immune system before (so-called neoantigens). CT or cancer germline antigens result from cancer germline genes and have been found in various cancer types (50). These genes are normally not expressed in normal tissues except in germ cells which do not express HLA, thus making cancer-specific CT antigens potential neoantigens (51). Gene expression is regulated by epigenetic modifications, such as methylation, and epigenetic changes have been observed in cancer (52). Furthermore, demethylation has been connected to expression of CT antigens in different cancers (53). Oncogenic viruses, such as the human papillomavirus (HPV) known to cause cervical carcinomas, insert DNA or RNA into cells resulting in cytosolic presence of viral proteins that are processed by the antigen presentation pathway (54). This results in viral non-self antigens that are unknown to the surrounding immune cells if no prior infection and recognition of these antigens occurred. Lastly, cancer-specific non-synonymous mutations can either produce new peptide sequences that were not able to bind to any of the HLA alleles prior to the amino acid change or result in a peptide that could be bound and presented by HLA without the mutation but, due to the new amino acid, provides an unknown epitope for T cell – recognition (55). Besides mutations, neo-epitopes can also contain cancer-driven frameshift mutations (DNA insertion or deletion) resulting in novel protein sequences recognized as non-self, which were seen in various cancer types such as colorectal cancer or leukemia (56, 57). Additionally, it is known that the proteasome, the main cleavage enzyme of the HLA-I peptidome, can splice peptides and thereby create an additional source of potential neoantigens. It was suggested that these spliced peptides can make up between 13 to 45 percent of the whole peptidome in studied cell lines (58, 59). This caused a controversial discussion in the field and different follow-up studies showed that these numbers were likely overestimated and that that spliced peptides constitute merely between 1 to maximal 11 percent of the HLA-I peptidome (60–62).

Antigens with high tumor specificity are attractive candidates for the development of immunotherapeutic strategies. However, direct identification of antigens and in particular detection of neoantigens is still challenging. Pipelines including high-sensitivity experimental workflows and cancer sequencing to identify cancer-specific genomic or proteomic alterations in combination with *in silico* models for prediction of HLA – ligand interactions have been introduced for (neo-)antigen detection (64–67).

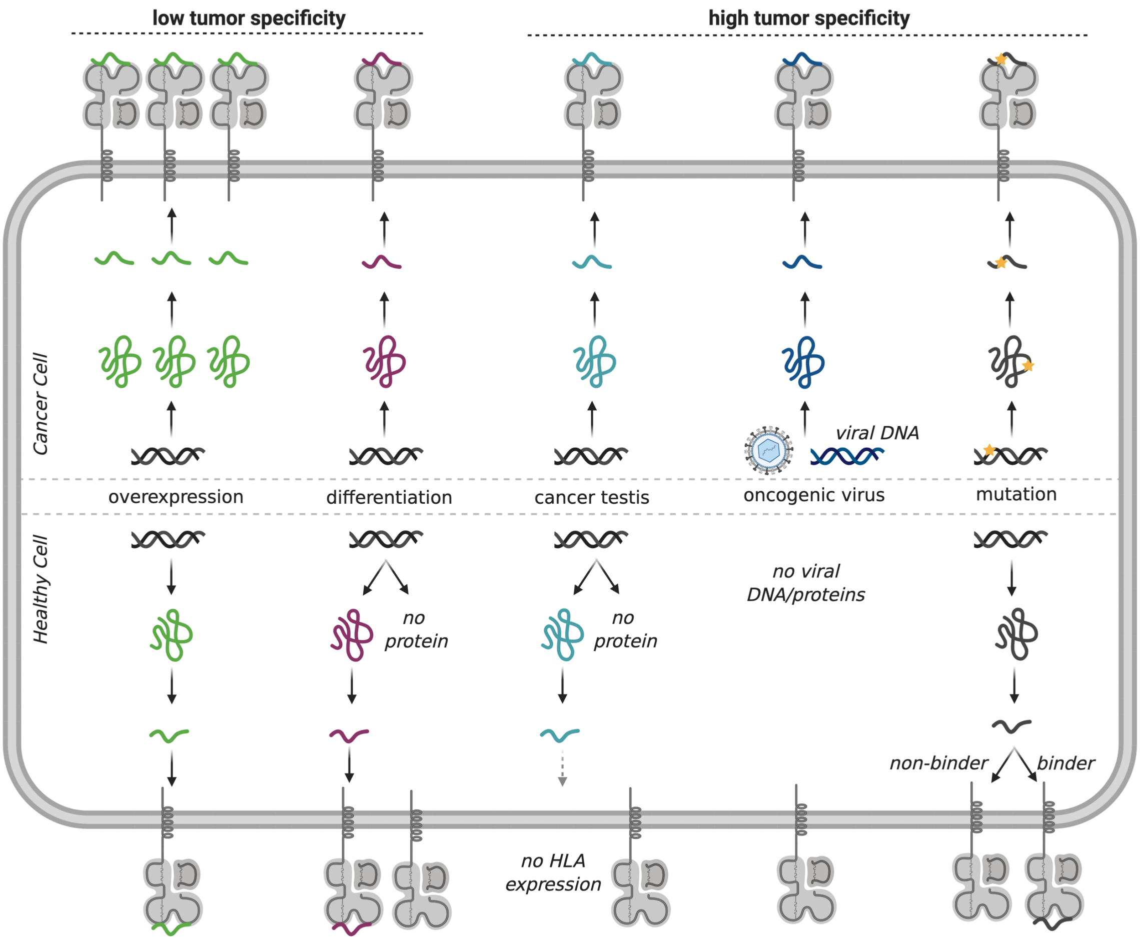


*Figure 1.4: **Overview cancer antigens in cancerous and normal cells.** Cancer antigens with low tumor specificity can result from gene overexpression or differentiation, resulting in a different presentation of such antigens in cancer cells than in healthy cells. Antigens with high tumor specificity can result from (1) expression of cancer testis antigens that are normally expressed in cells not expressing HLA or not expressed in healthy cells, (2) oncogenic viruses, or (3) non-synonymous mutations that either create an antigen previously unable to be presented or give rise to a novel, unknown sequence for in a binder. [Cancer antigens adapted from N. Vigneron 2015* (63) *and created with Biorender.com]*

1.3 *In silico* Models to Predict HLA – Ligand Interactions

Identification of HLA – ligand interactions is essential to understand mechanisms of the immune system such as antigen processing and presentation as well as recognition and elimination by T cells and furthermore help to identifying important targets for cancer immunotherapies or vaccines.

HLA alleles have unique peptide binding repertoires, depicted by their binding motifs (see Figure 1.3C and F), and different methods have been established for identification of HLA ligands. The HLA peptide repertoire, also known as the HLA peptidome or HLA immunopeptidome, can be identified using *in vitro* binding assays (68, 69) or by mass spectrometry (MS) of naturally presented HLA ligands (70–73). While identification of HLA ligands by binding assays requires a priori determination and synthesis of sequences for testing, thereby limiting and biasing screening of the binding repertoire, and furthermore only considers the binding affinity of the peptide to the alleles, MS immunopeptidomics data of naturally presented ligands also captures other stages of the antigen presentation pathway, such as proteasomal processing, transportation, and binding stability of the HLA – ligand complex. In recent years, technical advances established MS protocols, for instance based on immunoaffinity purification (IP), for high throughput, unbiased, and one-experiment identification of big sets of HLA ligands (73). Briefly, HLA – ligand complexes are eluted by IP followed by high resolution liquid chromatography – mass spectrometry (LC-MS) or liquid chromatography – tandem mass spectrometry (LC-MS/MS) to separate and analyze peptide sequences. Resulting MS spectra are then analyzed to identify sequences using *in silico* methods that compare MS spectra with computationally determined spectra of a reference database, such as the human proteome.

One major challenge of MS-based immunopeptidomics data is to determine allelic restriction of HLA molecules expressed in a specific sample or cell line. This can be avoided by using mono-allelic cell lines, which are cell lines engineered to express only one allele, hence MS spectra from isolated HLA – ligand complexes of these cells correspond to the binding repertoire of one HLA allele (70, 71). However, antigen presentation might not be naturally reflected in mono-allelic cell lines since competition to bind a specific peptide between different alleles is lost and alleles potentially present lower affinity peptides (74). Processing of natural, unmodified multi-allelic cell lines or tissue samples of patients requires efficient and precise tools to assign the HLA peptidome to the corresponding allele. Different computational mechanisms have been established to solve this by assigning each peptide to its respective allele using predictors for HLA – ligand interactions or unsupervised clustering of the HLA peptidome to determine binding motifs.

1.3.1 Predictors for HLA Ligand Interactions

To reduce MS identification of HLA peptidomics data that requires effort- and cost-intensive experimental work, *in silico* prediction methods for HLA – ligand interactions provide a faster and cheaper solution. HLA peptidomics data is accessible in databases such as the immune epitope database (IEDB) (75) and present a powerful resource for training of computational models. Initially, predictors for HLA – ligand interaction were mainly trained on *in vitro* binding affinity ligands (76–78). With increasing accessibility of eluted HLA-I and HLA-II ligands as well as efficient tools to determine allelic restriction, predictors solely trained on eluted HLA ligands (33, 37, 70, 71, 79, 80) or on a combination of binding affinity and eluted ligands (81–88) were developed. It was shown that predictors trained on eluted, naturally presented HLA ligands identified HLA binding motifs more accurate than predictors trained on binding affinity data (81).

Different predictors for HLA – ligand interactions were developed, among others motif – based predictors such as SYFPEITHI (89) or MixMHCpred (33, 90) as well as more complex machine learning models using neural networks, such as MHCFlurry (86, 87), MSIntrinsic (70), or different NetMHC predictors (88, 91–94). On the one hand, prediction methods can be trained separately for specific alleles, producing robust results for alleles with known peptide binding repertoires (37, 70, 71, 77, 79, 80). On the other hand, with the introduction of pan-specific models, which are trained simultaneously for multiple alleles, predictions could also be performed for alleles with little or unknown HLA binding repertoires (76, 78, 85–87). This is of particular interest for HLA – ligand interactions given the highly polymorphic HLA genes. Overall, it was shown that introducing pan-specific predictors improves prediction accuracy for HLA – ligand interaction (95, 96).

The model NNAlign can use any kind of receptor – ligand data, such as HLA – peptide interactions, to identify sequence alignment and binding motifs and subsequently build an artificial neural network (ANN) framework for predictions of new ligands. NNAlign is the framework for all NetMHC HLA – ligand predictors (97). The ANN of the pan-specific HLA-I – ligand predictor NetMHCpan is trained on MS as well as *in vitro* binding affinity ligands in combination with sequences of HLA molecules (94). Similarly, NetMHCII (92) and NetMHCIIpan (88), predicting HLA-II – ligand interactions, are allele- and pan-specific ANN – based models trained on human and murine data. MARIA, a recurrent neural network model, is trained on binding affinity as well as eluted MS peptidomics data in combination with expression levels of antigen source genes as well as signals of protease cleavage (83). For neonmhc2, the authors translated HLA-II ligands into amino acid proximity matrices. These were used to train a convolutional neural network separately for each allele in combination with additional binary encoded features on each amino acid, such as hydrophobicity, amino acid charge, and position in the peptide sequence (71). Models based on HLA binding motifs use position weight matrices (PWMs), a mathematical way of describing sets of aligned sequences by representing each position by amino acid frequencies (i.e. a 9x20 matrix for a set of aligned 9-mer sequences over the alphabet of the 20 canonical amino acids), to describe HLA binding motifs. In addition to PWMs, MixMHC2pred further relies on information of the N- and C-terminus of HLA-II ligands by building allele-independent N-/C-terminal motifs of the HLA-II peptidome as well as peptide length distribution and peptide binding cores to train an allele-specific predictor (37).

1.3.2 Identification of Binding Motifs in HLA Peptidomics Data

DNA or protein binding sites are often characterized by very specific motifs and identification of these motifs has been an important and crucial issue to better understand how biological processes and signaling is monitored and regulated (98, 99). Efficient computational methods have been developed enabling identification of transcription factor binding sites or protein – protein and protein – ligand interaction sites without heavy experimental work of testing the binding of potential ligands (100–102). Expectation maximization (EM) algorithms, such as MEME (103, 104), probabilistic frameworks based on PWMs (105, 106), or neural network models (107) are widely used methods to detect patterns in unaligned DNA or protein sequences.

Unsupervised clustering of HLA peptidomics data is an efficient tool to identify binding motifs of HLA alleles without prior knowledge of the binding motif (33, 37, 79, 108, 109). This is particularly useful to apply to samples containing more rare alleles without a widely established peptide binding repertoire. For instance, GibbsCluster uses the Gibbs sampling approach to align and cluster peptides simultaneously and can be applied, among other peptide data, to HLA-I or HLA-II peptidomes (108, 109). In short, GibbsCluster groups peptides into clusters, initially starting with a random cluster assignment, and in each iteration optimized clusters are found by (1) realigning peptide binding cores within a cluster, (2) moving peptides between clusters, or (3) shifting the binding core of an optimally aligned cluster to receive the best binding core. Using Kulback-Leibler divergance (KLD), the distance within peptide clusters is minimized and at the same time maximized between peptide clusters to identify the optimal alignment and clustering of the peptide data. For peptide sequences that do not match any of the clusters inferred by the algorithm, GibbsCluster further builds a so-called trash cluster, which can identify potential falsely identified peptides. Another tool developed to identify binding motifs in HLA-I peptidomics data is MixMHCp, a mixture model – based motif deconvolution method (33, 79). MixMHCp identifies peptide clusters that correspond to HLA-I binding motifs, represented as PWMs, using a log likelihood framework and assigns so-called responsibility values to each peptide describing the likelihood of the peptide belonging to each peptide cluster. Similarly, the probabilistic machine learning model MoDec provides a powerful tool to identify HLA-II binding motifs and binding cores simultaneously with the advantage of not requiring peptide alignment a priori, which is a crucial step in identifying binding motifs of HLA-II ligands due to the preference for longer peptides and the open binding pocket of HLA-II molecules (37). MoDec detects optimal peptide clusters and peptide alignment by determining maximum likelihood using an EM algorithm.

Annotation of identified peptide clusters can be either done manually relying on known binding motifs from previous studies or using fully unsupervised approaches based on Euclidean distances measured for PWMs of alleles (90) or KLD between different PWMs (37). These unbiased and automatic methods make use of co-occurring alleles between samples and can be applied to clusters of HLA binding motifs resulting from two or more different samples that (1) contain the same alleles except for one (to identify motifs of alleles that are only present in one of the samples) or (2) have only one allele in common, which can be used to identify the unknown motif of an allele by finding re-occurring motifs in both samples. These approaches provide effective and unbiased models for the identification of HLA binding motifs from MS – based immunopeptidomics data.

1.4 Phosphorylation

Post-translational modifications (PTMs), such as acetylation, glycosylation, and phosphorylation, are reversible processes on proteins and play a role in establishing the diversity of the human proteome (110). PTMs regulate many cellular processes (111–113), including protein localization, activation, de-activation, and degradation of proteins as well as mediating protein-protein interactions. Among others, glycosylation and ubiquitination are involved in the regulation of proteolysis (114, 115), protein folding is monitored by glycosylation of synthesized proteins in the ER (116), and phosphorylation, one of the most common PTMs, plays an important role in regulating signaling pathways as well as activation and deactivation of many proteins (117–119). For instance, the transcription factor OCT4, known to play an important role in maintaining pluripotency and evading differentiation (120), is known to be regulated by phosphorylation (121). The mitogen-activated protein kinase (MAPK) signaling pathway is an important mediator for cell growth, proliferation, and survival (122, 123). The MAPK pathway is initiated by extracellular growth factors (GF) binding to growth factor receptors, such as the epidermal growth factor receptor (EGFR), and regulates the activity of intracellular receptor tyrosine kinases (RTKs). RTKs initiate the downstream signal transduction cascade of the MAPK signaling pathway which consists of succeeding phosphorylation events and results in the regulation of expression of effector genes (see Figure 1.5A).

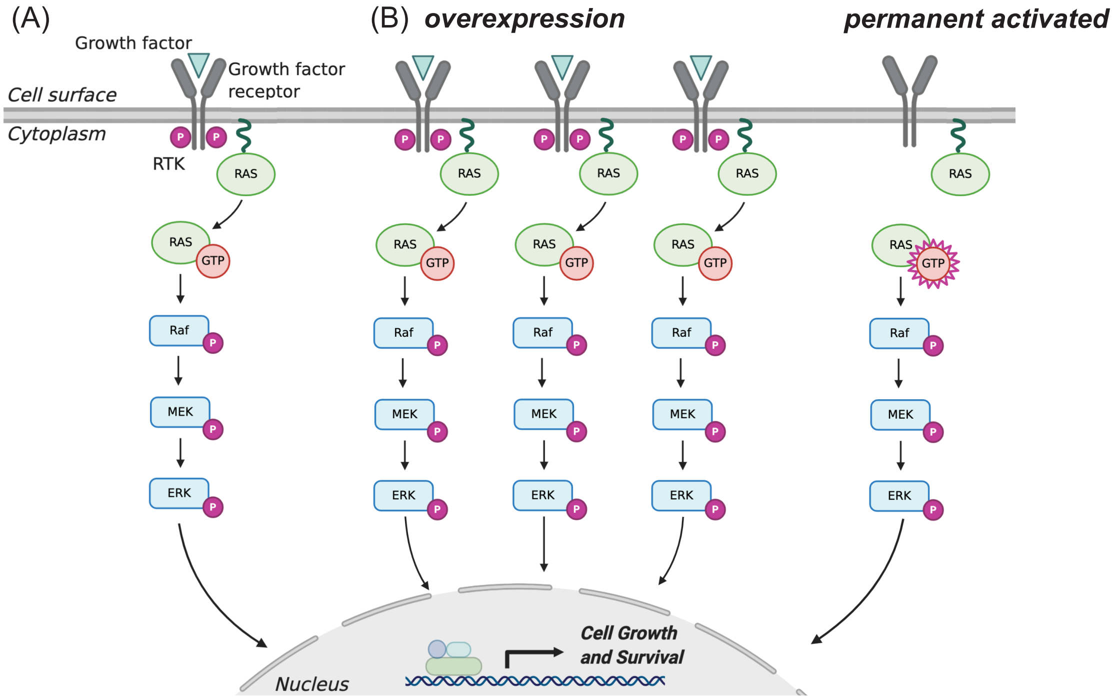


Figure 1.5: MAPK pathway and disruption in cancer. *(A) Simplified representation of phosphorylation-driven MAPK pathway. Activated by binding of GF, intracellular RTK initiate a signaling cascade controlled by phosphorylation events on RAS, Raf, MEK, and ERK and resulting in transcription of cell growth and survival regulators. (B) Disruption of MAPK signaling in cancer through overexpression of growth factor receptors (left) or permanent binding of GTP to RAS resulting in activation of downstream signaling without GF activation.*

It was estimated that ~30% of proteins in the human proteome can contain temporary phosphorylation events (124) and a more recent analysis of the human phosphoproteome *in vivo* suggested that up to three-fourth of all proteins hold phosphosites (125). Protein phosphorylation is the ligation of an additional phosphate group from adenosine triphosphate (ATP) to the hydroxyl group of serine, threonine, and tyrosine residues and is performed by protein kinases. Furthermore, kinase activity itself is regulated by phosphorylation (126). More than 500 different kinases are known and it was estimated that their coding genes make up approximated two percent of the human genome (127). Kinases show distinct substrate specificity, such as serine/threonine or tyrosine kinases, and further depend on distinct substrate binding motifs, for instance the proline-dependent serine/threonine kinase MAPK requiring a proline adjacent to the substrate residue ([pS/pT]P). Dephosphorylation, the enzymatic cleavage of the phosphate group from phosphorylated residues, is performed by so-called phosphatases. Thus, phosphatases are important opposing players to kinases with a crucial role in regulating signaling pathways by resuming substrates to their pre-phosphorylation functionality, and in maintaining homeostatic phosphorylation levels in cells (128).

1.4.1 Phosphorylation in HLA Ligands

Cancer-specific alterations resulting in upregulated kinase or downregulated phosphate activity can result in cancer-specific phosphosites, as it could be observed in a proteogenomics study of breast cancer (135). This suggests that phosphorylated HLA ligands should be considered as potential targets for the development of cancer immunotherapeutic strategies. Phosphorylation of serine, threonine, or tyrosine changes the biophysical properties of HLA ligands through the addition of a negatively charged phosphate moiety to the residue. Depending on the position of the phosphorylation in the peptide, this impacts the binding of the peptide to the allele or the surface potential for T cell recognition of the presented epitope.

Phosphorylated HLA-I Ligands

Multiple studies established that phosphorylated residues in peptides do not interfere with the HLA-I antigen presentation machinery and phosphorylated ligands can be naturally processed. Phosphorylated peptides were shown to be bound by HLA-I molecules, transported to the cell surface, and furthermore HLA-I – phosphorylated peptide complexes were recognized by CD8+ cells (136–140). It was observed that the addition of a phosphate moiety to the peptide provides a distinct peptide surface potential for T cell recognition compared to the unmodified version of the peptide (141–143). This as well as further experimental validation led to the conclusion that T cell recognition of phosphorylated HLA-I peptides was both, sequence-specific and phosphosite-dependent, i.e. specific T cells did not recognize other phosphorylated peptides nor unmodified counterparts of the tested phosphorylated peptides (138, 139, 141, 142, 144). This suggest, that phosphorylated HLA-I ligands can show specific immunogenicity and are of potential interest for the development of immunotherapeutic targets (140). This was supported by other work identifying cancer-specific phosphorylated HLA-I ligands. For instance, phosphorylated peptides specifically found in tumor cells but not in healthy cell lines or tissue could be identified (137), the HLA-I phosphopeptidome of melanoma samples was determined (145), and specific immunity in healthy donors against a Leukemia – associated phosphorylated peptide was shown (144). Recently, a first clinical trial was performed on melanoma patients using two phosphorylated peptides in a vaccine (146). Prior to the development and testing of the vaccination, the phosphorylated peptides showed immunogenicity in HLA-A2 transgenic mice *in vivo* as well as in healthy human tissues *in vitro*. These results further outline the importance to include phosphorylated HLA-I ligands in the HLA-I immunopeptidome and therefore consider them in development of cancer immunotherapies.

Additional specific characteristics of phosphorylated HLA-I peptides were identified and validated in many studies. For instance, in 9- to 12-mer ligands the phosphorylated residue was observed to be positioned mainly at P4 of the peptide (141, 144, 145, 147, 148). Further, it was seen that phosphorylated HLA-I ligands often show enrichment of a basic residue at P1 (137, 138, 141, 144, 145, 147, 148). One group analyzed this further by performing crystallography of different phosphorylated HLA-I peptides in both unmodified and phosphorylated versions (141, 142). They saw that phosphorylated residues could enhance the stability of the HLA-I – bound peptides by interaction of the phosphate group with specific residues of the HLA-I binding pocket (Arg66 or Lys65) as well as intermolecular bonds between the basic residue at P1 and the phosphate-moiety at P4. When the conformation of the phosphorylated peptide was compared to its unmodified counterpart, two out of three phosphorylated peptides showed distinct conformational changes compared to the unmodified peptide, suggesting that this could help the binding of the phosphorylated peptide and explain higher binding affinity observed for peptides with phosphorylation compared to unmodified versions of the peptides (142). Furthermore, enrichment of proline next to phosphorylated serine and threonine residues was detected in phosphorylated HLA-I ligands, a signal of the kinase binding motif of proline – dependent serine/threonine kinases such as CDK1 or MAPK1 (138, 141, 144, 145, 147, 149).

Studies on phosphorylated HLA-I ligands mostly identified a limited amount of ligands and predominantly focused on the common HLA-A*02:01 or HLA-B*07:02 alleles (138–144, 150). Some work collected phosphorylated peptides for different mono- or multi-allelic samples (70, 136, 137, 145, 147–149), but overall to date the allelic-coverage as well as the amount of identified phosphorylated HLA-I ligands from these studies are limited.

Phosphorylated HLA-II Ligands

Phosphosites presented on HLA-II ligands have been less well studied than phosphorylated HLA-I peptides and so far, only three studies directly analyzed phosphorylated HLA-II ligands. The first study to identify naturally presented phosphorylated HLA-II ligands used an EBV – transformed B-lymphoblastoid and a melanoma cell line and collected 27 and 20 different phosphorylated HLA-DRB1 peptides in these samples, respectively (137). They saw that source proteins for more than half of these phosphorylated peptides were transmembrane proteins, while the rest resulted from cytoplasmic and nucleic proteins, outlining that phosphorylated peptides can be processed by the endocytic pathway for presentation by HLA-II molecules.

Another study investigated the mutant melanoma antigen BRAF^{V600E} (BRAF mut), a mutation common in more than 60% of melanoma patients (151), and reported CD4+ T cell recognition of the phosphorylated BRAF mut antigen (152). Furthermore, they saw that CD4+ T cells specific against the non-phosphorylated BRAF mut failed to recognize the phosphorylated version of the antigen, while phosphorylation – specific T cells against the BRAF antigen could detect it. However, *in vitro* testing of phosphorylated BRAF mut – specific T cells against melanoma cell lines could not be observed, potentially caused by failed processing or HLA-II – binding of phosphorylated BRAF mut. The study further identified 150 unique phosphorylated HLA-II peptides from four cell lines (two melanoma and two EBV-B cell lines) using MS. Within these phosphorylated peptides, ~30% of the phosphosites were previously reported, all peptides except one contained one phosphosite, and the phosphorylation was distributed with 93.0, 5.3, and 1.7 percent among serine, threonine, and tyrosine residues, respectively. The analysis of source proteins of the HLA-II phosphopeptidome supported evidence that lysosomal HLA-II pathways can process intracellular proteins through autophagy (17, 26). In the two melanoma samples a phosphosite could be seen in the melanocytic antigen Melan-A/MART-1, which in its unmodified version is known to be immunogenic and therefore of interest for immunotherapeutic strategies (153–155). T cell recognition of the phosphorylated Melan-A/MART-1 (pMelan-A/pMART-1) peptide was seen *in vitro* and the specificity of the peptide could be attributed to the phosphorylated serine, as unmodified Melan-A/MART-1 did not result in INF-γ secretion.

Finally, the third study looked into structural properties of phosphorylated HLA-II ligands and performed crystallography of the pMelan-A/pMART-1 antigen (156). Crystal structures of a 15-mer pMelan-A/pMART-1 in complex with HLA-DRB1*01:01 revealed that the peptide's conformation at anchor positions (P1, P4, P6, and P9 for DRB1*01:01) was conserved as expected from unmodified ligands. Furthermore, the phosphorylated residue was positioned at a non-anchor position (P5) and could therefore be in direct contact with the T cell for recognition. The authors tested the binding and recognition of different peptide versions of pMelan-A/pMART-1 with various length and phosphosite position, and concluded that this can affect both, HLA-II binding and T cell recognition of the pMelan-A/pMART-1 antigen. These results illustrate the restrictions for ligand binding imposed by main and secondary anchor positions to HLA-II alleles, which is also crucial for the binding of phosphorylated ligands.

Cancer-specific phosphosites comprised in the HLA-I and HLA-II immunopeptidomes can potentially act as attractive targets for the development of cancer immunotherapies. The presented studies on phosphorylated HLA-I and HLA-II ligands showed the importance of including phosphorylated ligands in the understanding of HLA-I and HLA-II ligand presentation as well as outlines the need to integrate phosphorylated HLA ligands in identification protocols of HLA peptidomes as well as prediction methods for HLA-I and HLA-II ligand interaction.

1.5 Aim and Objective of this Thesis

Little is known about the role of phosphorylated peptides in HLA ligand repertoires, despite the advancements in experimental and computational workflows for HLA ligand identification, the increasing growth of HLA immunopeptidomics data available in public databases as well as the clear evidence of processing, presentation, and T cell recognition of phosphorylated HLA-I and HLA-II ligands. This thesis builds on these findings and aims to contribute to a better understanding of HLA-I and HLA-II phosphopeptidomes as well as provide prediction models for phosphorylated HLA ligands.

Identification of the HLA-I phosphopeptidome and the development of a predictor for phosphorylated HLA-I – ligand interaction

Extensive knowledge on phosphorylated HLA-I ligands are valuable contributions to fully comprehensive understanding of HLA immunopeptidomes and expand the pool of potential immunotherapeutic targets. Until now, binding motifs of phosphorylated HLA-I ligands are undefined, specific phosphorylation – dependent features in HLA-I ligands unknown, and none of the existing tools for HLA-I – ligand prediction are specifically trained on modified sequences, thus potentially lacking relevant information for the prediction of phosphorylated HLA-I ligands.

The first objective of this thesis is to curate MS – based immunopeptidomics data to identify the phosphorylated HLA-I binding repertoire at a high allelic – coverage and further analyze this data to detect specific characteristics in the HLA-I phosphopeptidome. Secondly, this thesis aims to exploit the identified HLA-I phosphopeptidome to develop a predictor for HLA-I – phosphorylated ligand interactions.

Understanding and prediction of phosphorylated HLA-II ligands

The second core aim of this thesis focusses on a better understanding of phosphorylated ligands presented by HLA-II molecules on professional APCs. As presented above in section 1.4.1, only few studies worked on phosphorylated HLA-II ligands, thus to date little is known about the HLA-II phosphopeptidome. Taking advantage of HLA-II immunopeptidomics data of multiple samples, phosphorylated HLA-II peptides will be collected and analyzed for the identification of phosphorylated HLA-II binding motifs as well as phosphorylation – specific characteristics in the HLA-II phosphopeptidome. The final objective is to retrain a model for HLA-II – ligand interaction specifically on the curated HLA-II phosphopeptidome for prediction of phosphorylated ligands.

Overall, defining the space of HLA-I and HLA-II phosphopeptidomes contributes to a better understanding of HLA presentation of phosphorylated ligands. This enables the development of prediction models specifically trained on phosphorylated ligands, providing tools for future studies on T cell epitope identification in infections and malignancies.

Chapter 2 Manuscript "Mass Spectrometry – Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands"

The first part of my thesis comprised the identification and analysis of phosphorylated HLA-I ligands and the development of a prediction method for HLA-I – phosphorylated peptide interactions. This work was published in *Molecular and Cellular Proteomics* in February 2020 and is attached in its published version as Appendix A. Figure references in this chapter refer to the original article.

Authors and affiliations

Marthe Solleder[1,2], Philippe Guillaume[1], Julien Racle[1,2], Justine Michaux[1,3], Hui-Song Pak[1,3], Markus Müller[2], George Coukos[1,3], Michal Bassani-Sternberg[1,3], and David Gfeller[1,2]

[1] Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University of Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3] Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University Hospital of Lausanne, Lausanne, Switzerland

Author Contributions

The experimental parts of this study were performed in collaboration with the group of Dr. Michal Bassani-Sternberg of the *Human Integrated Tumor Immunology Discovery Engine* at the Department of Oncology UNIL CHUV, who conducted MS analysis of new samples and further curated publicly available immunopeptidomics samples to identify phosphorylated peptides. Furthermore, binding assays of HLA-I – ligand interactions were performed by Dr. Philippe Guillaume.

The computational pipeline to analyze the phosphorylated HLA-I peptidomics data, including curation of phosphorylated HLA-I binding motifs, data analysis, and the development of the predictor, was performed by myself under the supervision and guidance of Prof. David Gfeller. The manuscript was written by myself together with Prof. David Gfeller and Dr. Michal Bassani-Sternberg. Furthermore, Dr. Philippe Guillaume contributed to the manuscript with a detailed description of the experiments he performed.

2.1 Summary of Results

The newly processed HLA-I peptidomics data from six additional samples in this study together with the curation of several publicly available immunopeptidomics studies (33, 72, 145, 157–159) resulted in the identification of 2,190 unique phosphorylated peptides. To determine allelic – restriction of identified HLA-I ligands for each of the 61 samples, the previous published motif deconvolution algorithm MixMHCp (33, 79) was expanded to be able to process phosphorylated residues. We could identify binding motifs of phosphorylated HLA-I ligands without a priori information on their interactions with HLA-I alleles through applying MixMHCp to the combined dataset of phosphorylated and unmodified HLA-I ligands. We saw direct similarity in binding motifs of phosphorylated HLA-I ligands compared to unmodified HLA-I ligands, in particular at the second and final position, which are anchor positions of HLA-I ligands (Figure 1). With the addition of phosphorylated HLA-I ligands with known allelic-restriction from previous studies (70, 137, 149, 160–163, 138, 139, 141–144, 147, 148), this work comprised in total 2,066 unique phosphorylated peptide sequences and 2,585 unique HLA-I-phosphorylated peptide interactions with 72 different HLA-I alleles.

The analysis of the HLA-I phosphopeptidome showed a higher frequency of phosphorylated peptides detected in HLA-C alleles (Figure 2A). This is likely explained by human phosphosites fitting binding motifs of HLA-C alleles better than those of HLA-A or HLA-B alleles. This was further supported by the fact that we could also see a higher fraction of phosphosites from the human phosphoproteome (125) in unmodified HLA-C ligands (see Figure 2C). Different characteristics of phosphorylated peptides in the HLA-I peptidome could be observed. We saw a similar length distribution to what was expected from unmodified HLA-I ligands (Figure 2D), a similar distribution of phosphorylated residues compared to the human phosphoproteome (Figure 2D), and a clear preference for phosphorylated residues at P4 (Figure 3A). The latter could be supported by binding assay of multiple peptides showing that other positions lacked good binding to the alleles (Figure 3B). Furthermore, we could confirm previous observations of enrichment of proline next to phosphorylated residues, which is explained by the very frequent binding motif [pS/pT]P of proline-dependent kinases such as MAPK1 (Figure 4A). We further detected a preference for arginine at the first position of phosphorylated HLA-I ligands (Figure 4D). Results of binding assays outlined that this is most likely due to the RXX[pS/pT] binding motif of kinases such as the family of protein kinase A (PKA) or B (PKB) (Figure 4E and F).

Finally, we used the identified and curated binding motifs of phosphorylated HLA-I ligands to train the first predictor for HLA-I interactions with phosphorylated peptides. Training of the model on a combined set of phosphorylated and unmodified ligands outperformed training solely on unmodified ligands, showing that information on phosphorylated residues can improve the predictions of phosphorylated HLA-I ligands (Figure 5).

Chapter 3 Deciphering the landscape of phosphorylated HLA-II ligands

The following chapter presents the first in-depth analysis of phosphorylated HLA-II ligands as well as the first method to predict HLA-II – phosphorylated ligand interactions specifically trained on phosphorylated peptides. The manuscript of this study is currently prepared for submission.

Authors and Affiliations

Marthe Solleder[1,2], Julien Racle[1,2], Philippe Guillaume[1], George Coukos[1,3], Michal Bassani-Sternberg[1,3], and David Gfeller[1,2]

[1] Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University of Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3] Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University Hospital of Lausanne, Lausanne, Switzerland

Author Contributions

Analysis of raw MS data to identify phosphorylated HLA-II peptides was run by Dr. Michal Bassani-Sternberg and binding assays of HLA-II – ligand interactions were performed by Dr. Philippe Guillaume. Data curation, the computational analysis of phosphorylated HLA-II ligands, and the expansion of computational tools for motif deconvolution and prediction of phosphorylated HLA-II ligands, was performed by myself under the supervision and guidance of Dr. Julien Racle and Prof. David Gfeller. The manuscript was written by myself with contributions and revisions by Dr. Michal Bassani-Sternberg, Dr. Philippe Guillaume, Dr. Julien Racle and Prof. David Gfeller.

3.1 Introduction

CD4+ T cells play a central role in adaptive immune responses against viral infections and cancer through the recognition of non-self peptides (i.e., from pathogens) or tumor-specific antigens (i.e., genetic/proteomic alterations in cancer). Antigen presentation to

CD4+ T cells is performed by HLA-II molecules, which are expressed on the cell surface of professional APCs such as dendritic cells or B lymphocytes. HLA-II molecules form heterodimers and are encoded by three pairs of genes (HLA-DRA/B, -DPA/B, -DQA/B). Except for HLA-DRA, these genes are highly polymorphic and thousands of alleles have been discovered in humans. HLA-II molecules bind mostly peptides of 12 to 20 amino acids with a 9-mer peptide binding core (see Figure 3.1A) (14, 36). HLA-II ligands can originate from both exogenous and intracellular proteins processed by endocytic pathways (164). Recently, HLA-II ligands have been shown to play an important role in the response to personalized cancer vaccines (165–168). Most HLA-II ligands are unmodified peptides, although PTMs can also be displayed on HLA-II molecules (169). Being able to identify post-translationally modified HLA-II ligands is therefore promising to expand the range of potential targets for cancer immunotherapy. HLA-II ligands can either be directly identified by MS, although such experiments are technically challenging (170), or using prediction methods followed by experimental validation. Several different predictors of HLA-II ligands have been developed (37, 71, 83, 94) and can contribute to reduce cost and efforts to identify novel HLA-II ligands, including class II neoantigens. However, none of these predictors specifically integrate PTMs.

PTMs of proteins are essential regulators in many biological processes (112, 113, 118). PTMs like phosphorylation were shown to be deregulated in cancer cells, causing aberrant cellular behavior (171–173). Therefore, phosphorylated peptides presented on HLA molecules provide potential targets for the development of immunotherapeutic strategies (137, 146, 150, 174). While many studies analyzed phosphorylated peptides presented on HLA-I molecules (136, 139, 141, 143, 144, 147, 175), phosphorylated HLA-II ligands have received much less attention. The first naturally presented phosphorylated HLA-II ligands were identified from an EBV – transformed B-lymphoblastoid and a tumor cell line (137). Shortly after, the first CD4+ T cell recognition of a phosphorylated HLA-II ligand was shown using the melanoma antigen Melan-A/MART-1 (152). Structural analysis of a phosphorylated peptide bound to HLA-DRB1 showed that the phosphorylated residue can in this case directly interact with the T-cell receptor (156). While these studies provide evidences for HLA-II presentation of phosphorylated peptides and show potential application as targets for immunotherapies, further characteristics such as binding motifs of phosphorylated HLA-II ligands on a large allelic coverage remain unknown and no HLA-II ligand predictor is specifically trained on modified sequences. Recently, we have shown that including the

HLA-I phosphopeptidome in the training of HLA-I ligand predictors could significantly improve the accuracy of HLA-I phosphorylated ligand predictions (175).

In this work, we capitalized on high quality MS HLA-II peptidomics datasets and identified 2,473 novel phosphorylated HLA-II ligands. Based on this data, we defined phosphorylated binding motifs of HLA-II alleles, identified specific molecular properties of phosphorylated HLA-II ligands, and investigated differences in kinase motifs between phosphorylated HLA-II and HLA-I ligands. Furthermore, we developed the first HLA-II ligand prediction method specifically considering phosphorylated peptides and demonstrated improved accuracy.

3.2 Results

MS-based HLA-II peptidomics identifies multiple phosphorylated HLA-II ligands

To identify a broad spectrum of phosphorylated HLA-II ligands across a wide range of HLA-II alleles, we reanalyzed raw MS HLA-II peptidomics data of 23 poly-allelic samples (37) with MaxQuant, allowing for phosphorylation on serine, threonine, and tyrosine as variable modifications (see Methods in 3.4). A total of 2,800 unique phosphorylated peptides were identified. To determine HLA-II allelic restriction, predict binding cores, and remove potential wrongly identified peptides, we expanded the motif deconvolution method MoDec (37) to phosphorylated residues and applied it to the pool of phosphorylated and unmodified HLA-II ligands for each sample (see Figure 3.1B and Methods in 3.4). 327 phosphorylated peptides were assigned to the flat motif and we consider these as co-eluted contaminants or wrongly identified peptides, as expected in HLA-II peptidomics studies (37). To support this hypothesis, we compared the score for peptide spectrum matches from the Andromeda search engine (peptide score, higher values for higher confidence in peptide identification) with the score difference to the second best peptide spectrum match (delta score, higher values for unambiguous distinction from other peptides). The distribution of these two scores for all 2,800 phosphorylated peptides is shown in Figure 3.1C. As expected, phosphorylated peptides that were assigned to the flat motif by MoDec showed lower peptide and delta scores than those that were assigned to other motifs. These peptides were therefore excluded from downstream analyses, similarly to what has been done for HLA-I (175). The remaining phosphorylated HLA-II ligands showed a length distribution similar to the one of unmodified HLA-II ligands (Figure 3.1D) and the majority contained one phosphorylation, with ~16% double phosphorylated and less than 3% triple phosphorylated (Figure 3.1E). Furthermore, phosphorylated residues were observed with 57.13, 28.74, and 14.13% for phosphorylated serine, threonine, and tyrosine, respectively.

Phosphorylated peptides bind to HLA-II molecules with specific motifs.

To assign phosphorylated peptides to their cognate HLA-II alleles and determine binding motifs of phosphorylated HLA-II ligands, we curated the output from MoDec for each sample (see Methods in 3.4 and example in Figure 3.1B). 1,579 unique phosphorylated peptides could be unambiguously assigned to 32 different alleles (including the bispecific binding allele HLA-DRB1*08:01), for a total of 1,644 unique interactions between phosphorylated peptides and HLA-II alleles. Binding motifs of phosphorylated HLA-II ligands showed conserved specificity at anchor residues P1, P4, P6, and P9 for most alleles (see Figure 3.2). The remaining 894 phosphorylated peptides came from motifs that could not be assigned to one specific allele (e.g., ambiguous motifs mixing multiple alleles) and were therefore not considered in allele-specific analyses to minimize the risk of wrong allelic assignment. We observed similar frequency of phosphorylated peptides for different HLA-II genes, with the only exception of the two HLA-DRB4 alleles which had higher fraction of phosphorylated ligands (Supplemental Figure 3.1). This enrichment may be explained by the presence of an anchor position at P7 in both HLA-DRB4 alleles that shows strong specificity for negatively charged residue (D, and to a lower extend E) (Figure 3.2).

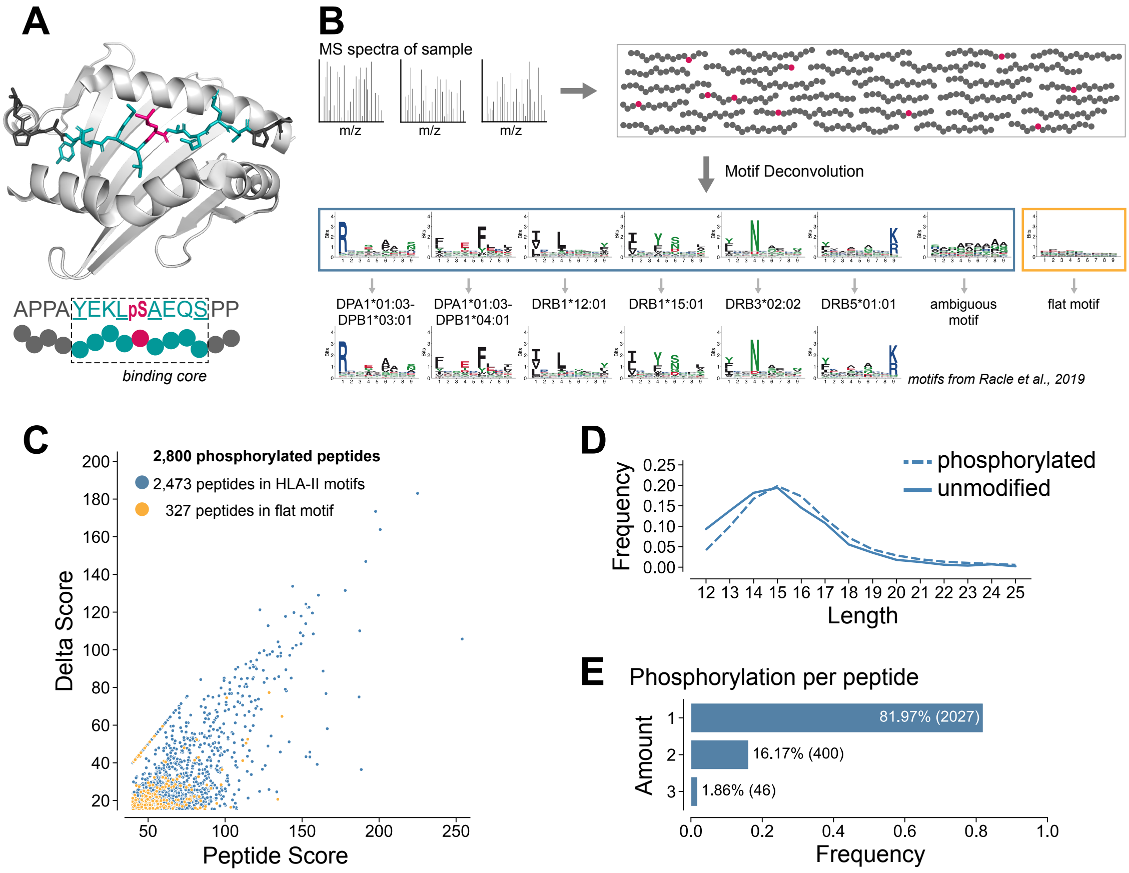


*Figure 3.1: MS-based HLA-II peptidomics identifies multiple phosphorylated HLA-II ligands. (A) Representative crystal structure of HLA-DRB1*01:01 molecule in complex with a phosphorylated ligand (PDB identification code 3L6F (Li et al., 2010)). The binding core of the peptide is shown in turquoise, the flanking regions in dark grey, the phosphorylated residue in pink, and the HLA-DR in light grey. Anchor positions P1, P4, P6, and P9 are underlined in the peptide sequence and point towards the HLA-II binding site. **(B)** HLA-II peptidomics MS spectra were analyzed for each sample separately to identify HLA-II ligands, including phosphorylated peptides. The peptides were then processed by MoDec and assigned to specific motifs, including a flat motif used to identify contaminants or wrongly identified peptides (yellow box) (Racle et al., 2019). Motifs were annotated to the HLA-II alleles present in each sample based on the similarity with known HLA-II binding motifs or left as 'ambiguous' when this annotation could not be unambiguously performed. **(C)** Distribution of Andromeda search engine peptide spectrum match scores ('Peptide score') vs. score differences to the second-best peptide spectrum match ('Delta scores') of the phosphorylated HLA-II ligands. Those assigned to the flat motif are shown in yellow, the others are shown in blue. **(D)** Comparison of length distribution of unmodified and phosphorylated HLA-II ligands. **(E)** Amount of detected phosphorylated residues per phosphorylated peptide in the HLA-II phosphopeptidome.*

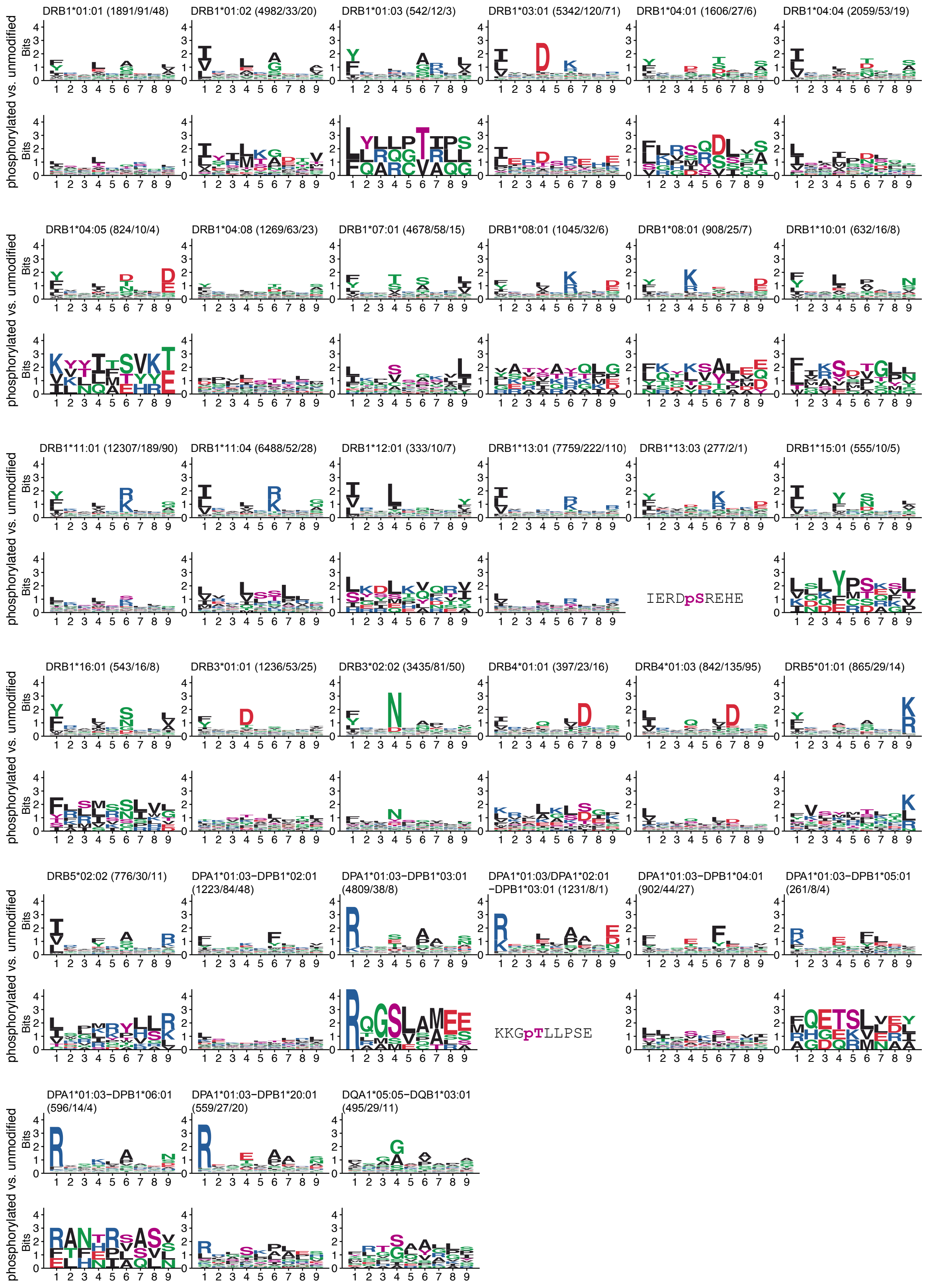


*Figure 3.2: **Phosphorylated peptides bind to HLA-II molecules with specific motifs. List of alleles with phosphorylated peptides.** For each allele, the HLA-II motif based on unmodified ligand is shown on top, and the motif of phosphorylated HLA-II ligands determined in this work is shown below. Numbers correspond to the number of peptides (unmodified peptides / all*

phosphorylated peptides / only phosphorylated peptides with the phosphorylated residue in the core). Phosphorylated residues are shown in pink.

Phosphorylated residues show positional specificity in HLA-II ligands

To investigate if there is any preference for phosphorylated residues in the peptide binding core and the peptide flanking regions (PFRs), we computed the amount of phosphorylated residues detected in these different parts of the peptide and compared it to the total amount of residues. We could see that phosphorylation is enriched outside of the peptide binding core, with 53.5% of all phosphorylated residues found in PFRs, while PFRs cover only ~41% of the positions of phosphorylated HLA-II ligands (Figure 3.3A). We then analyzed the occurrences of phosphorylation in PFRs and in particular how they are distributed in the first and last three amino acids of the PFRs at the N- and C-terminus of the phosphorylated peptides. Specific preferences for amino acids in these regions have been attributed to peptide processing and cleavage (37, 82, 176). Phosphorylation at the N-terminal region of phosphorylated HLA-II ligands are mostly found at P3 (49.5%), followed by P2 (29%) and P1 (21.5%) which stands in contrast to the evenly distributed phosphorylated residues at the C-terminal PFR (Figure 3.3B). We then looked at the distribution of phosphorylated residues within the 9-mer binding core. We could clearly see less phosphorylated residues at the main anchor positions P1 and P9, which is consistent with the higher specificity observed at these positions in unmodified HLA-II ligands (Figure 3.3C). Less expectedly, a relatively high frequency of phosphorylated residues was observed at secondary anchor positions (especially P4 and P6).

To further investigate the preference for phosphorylated residues at specific positions in the core, we performed competitor binding assays for two different HLA-DR alleles testing different versions of the same peptide containing the phosphorylated residues at all possible positions within the core (see Methods in 3.4). The two peptides were selected among the set of phosphorylated HLA-II ligands identified by MS with the phosphorylated residue at the non-anchor positions P2 and P5, respectively. The results of the binding assays showed that for both alleles, the version of the peptide that was found in our MS data showed good binding (see Figure 3.3D for HLA-DRB1*01:01 with pS at P2 and Figure 3.3E for HLA-DRB1*07:01 with pS at P5). Furthermore, the unmodified version of the peptide bound similarly well. The presence of the phosphorylated residue at other positions showed inferior binding, especially at anchor positions P1, P4, P6, and P9 (Figure 3.3D, E). These positions could clearly be identified as anchor positions of the alleles (see binding motifs Figure 3.3D, E left panels). We then selected a second peptide for HLA-DRB1*01:01 that was found in our MS data with a phosphorylated residue predicted at the secondary anchor position P4. The low binding with the phosphorylated residue at P1 and P9 could be confirmed. However, for other core positions the results did not fully recapitulate those of Figure 3.3D and showed a good binding of this peptide with a phosphorylated residue only at P2 and P4 (Figure 3.3F). Overall, these observations suggest that the preference for the position of the phosphorylated residue in the middle of the core may be different for different peptides, which could explain the relatively broad distribution in Figure 3.3C, and the lack of exclusion of P4 and P6 secondary anchor positions.

The preference for the negatively charged aspartic acid (D) at secondary anchor positions (e.g., P4 and P6) in several HLA-II alleles and the ability to bind peptides with phosphorylated residues at these positions (see example of Figure 3.3F) may further explain why the distribution of phosphorylated residues in the middle of the core does not show a strong preference for non-anchor positions only. To understand if there is a relationship between the specificity for aspartic acid at anchor positions and the ability to bind phosphorylated residues at these positions, we computed the frequency of aspartic acid and of phosphorylated residues found at secondary anchor positions (see Methods in 3.4). We could see a trend of higher frequency of phosphorylated residues at these secondary anchor positions for alleles with aspartic acid in their unmodified HLA-II peptidome (Figure 3.3G). These results suggest that alleles with aspartic acid at secondary anchor positions are especially prone to accommodate phosphorylated residues at these positions, as can be seen for HLA-DRB3*02:02 at P4 or HLA-DRB4*01:01 at P7 (Figure 3.3H).

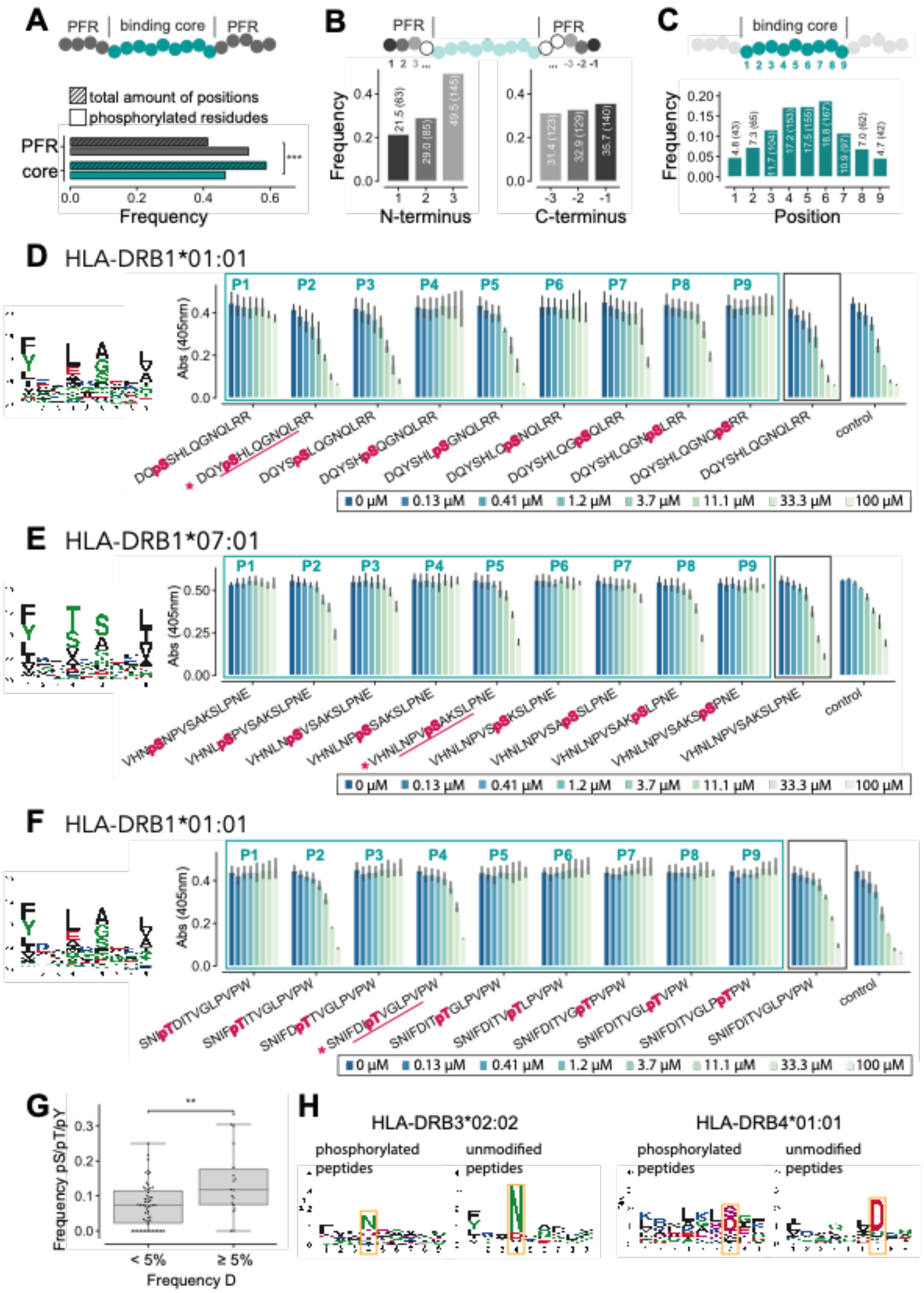


*Figure 3.3: **Phosphorylated residues show positional specificity in HLA-II ligands.** **(A)** Distribution of phosphorylated residues and total residues in the binding core vs PFRs of phosphorylated HLA-II ligands. **(B)** Distribution of phosphorylated residues in the first and last three residues of N- and C-terminal PFRs, respectively. **(C)** Positional distribution of phosphorylated residues in the*

binding core of phosphorylated HLA-II ligands. ***(D-F)*** *Competitor binding assays for peptides with a phosphorylated residue at each of the different core positions (turquoise box) and without phosphorylated residue (black box). The peptide initially found by MS is marked by a pink asterisk and the core predicted by MoDec is underlined.* ***(G)*** *Frequencies of phosphorylation at secondary anchor position for positions with an aspartic acid frequency lower or higher than 5%.* ***(H)*** *Binding motifs of HLA-DRB3*02:02 and HLA-DRB4*01:01 showing specificity for aspartic acid in unmodified peptides and phosphorylated residues in phosphorylated peptides.*

HLA-II ligands are phosphorylated by a broader repertoire of kinases than HLA-I ligands

To investigate the presence of kinase motifs in the HLA-II phosphopeptidome, we specifically searched for known kinase motifs from the PhosphoMotif Finder of the Human Protein Reference Database (177) in phosphorylated and unmodified HLA-II ligands as well as in the human proteome and the human phosphoproteome (125) (see Methods in 3.4). Most kinase motifs are present at a similar or lower frequency than in the human phosphoproteome with only a few exceptions (see Figure 3.4A, e.g. [K/R]X[pS/pT] or [pS/pT]X[K/R]). In particular, the very frequent kinase motif [pS/pT]P that is seen at ~32% of phosphorylated serine and threonine in the human phosphoproteome was only detected for 7.9% of phosphorylated serine and threonine in phosphorylated HLA-II ligands. This reflects roughly the frequency of proline after serine or threonine in the human proteome (7.2%). As expected from binding motifs of phosphorylated HLA-II ligands (see Figure 3.2), no clear motif enrichment was found comparing phosphorylated and unmodified HLA-II ligands (Figure 3.4B). In our previous work on phosphorylated HLA-I ligands, we could see a high enrichment of [pS/pT]P motifs, which are phosphorylated by proline-dependent serine/threonine kinases such as MAPK1, as well as an enrichment of the RXX[pS/pT] motif (Figure 3.4B), which corresponds to kinases such as PKA or PKB (175). To assess what could be the reasons for the lack of enrichment of these motifs in the HLA-II phosphopeptidome, we investigated if this may reflect a gene bias of source proteins in the HLA-II peptidome (i.e., peptides coming from proteins with such phosphorylation sites are under-represented in HLA-II ligands, irrespective of the phosphorylation status). To explore this hypothesis, we computed the overlap between the source genes of all phosphorylated HLA-II ligands with source genes of proteins containing phosphosites with the [pS/pT]P/RXX[pS/pT] motifs in the human phosphoproteome. This overlap was either as expected (odds ratio: 1.01 for [pS/pT]P-motifs), or slightly higher than expected (odds ratio: 1.22 for RXX[pS/pT]-motifs). Thus, we conclude that the lack of enrichment of [pS/pT]P and RXX[pS/pT] phosphorylation sites in phosphorylated HLA-II ligands is not due to a source gene bias and hypothesize that such phosphosites are present without phosphorylation in the pool of HLA-II – bound peptides.

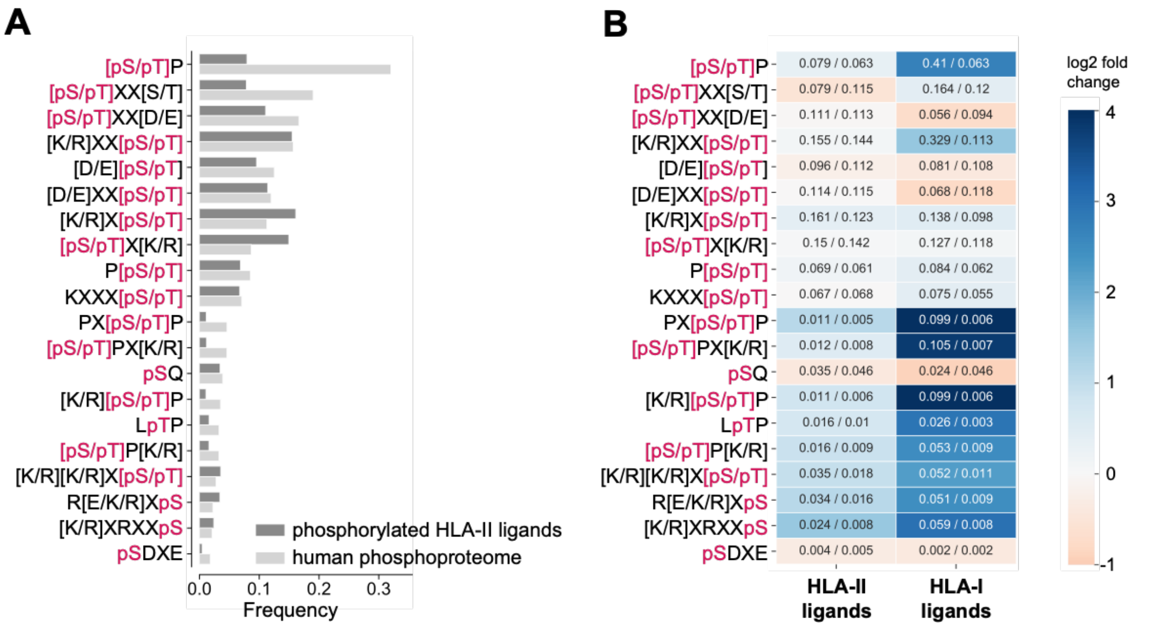


Figure 3.4: HLA-II ligands are phosphorylated by a broader repertoire of kinases than HLA-I ligands. (A) Kinase motif frequency in the HLA-II phosphopeptidome and the human phosphoproteome. (B) Kinase motif frequency in the HLA-II peptidome (1st column, frequency in phosphorylated / unmodified peptides) and the HLA-I peptidome (2nd column, frequency in phosphorylated / unmodified peptides). Heatmap colors show the log2 fold change between phosphorylated and unmodified peptides. In A and B, kinase motifs are sorted according to the frequency in the human phosphoproteome.

The HLA-II phosphopeptidome improves prediction of phosphorylated HLA-II ligands

We then used our HLA-II phosphopeptidome to expand our HLA-II ligand prediction method MixMHC2pred (37) to phosphorylated peptides. To this end, MixMHC2pred was retrained combining both unmodified and phosphorylated peptides (see Methods in 3.4). To benchmark its performance, we performed a 5-fold cross-validation with each HLA-II allele for which we could find at least five phosphorylated ligands in our dataset. We then compared the new version of MixMHC2pred with the existing tools NetMHCIIpan4.0 (88) and MARIA (83) (see Methods in 3.4). Our results show that the predictions of phosphorylated peptides significantly improved with the new version of

MixMHC2pred (Figure 3.5A and Supplemental Figure 3.2A). Similar results were obtained when predicting the same set of peptides with phosphorylated residues substituted by glutamic acid. As a second benchmarking, we performed a leave-one-sample-out cross-validation by excluding all the phosphorylated peptides found in one sample from the training of MixMHC2pred and using them as test set. Predictions with MixMHC2pred were compared to those with NetMHCIIpan4.0 and MARIA and results showed significantly improved predictions with the new version of MixMHC2pred (Figure 3.5B and Supplemental Figure 3.2B).

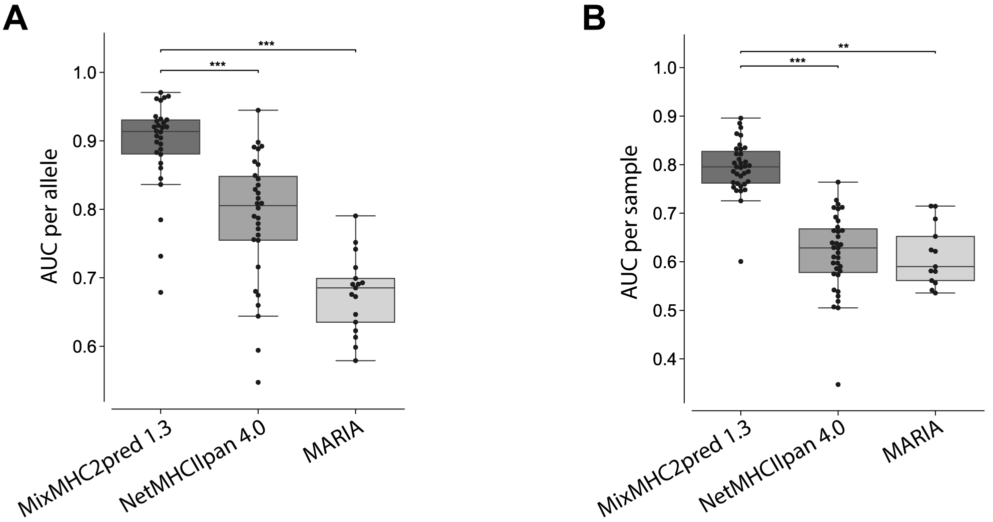


Figure 3.5: The HLA-II phosphopeptidome improves prediction of phosphorylated HLA-II ligands. (A) *AUC values for all alleles in the cross-validation of the new version of MixMHC2pred (v1.3) trained on a combination of phosphorylated and unmodified peptides. For comparison AUC values of NetMHCIIpan4.0 and MARIA are shown.* ***(B)*** *AUC values for leave-one-sample-out cross-validation for all samples in this work. AUC values of MixMCH2pred1.3 are shown in comparison to AUCs of NetMHCIIpred4.0 and MARIA. P-values between the different predictors were calculated using the paired two-sided Wilcoxon signed rank-test.*

3.3 Discussion

A better understanding of the repertoire and the properties of HLA-II ligands is promising for the development of personalized cancer immunotherapies such as cancer vaccines (166–168). As cancer can cause aberrant PTMs which can be presented on HLA-II molecules and recognized by CD4+ T cells (137, 152, 156), including PTMs such as phosphorylation in HLA-II ligand predictions is powerful to expand the list of potential targets for cancer immunotherapy.

In this work, we provided an in-depth analysis of the HLA-II phosphopeptidome. We could identify binding motifs of phosphorylated HLA-II ligands for more than 30 alleles. These binding motifs showed high similarity with those of unmodified HLA-II ligands at anchor positions, in particular the main anchors at P1 and P9.

Our analysis of the position of phosphorylated residues in HLA-II ligands revealed a preference of phosphorylation either at PFRs or in the middle of the core, and a very low frequency of phosphorylated residues at the main anchor positions P1 and P9 (Figure 3.3A, C). These results could be confirmed with binding assays and are consistent with the low frequency of phosphorylated residues at anchor positions in HLA-I ligands (175). The presence of phosphorylated residues at secondary anchor positions (mainly P4 and P6) was less expected. However, our binding assays confirmed that specific peptides can accommodate phosphorylated residues at such anchor positions, especially when negatively charged residues (mainly aspartic acid) are present in the binding motifs based on unmodified ligands (Figure 3.3F). This clearly shows that the charged properties of phosphorylated residues also play a role in the binding of phosphorylated HLA-II ligands. We speculate that a lower frequency of phosphorylation at P2 and P8 compared to central positions could possibly be due to some kinase motifs at these positions being less compatible with HLA-II binding motifs. For instance, the phosphorylated residues corresponding to the [pS/pT]P kinase motif are unlikely to be found at P8 since HLA-II motifs strongly disfavor Pro at P9. Overall, our results suggest that the preference for phosphorylated residues at positions middle positions of the core may be context dependent and that phosphorylation is not depleted at anchor positions P4 and P6. The low frequency of phosphorylated residues at the N-terminus of the HLA-II ligands (Figure 3.3C) could suggest that this may not be favorable for protein cleavage or transport, although additional work will be warranted to confirm this hypothesis.

Our analysis of kinase motifs did not detect a strong over-representation of common kinase motifs seen in the intracellular phosphoproteome or the HLA-I phosphopeptidome (e.g., [pS/pT]P or RXX[pS/pT]). We speculate that many of these potential phosphosites are simply not phosphorylated in the pool of ligands available for loading onto HLA-II molecules or that these are efficiently removed by phosphatases before or after binding to the HLA-II molecules. This supports the idea that phosphorylated residues observed among HLA-II ligands come from a more diverse repertoire of kinases compared to the one observed in both the phosphoproteome and the HLA-I phosphopeptidome. This hypothesis is consistent with the differences between class I and class II antigen presentation pathways and the fact that many HLA-II ligands come from endocytosis of proteins in the extracellular matrix, which may undergo phosphorylation by different sets of kinases compared to intracellular proteins displayed on HLA-I molecules.

Finally, we used our data to build a predictor for phosphorylated HLA-II ligands by including the HLA-II phosphopeptidome in the training data of our HLA-II ligand prediction method MixMHC2pred. The results of the cross-validation showed that our expanded HLA-II ligand predictor could improve predictions for phosphorylated HLA-II ligands compared to existing tools (Figure 3.5A, B). The motifs of phosphorylated HLA-II ligands suggest that binding of phosphorylated peptides is shaped by the binding motif of the HLA-II allele and some positional specificity for the phosphorylated residues (e.g., exclusion of P1 and P9).

Altogether, our work represents the first in-depth analysis of the repertoire of phosphorylated HLA-II ligands. We anticipate that this unique resource and the associated computational tools to predict phosphorylated HLA-II ligands in different contexts will facilitate the discovery of potential new targets for CD4+ T-cell recognition in infectious diseases and cancer immunotherapy.

3.4 Methods

Curation of immunopeptidomics HLA-II MS datasets

The MaxQuant platform (178) version 1.5.5.1 was employed to search the MS peak lists of 23 samples from (37) against a fasta file containing the human proteome (Homo_sapiens_UP000005640_9606, the reviewed part of UniProt, with no isoforms, including 21,026 entries downloaded in March 2017) and a list of 247 frequently observed contaminants. Peptides with a length between 8 and 25 amino acids were allowed. The second peptide identification option in Andromeda was enabled and the enzyme specificity was set as unspecific. A false-discovery rate of 5% was required for peptides and no protein false-discovery rate was set. The initial allowed mass deviation of the precursor ion was set to 6 ppm and the maximum fragment mass deviation was set to 20 ppm. Methionine oxidation, N-terminal acetylation and phosphorylation on serine, threonine, and tyrosine were set as variable modifications. The resulting list of msms identifications were further filtered to include phosphorylated peptides with identification score ≥ 40, score difference to the second best peptide spectrum match (delta score) ≥ 10, and localization probability for phosphorylation of >0.75 as well as peptide lengths restricted to 12 to 25 amino acids. To obtain better specificity for unmodified peptides, sample-specific unmodified sequences identified with 1% FDR were obtained from (37).

HLA-II Motif Deconvolution for identification of HLA-I binding motifs

To determine allelic restriction and identify phosphorylated HLA-II binding motifs, the motif deconvolution method MoDec (37) was expanded to allow for phosphorylated residues within sequences. This was done by expanding the alphabet from 20 essential amino acids to include the three phosphorylated residues, giving an alphabet of size 23. For each sample, MoDec was applied to the combined set of phosphorylated HLA-II ligands identified in this work and unmodified HLA-II ligands (37). Motifs were manually assigned to alleles by using previously identified binding motifs of unmodified HLA-II ligands (see example in Figure 3.1B). For some samples, some motifs could not be unambiguously assigned to a single allele, as previously observed with unmodified HLA-II ligands (37). The corresponding peptides were not assigned to any allele and were not considered in the allele-specific analyses. MoDec also includes a flat motif that is useful to model potential contaminants or wrongly identified peptides (37, 175). Peptides assigned to this flat motif were not considered in any analysis. Sequence logos including phosphorylated ligands were drawn with the extended version of ggseqlogo (https://github.com/GfellerLab/ggseqlogo) (40) and phosphorylated residues are shown in purple (Figure 3.2).

General analysis of phosphorylated HLA-II ligands

The frequency of phosphorylated residues inside of the binding core and in PFRs were compared to the fraction of positions in these two regions of the peptides and a two-sided Fisher's exact test was applied to calculate the p-value (Figure 3.3A). Only peptides with phosphorylated residues in the first three position of the N-terminal region or in the last three positions of the C-terminal region were used to compute the distribution of phosphorylation in PFRs (Figure 3.3B). The distribution of phosphorylated residues per position in the core was computed position-wise for all peptides that contained at least one phosphorylation in the binding core (Figure 3.3C).

To analyze the correlation between aspartic acid and phosphorylated residues at anchor positions, each allele with at least 20 phosphorylated peptides was considered. Anchor positions per allele were identified in unmodified peptides if any of the secondary anchor positions (positions 4, 6, or 7) had an entropy higher than the median entropy of all positions of the allele. At these allele-specific anchor positions the frequency of aspartic acid in unmodified peptides and the frequency of phosphorylated residues in phosphorylated peptides was measured. p-value was calculated with independent two-sample t-test (Figure 3.3G).

Competition Binding Assays

To test binding of different phosphorylated HLA-II ligands, competition assays were performed for HLA-DRB1*01:01 and HLA-DRB1*07:01 with two and one different peptides detected by MS in the samples, respectively. The competition assays were performed by mixing in v-bottom 96-well plate (Greiner Bio-One) in a citrate saline buffer (100 mM citrate, pH 6.0), with 0.2% β-octyl-glucopyranoside (Calbiochem), 1×complete protease inhibitors (Roche), and 1 mg of the biotinylated empty allele with a FLAG-tagged peptide at fixed concentration of 2 µM (Influenza $HA_{307-319}$ for HLA-DRB1*01:01 and NY-ESO-1_{87-99} for HLA-DRB1*07:01). The peptide of interest was added to this mix into each well at a final concentration of 0, 0.13, 0.41, 1.3, 3.7, 11.1, 33.3, and 100 µM. For the control, untagged peptide (Influenza $HA_{307-319}$ or NY-ESO-1_{87-99}) were added at the respective concentrations to the mix of allele and FLAG-tagged peptide. After incubation at 37°C overnight, the binding of the tagged peptides to HLA-II molecule was measured by ELISA. The mix was transferred to a plate coated with avidin and the FLAG-peptide was detected with an anti-FLAG-alkaline phosphatase conjugate (Sigma), developed with pNPP SigmaFAST substrate and absorbance was read with a 405nm – filter (Figure 3.3D-F).

Kinase motifs

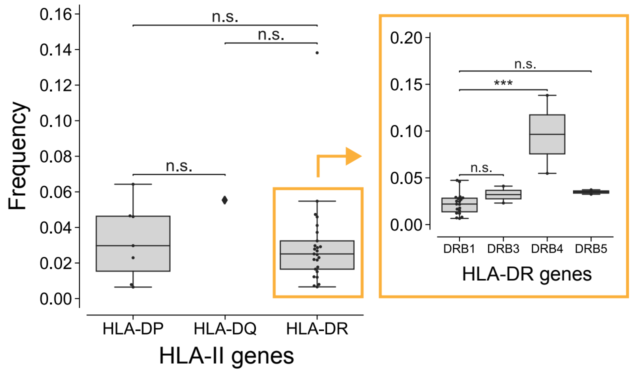
To detect enrichment of kinase motifs in phosphorylated HLA-II ligands, occurrences of all motifs from the PhosphoMotif Finder of the Human Protein Reference Database (177) were searched in phosphorylated as well as unmodified HLA-II ligands (Figure 3.4A, B). To be able to search each motif on all peptides, including those that had the phosphorylated residue at the first or last positions, resulting in the motif not to be entirely contained in the HLA-II ligands, each phosphorylated and unmodified peptide was mapped to its source protein and N'- and C'-terminally extended. Occurrences of kinase motifs were normalized by the amount of phosphorylated residues of the corresponding motif in all phosphorylated peptides (e.g., amount of pS in phosphorylated peptides for motif pSP, amount of pS and pT in all phosphorylated peptides for motif [pS/pT]P). Similarly, frequencies of kinase motifs in unmodified peptides were determined by normalization with the amount of the unmodified counterpart of the phosphorylated resides of the corresponding motif in all unmodified peptides (e.g., amount of S in unmodified peptides for motif SP, amount of S and T in unmodified peptides for motif [S/T]P). For comparison, the same analysis was also performed on phosphorylated HLA-I peptides from our previous work (175) as well as the human proteome (Uniprot accession number UP000005640) and a human phosphoproteome (125). The most common and non-redundant kinase motifs that showed a p-value $p \leq 0.01$ between phosphorylated and unmodified HLA-II peptides (computed with one-sided Fisher's exact test) are shown in Figure 3.4A, B. To analyze whether the difference in kinase motifs between phosphorylated HLA-II and HLA-I ligands is due to a gene bias of source proteins, a universal set of source genes of MS-detected sequences was defined. This universal gene set contained all source genes of phosphorylated and unmodified HLA-II sequences, source genes from a phosphoproteome (125), and a MS-based human proteome (179). Next, source genes of known phosphosites from the phosphoproteome containing the [pS/pT]P or RXX[pS/pT] motif were identified and the overlap with unmodified HLA-II ligands was computed. p-values were computed with one-sided Fisher's exact tests.

Predictor

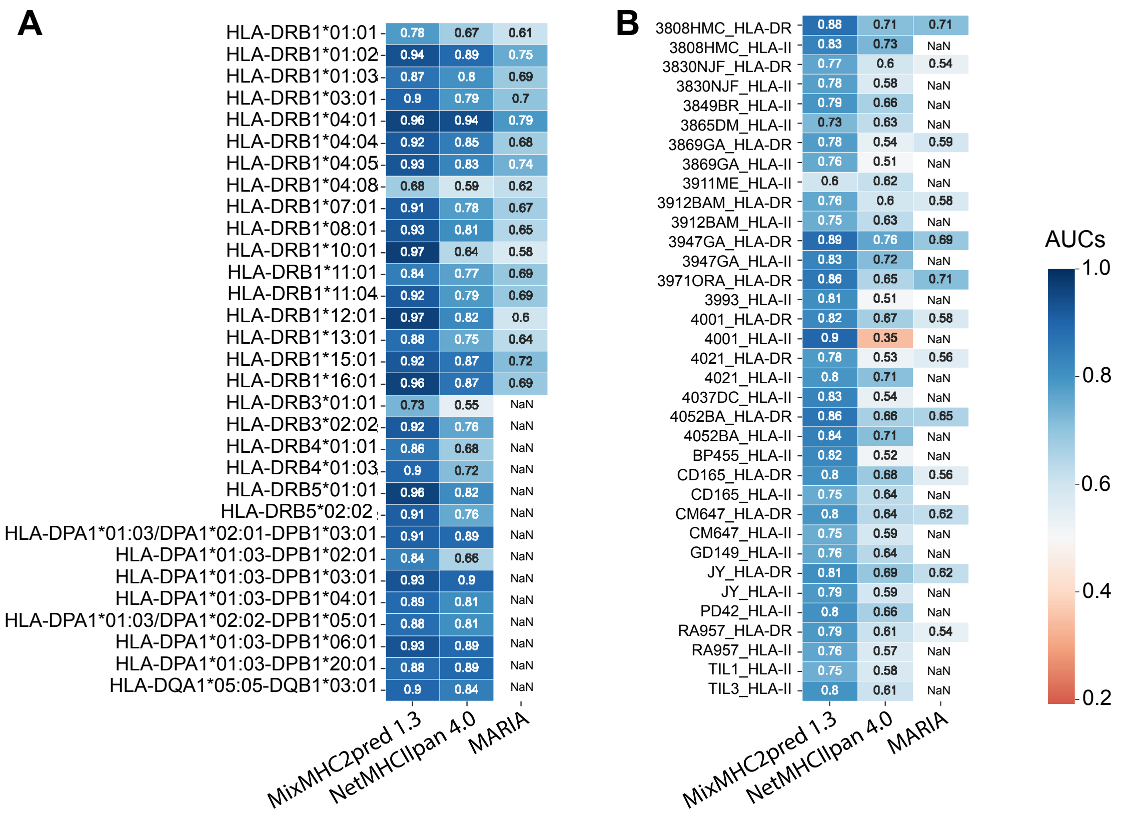
Predictions of interactions between HLA-II alleles and phosphorylated peptides were based on the previously developed HLA-II prediction method MixMHC2pred (37). Following our previous work on phosphorylated HLA-I ligands (175), the MixMHC2pred training framework was extended to consider 23 amino acids and the phosphorylated peptides were added to the training set used in (37). MixMHC2pred was then retrained on this combined dataset of both phosphorylated and unmodified HLA-II ligands. A 5-fold cross-validation was performed for each allele with at least five phosphorylated peptides by splitting the phosphorylated peptides randomly into testing and training data (one fifth and four fifth of the phosphorylated peptides, respectively). In each round of the cross-validation, the set of phosphorylated peptides used for training was added to the existing (unmodified) training data previously used (37). For the leave-one-sample-out cross-validation, each sample from the dataset was iteratively used as test set and all phosphorylated peptides that were found in this sample were removed from the training data of MixMHCpred. Five times the amount of positive phosphorylated peptides were added to the testing data as negative peptides. Peptides used as negative in the test set were of lengths 12 to 25 amino acids and contained a phosphosite from the human phosphoproteome (the phosphosite itself, the length of the peptide as well as the position of the phosphosite in the 12 to 25-mer were randomly chosen).

Other existing HLA-II predictors (MARIA (83) and NetMHCIIpan4.0 (88)) were used to benchmark the prediction results (Figure 3.5 and Supplemental Figure 3.2). MARIA was used with the unmodified version of the phosphorylated peptides (S, T, Y instead of pS, pT, pY) as well as gene names of the peptides' source proteins for all available alleles and only applied to alleles given in the list of alleles supported by MARIA. Phosphorylated residues in HLA-II ligands were substituted by 'X' for predictions with NetMHCIIpan4.0. For comparison of the predictions with each method, the area under the curve (AUC) of the receiver operating characteristic (ROC) was computed for each allele and each predictor. Due to limited allele availability, MARIA was only applied to the HLA-DR alleles (Figure 3.5A and Supplemental Figure 3.2A) and HLA-DR specific samples (Figure 3.5B and Supplemental Figure 3.2B).

3.5 Supplemental Figures



Supplemental Figure 3.1: HLA-II alleles present similar fractions of phosphorylated ligands. *Frequency of phosphorylated peptides in the HLA-II peptidome for the alleles of each gene (-DR, -DP, or -DQ).*



Supplemental Figure 3.2: The HLA-II phosphopeptidome improves prediction of phosphorylated HLA-II ligands. (A) *Heatmap showing AUC values from the 5-fold cross-validation for each allele and each predictor. NaN for MARIA denotes alleles not available for predictions.* ***(B)*** *Heatmap showing AUC values from leave-one-sample-out cross-validation for each sample and each predictor. MARIA was only applied to HLA-II peptidomics samples analyzed with HLA-DR antibodies.*

Chapter 4 Conclusions & Discussion

Ever since the first discovery on how to use mechanisms of the immune system to treat malignancies, immunotherapies have received increasing interested for the development of targeted and efficient cancer treatments and have been rapidly advancing over the last decade (180). Cancer immunotherapies, including immune checkpoint inhibitors or chimeric antigen receptor (CAR) T cell therapies, have shown promising results (181–183). Identification of neoantigens for clinical application is crucial for the development of antigen-specific therapeutic strategies and is facilitated by efficient pipelines based on high-throughput sequencing data and bioinformatics models to predict HLA – ligand interactions (67). The presented thesis summarizes two in-depth analyses of (1) phosphorylated HLA-I and (2) HLA-II ligands as well as (3) prediction models for HLA – ligand interactions specifically trained on HLA-I/HLA-II phosphopeptidomes and concludes the following:

Phosphorylated HLA-I binding motifs can be identified from MS – based immunopeptidomics data and are shaped by a combination of HLA-I binding motifs, intrinsic HLA-I binding properties of phosphorylated peptides, and kinase motifs

The first part of this thesis comprises a comprehensive analysis of phosphorylated HLA-I ligands with a large allelic-coverage using different MS – based immunopeptidomics studies. Motif deconvolution in combination with unmodified HLA-I ligands enabled the determination of allelic restriction for each sample and allowed us to identify HLA-I binding motifs of phosphorylated ligands. The data used in this study was collected with a less conservative FDR of 5% in contrast to the usually applied FDR of 1%. However, to maintain peptide spectrum matches with relatively high confidence, stringent cutoffs of identification parameters were employed. Our results showed that less stringent FDR of 5% together with motif deconvolution resulted in high confidence HLA-I ligands. Thus, we propose that motif deconvolution acts as additional filtering for the identification of high confidence peptides, including those containing phosphorylation.

Clear properties of HLA-I binding of phosphorylated ligands could be detected in the HLA phosphopeptidome. For instance, a strong enrichment of phosphorylation at P4 was seen in peptides of length 8-12 amino acids. Previous studies observed interactions between the phosphate moiety at P4 of the peptide with different positions in the heavy chain of the allele (142, 147). One study on phosphorylated HLA-B peptides concluded that the interaction between the phosphorylated ligand and Arg62 of the HLA heavy chain, a position highly conserved in HLA-B alleles, was specific for phosphorylated ligands (147). Comparable results were observed for HLA-A*02:01, where phosphorylated ligands interacted with Arg65 of the heavy chain of the allele (142). Additionally, these studies concluded that the high frequency of basic residues at P1 of the phosphorylated peptides provided additional binding stability for the peptides through intermolecular bonds between the phosphate moiety and the basic residue and further suggested that this was linked to the structure of the HLA binding pocket. Here, the analysis of the HLA-I phosphopeptidome could confirm an enrichment of phosphorylation at P4 as well as arginine at P1 in phosphorylated ligands beyond HLA-A*02:01 and HLA-B alleles. Binding assays performed for multiple phosphorylated ligands with HLA-A*02:01 and HLA-B*07:02 showed no difference for peptides with or without arginine at P1. Furthermore, [K/R]XX[pS/pT] is the binding motif for different kinases, thus we propose that the enrichment of basic residues at P1 in phosphorylated HLA-I ligands is likely a result of kinase motifs. Nevertheless, we cannot exclude that some peptides show improved binding due to intermolecular bonds between the basic residue at P1 and the phosphate moiety at P4.

While it was shown that HLA-A and HLA-B alleles are expressed at higher levels than HLA-C alleles (11, 12), we observed that HLA-C alleles expressed on average the highest fraction of phosphorylated ligands. This suggests that while HLA-C expression and peptide binding are usually limited due to different factors, phosphorylated peptides actually fit the binding motifs of HLA-C alleles better than those of HLA-A or HLA-B alleles. For instance, unmodified binding motifs of some HLA-C alleles including HLA-C*07:01 show an enrichment for arginine at P1 and thus are a good fit for peptides containing phosphosites with the very frequent [K/R]XX[pS/pT] kinase motif.

Differences in HLA binding properties and kinase motifs in the HLA-II phosphopeptidome compared to the HLA-I phosphopeptidome

Chapter 3 presents the first comprehensive work on the HLA-II phosphopeptidome including the identification of binding motifs of phosphorylated HLA-II ligands for more than 30 alleles. In line with what was observed for HLA-I, phosphorylated HLA-II binding motifs show high similarities to unmodified HLA-II ligands at anchor positions, in particular the main anchors P1 and P9, as well as low frequencies of phosphorylation at these main anchor positions. Contrarily, phosphorylated residues within the binding core of HLA-II ligands are evenly distributed among central positions and no preference for one specific position was observed. Furthermore, phosphorylation was also seen at secondary anchor positions (particularly P4 and P6) of HLA-II ligands. Binding assays of one HLA-DRB1*01:01 – restricted peptide with a phosphorylated residue at P4 showed that phosphorylation at this position did not interfere with the binding, potentially a result of the preference for negatively charged aspartic acid for the allele at P4. Additionally, only little amount of phosphorylation was observed at P2 and P8 in HLA-II ligands, which stands in contradiction to phosphorylation observed at higher levels at other non-anchor positions of HLA-II. Similarly, in HLA-I ligands a low frequency of phosphorylation was observed at non-anchor position P8. Furthermore, binding assays of HLA-I ligands with phosphorylation at P8 showed better binding compared to other non-anchor positions (i.e. P3, P5, P6, and P7). These observations are possibly explained by phosphorylation – specific characteristics influencing the binding of phosphorylated ligands to HLA molecules. In particular, the very frequently observed kinase motif [pS/pT]P and the incompatibility of proline at P9 in both, HLA-I and HL-II binding motifs, could likely explain the lack of phosphorylation at P8 in HLA ligands. This outlines that in addition to the HLA binding motifs, the binding of phosphorylated ligands to HLA molecules is restricted by phosphorylation – specific properties.

Additionally, and in contrast to what could be observed in the HLA-I phosphopeptidome, phosphorylated HLA-II ligands did not show any clear enrichment of kinase binding motifs compared to the unmodified HLA-II peptidome. In particular, the very frequently observed motifs [pS/pT]P and [K/R]XX[pS/pT] showed no enrichment in phosphorylated compared to the unmodified HLA-II peptidome. This suggests that a broader repertoire of kinases is responsible for the phosphorylation of source proteins of HLA-II – presented ligands. This is most likely explained by HLA-II molecules expressing peptides from endocytosed proteins as well as different kinases responsible for the extracellular phosphoproteome. For instance, recent studies have identified a secreted kinase which is active in the extracellular matrix as well as a secretory pathway kinase that is responsible for phosphorylation of secreted proteins in the Golgi apparatus (184, 185).

Training on HLA-I and HLA-II phosphopeptidomes results in robust predictions of HLA – phosphorylated peptide interactions

The identified HLA-I and HLA-II phosphopeptidomes enabled us to specifically train prediction models for HLA – ligand interactions. Comparing the prediction model for phosphorylated HLA-I ligands trained on both phosphorylated and unmodified peptides with a model trained only on the unmodified HLA-I peptidome showed that the additional information derived from the phosphorylated residues in the training data improved the predictions of such phosphorylated peptides. Similarly, the existing HLA-II ligand predictor (37) was expanded by including the HLA-II phosphopeptidome in the training data and showed robust predictions of phosphorylated HLA-II ligands. Both prediction models are the first to specifically train on post-translationally modified HLA ligand data and show promising results to include HLA-I and HLA-II phosphopeptidomes in future developments of such tools for comprehensive methods to further study HLA ligand interaction.

Limitations and Future Perspectives

The limited amount of phosphorylated HLA ligands as well as the smaller allelic-coverage compared to unmodified HLA ligands presents one of the main constraint of this work. This lack of phosphorylated HLA ligands in combination with the allele-specificity of the prediction models for HLA – ligand interactions limits their application in future research on phosphorylated HLA ligands. Previous HLA peptidomics studies either did not include phosphorylated peptides in the identification of HLA ligands or focused on specific alleles or phosphorylated peptides. Here we saw that phosphorylated ligands curated in the course of this study on average make up less than two percent of the HLA-I peptidome, with some variations for different alleles. Similarly, searching for known phosphosites in the unmodified HLA-I peptidome resulted in a comparable number, outlining that the HLA-I phosphopeptidome is naturally limited by the phosphorylation events of the phosphoproteome. Nevertheless, considering post-translationally modified ligands in future identification of HLA peptidomes will further add to phosphorylated HLA binding motifs as well as establish phosphorylation – specific characteristics of phosphorylated HLA ligands, such as kinase motifs, and therefore contribute to the training data of the prediction models. Furthermore, pan-specific predictors have been shown to improve predictions of unmodified ligands by overcoming the lack of or the limited availability of training data for rare and less studied alleles (78). Thus, the development of robust models for pan-allelic and pan-length predictors is one the main objectives in the field of HLA – ligand predictions, providing models applicable to a wide range of patient data independent of HLA typing. Predictors for phosphorylated HLA-I and HLA-II ligands would likewise benefit from robust pan-specific models, since HLA-I and HLA-II phosphopeptidomes have been sparsely studied, and help improve in particularly predictions for less frequent alleles.

Lastly, the thesis focused on the identification and prediction of phosphorylated HLA ligands, one of the most common and well-studied PTMs, also in regard to antigen processing, presentation and T cell recognition. However, other PTMs have also been observed in HLA peptidome, including deamidated HLA-I ligands, citrullinated self-peptides bound to HLA-DR alleles, or glycosylated HLA-II ligands in melanoma cell lines (169, 186, 187). Future work on other post-translationally modified HLA peptides can take advantage of motif deconvolution and prediction models presented in this work. This is of particular interest for PTMs that have been seen to play a role in malignancies, such as arginine methylation in cancer (188). Further exploring the space of post-translationally modified HLA-I and HLA-II ligands can contribute to a better understanding of antigen processing and HLA immunopeptidomes of modified ligands as well as their role in the development for therapeutic strategies.

One of the main challenges for applying the presented predictors in future studies on phosphorylated HLA ligands lies in confidently determining immunogenicity of potential phosphorylated HLA ligands for the development of immunotherapies. It was shown that T cell recognition of phosphorylated HLA-I ligands are both sequence-specific and phosphorylation-dependent (138, 139, 141, 142, 144). However, the question remaining is can a phosphosite be exclusively cancer-specific and can HLA ligands containing this site therefore be robust targets for immunotherapies? Phosphorylation is actively involved in many cell regulatory processes and expression of phosphorylated ligands by HLA-I and HLA-II molecules was observed in healthy as well as tumor tissue, thus T cell tolerance against phosphorylated HLA ligands exists. With somatic mutation creating a novel phosphorylation event (189), HLA – presented antigens containing this cancer-specific phosphosite can be potentially immunogenic. Recent work to determine immunogenicity of phosphorylated HLA-I ligands, derived from proteins linked to cell growth and survival, observed decelerated tumor growth *in vivo* (140) and further lead to a first clinical trial for vaccines with cancer-associated phosphorylated peptides in melanoma patients (146). Various methods and databases to study the phosphorylation landscape to detect differences caused by malignancies have been proposed and provide useful tools for the identification of cancer-specific phosphosites (189–192).

From a global perspective, providing access to robust tools for direct identification of phosphorylated HLA ligands without performing intensive experimental work will facilitate future research on phosphorylated HLA ligands as well as help defining their role in cancer-specific antigen presentation and T cell recognition. The most recent advancement in immunotherapies have reshaped cancer therapies in the last years. Future developments will further benefit from efficient pipelines for antigen identification including time- and effort-reduced methods such as *in silico* predictors. Identification of cancer-specific antigens, including antigens containing cancer-specific phosphosites, expand the pool of potential targets for immunotherapeutic approaches and show promising potential for the development of a more targeted and specific treatment.

References

1. Murphy, K., and Weaver, C. (2017) *Janeway's Immunobiology 9th Edition*
2. Larsen, S. B., Cowley, C. J., and Fuchs, E. (2020) Epithelial cells: liaisons of immunity. *Curr. Opin. Immunol.*,
3. Janeway, C. A., and Medzhitov, R. (2002) Innate immune recognition. *Annu. Rev. Immunol.*,
4. Kawai, T., and Akira, S. (2010) The role of pattern-recognition receptors in innate immunity: Update on toll-like receptors. *Nat. Immunol.*,
5. Heesters, B. A., van der Poel, C. E., Das, A., and Carroll, M. C. (2016) Antigen Presentation to B Cells. *Trends Immunol.*,
6. Laidlaw, B. J., Craft, J. E., and Kaech, S. M. (2016) The multifaceted role of CD4+ T cells in CD8+ T cell memory. *Nat. Rev. Immunol.*,
7. Zhang, N., and Bevan, M. J. (2011) CD8+ T Cells: Foot Soldiers of the Immune System. *Immunity*,
8. Palm, N. W., and Medzhitov, R. (2009) Pattern recognition receptors and control of adaptive immunity. *Immunol. Rev.*,
9. Kaech, S. M., Wherry, E. J., and Ahmed, R. (2002) Effector and memory T-cell differentiation: Implications for vaccine development. *Nat. Rev. Immunol.*,
10. Kappler, J. W., Roehm, N., and Marrack, P. (1987) T cell tolerance by clonal elimination in the thymus. *Cell*,
11. Neefjes, J. J., and Ploegh, H. L. (1988) Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. *Eur. J. Immunol.*,
12. Kulkarni, S., Savan, R., Qi, Y., Gao, X., Yuki, Y., Bass, S. E., Martin, M. P., Hunt, P., Deeks, S. G., Telenti, A., Pereyra, F., Goldstein, D., Wolinsky, S., Walker, B., Young, H. A., and Carrington, M. (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*,
13. Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2015) The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* 43, D423–D431
14. Neefjes, J., Jongsma, M. L. M., Paul, P., and Bakke, O. (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* 11, 823–836
15. Ackerman, A. L., and Cresswell, P. (2004) Cellular mechanisms governing cross-presentation of exogenous antigens. *Nat. Immunol.* 5, 678–684
16. Remesh, S. G., Andreatta, M., Ying, G., Kaever, T., Nielsen, M., McMurtrey, C., Hildebrand, W., Peters, B., and Zajonc, D. M. (2017) Unconventional peptide presentation by major histocompatibility complex (MHC) class i allele HLA-A*02:01: Breaking confinement. *J. Biol. Chem.*,
17. Dengjel, J., Schoor, O., Fischer, R., Reich, M., Kraus, M., Müller, M., Kreymborg, K., Altenberend, F., Brandenburg, J., Kalbacher, H., Brock, R., Driessen, C., Rammensee, H.-G., and Stevanovic, S. (2005) Autophagy promotes MHC class II presentation of peptides from intracellular source proteins. *Proc. Natl. Acad. Sci.*

U. S. A. 102, 7922–7927

18. Blees, A., Januliene, D., Hofmann, T., Koller, N., Schmidt, C., Trowitzsch, S., Moeller, A., and Tampé, R. (2017) Structure of the human MHC-I peptide-loading complex. *Nature*,
19. Ortmann, B., Copeman, J., Lehner, P. J., Sadasivan, B., Herberg, J. A., Grandea, A. G., Riddell, S. R., Tampé, R., Spies, T., Trowsdale, J., and Cresswell, P. (1997) A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes. *Science (80-.).*,
20. Radcliffe, C. M., Diedrich, G., Harvey, D. J., Dwek, R. A., Cresswell, P., and Rudd, P. M. (2002) Identification of specific glycoforms of major histocompatibility complex class I heavy chains suggests that class I peptide loading is an adaptation of the quality control pathway involving calreticulin and ERp57. *J. Biol. Chem.*,
21. Serwold, T., Gonzalez, F., Kim, J., Jacob, R., and Shastri, N. (2002) ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature*,
22. York, I. A., Chang, S. C., Saric, T., Keys, J. A., Favreau, J. M., Goldberg, A. L., and Rock, K. L. (2002) The Er aminopeptidase ERAP I enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat. Immunol.*,
23. Williams, A. P., Peh, C. A., Purcell, A. W., McCluskey, J., and Elliott, T. (2002) Optimization of the MHC class I peptide cargo is dependent on tapasin. *Immunity*,
24. Hughes, E. A., Hammond, C., and Cresswell, P. (1997) Misfolded major histocompatibility complex class I heavy chains are translocated into the cytoplasm and degraded by the proteasome. *Proc. Natl. Acad. Sci. U. S. A.*,
25. Parmentier, N., Stroobant, V., Colau, D., De Diesbach, P., Morel, S., Chapiro, J., Van Endert, P., and Van Den Eynde, B. J. (2010) Production of an antigenic peptide by insulin-degrading enzyme. *Nat. Immunol.* 11, 449–454
26. Crotzer, V. L., and Blum, J. S. (2005) Autophagy and intracellular surveillance: Modulating MHC class II antigen presentation with stress. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7779–7780
27. Roche, P. A., and Cresswell, P. (1990) Invariant chain association with HLA-DR molecules inhibits immunogenic peptide binding. *Nature*,
28. Roche, P. A., and Cresswell, P. (1991) Proteolysis of the class II-associated invariant chain generates a peptide binding site in intracellular HLA-DR molecules. *Proc. Natl. Acad. Sci. U. S. A.*,
29. Teyton, L., O'Sullivan, D., Dickson, P. W., Lotteau, V., Sette, A., Fink, P., and Peterson, P. A. (1990) Invariant chain distinguishes between the exogenous and endogenous antigen presentation pathways. *Nature*,
30. Costantino, C. M., Ploegh, H. L., and Hafler, D. A. (2009) Cathepsin S Regulates Class II MHC Processing in Human CD4 + HLA-DR + T Cells . *J. Immunol.*,
31. Villadangos, J. A., and Ploegh, H. L. (2000) Proteolysis in MHC class II antigen presentation: Who's in charge? *Immunity*,
32. Denzin, L. K., and Cresswell, P. (1995) HLA-DM induces clip dissociation from MHC class II αβ dimers and facilitates peptide loading. *Cell*,
33. Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R. T., Racle, J., Coukos,

G., and Bassani-Sternberg, M. (2018) The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. *J. Immunol.* 202, 1–12

34. Trolle, T., McMurtrey, C. P., Sidney, J., Bardet, W., Osborn, S. C., Kaever, T., Sette, A., Hildebrand, W. H., Nielsen, M., and Peters, B. (2017) The length distribution of class I restricted T cell epitopes is determined by both peptide supply and MHC allele specific binding preference. *J Immunol* 196, 1480–1487
35. Andreatta, M., Jurtz, V. I., Kaever, T., Sette, A., Peters, B., and Nielsen, M. (2017) Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology* 152, 255–264
36. Chicz, R. M., Urban, R. G., Lane, W. S., Gorga, J. C., Stern, L. J., Vignali, D. A. A., and Strominger, J. L. (1992) Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature*,
37. Racle, J., Michaux, J., Rockinger, G. A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., Bassani-Sternberg, M., and Gfeller, D. (2019) Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* 37, 1283–1286
38. Tynan, F. E., Borg, N. A., Miles, J. J., Beddoe, T., El-Hassen, D., Silins, S. L., Van Zuylen, W. J. M., Purcell, A. W., Kjer-Nielsen, L., McCluskey, J., Burrows, S. R., and Rossjohn, J. (2005) High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I: Implications for T-cell receptor engagement and T-cell immunodominance. *J. Biol. Chem.*,
39. Guillaume, P., Picaud, S., Baumgaertner, P., Montandon, N., Schmidt, J., Speiser, D. E., Coukos, G., Bassani-Sternberg, M., Filippakopoulos, P., and Gfeller, D. (2018) The C-terminal extension landscape of naturally presented HLA-I ligands. *Proc. Natl. Acad. Sci. U. S. A.* 115, 5083–5088
40. Wagih, O. (2017) Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* 33, 3645–3647
41. Dendrou, C. A., Petersen, J., Rossjohn, J., and Fugger, L. (2018) HLA variation and disease. *Nat. Rev. Immunol.* 18, 325–339
42. Bulek, A. M., Cole, D. K., Skowera, A., Dolton, G., Gras, S., Madura, F., Fuller, A., Miles, J. J., Gostick, E., Price, D. A., Drijfhout, J. W., Knight, R. R., Huang, G. C., Lissin, N., Molloy, P. E., Wooldridge, L., Jakobsen, B. K., Rossjohn, J., Peakman, M., Rizkallah, P. J., and Sewell, A. K. (2012) Structural basis for the killing of human beta cells by CD8 + T cells in type 1 diabetes. *Nat. Immunol.*,
43. Thomas, R., Apps, R., Qi, Y., Gao, X., Male, V., O'Huigin, C., O'Connor, G., Ge, D., Fellay, J., Martin, J. N., Margolick, J., Goedert, J. J., Buchbinder, S., Kirk, G. D., Martin, M. P., Telenti, A., Deeks, S. G., Walker, B. D., Goldstein, D., McVicar, D. W., Moffett, A., and Carrington, M. (2009) HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat. Genet.*,
44. Apps, R., Qi, Y., Carlson, J. M., Chen, H., Gao, X., Thomas, R., Yuki, Y., Del Prete, G. Q., Goulder, P., Brumme, Z. L., Brumme, C. J., John, M., Mallal, S., Nelson, G., Bosch, R., Heckerman, D., Stein, J. L., Soderberg, K. A., Moody, M. A., Denny, T. N., Zeng, X., Fang, J., Moffett, A., Lifson, J. D., Goedert, J. J., Buchbinder, S., Kirk, G. D., Fellay, J., McLaren, P., Deeks, S. G., Pereyra, F., Walker, B., Michael, N. L., Weintrob, A., Wolinsky, S., Liao, W., and Carrington, M. (2013) Influence of

HLA-C expression level on HIV control. *Science (80-.).*,

45. Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M. J., van de Haar, J., Engin, H. B., de Prisco, N., Ideker, T., Hildebrand, W. H., Font-Burgada, J., and Carter, H. (2017) MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*,
46. McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., Birkbak, N. J., Veeriah, S., Van Loo, P., Herrero, J., Swanton, C., Jamal-Hanjani, M., Shafi, S., Czyzewska-Khan, J., Johnson, D., Laycock, J., Bosshard-Carter, L., Gorman, P., Hynds, R. E., Horswell, S., Mitter, R., Escudero, M., Stewart, A., Xu, H., Turajlic, S., Abbosh, C., Goldman, J., Stone, R. K., Denner, T., Matthews, N., Elgar, G., Ward, S., Costa, M., Begum, S., Phillimore, B., Chambers, T., Nye, E., Graca, S., Al Bakir, M., Joshi, K., Furness, A., Ben Aissa, A., Wong, Y. N. S., Georgiou, A., Quezada, S., Hartley, J. A., Lowe, H. L., Lawrence, D., Hayward, M., Panagiotopoulos, N., Kolvekar, S., Falzon, M., Borg, E., Marafioti, T., Simeon, C., Hector, G., Smith, A., Aranda, M., Novelli, M., Oukrif, D., Janes, S. M., Thakrar, R., Forster, M., Ahmad, T., Lee, S. M., Papadatos-Pastos, D., Carnell, D., Mendes, R., George, J., Navani, N., Ahmed, A., Taylor, M., Choudhary, J., Summers, Y., Califano, R., Taylor, P., Shah, R., Krysiak, P., Rammohan, K., Fontaine, E., Booton, R., Evison, M., Crosbie, P., Moss, S., Idries, F., Joseph, L., Bishop, P., Chaturved, A., Quinn, A. M., Doran, H., Leek, A., Harrison, P., Moore, K., Waddington, R., Novasio, J., Blackhall, F., Rogan, J., Smith, E., Dive, C., Tugwood, J., Brady, G., Rothwell, D. G., Chemi, F., Pierce, J., Gulati, S., Naidu, B., Langman, G., Trotter, S., Bellamy, M., Bancroft, H., Kerr, A., Kadiri, S., Webb, J., Middleton, G., Djearaman, M., Fennell, D., Shaw, J. A., Le Quesne, J., Moore, D., Nakas, A., Rathinam, S., Monteiro, W., Marshall, H., Nelson, L., Bennett, J., Riley, J., Primrose, L., Martinson, L., Anand, G., Khan, S., Amadi, A., Nicolson, M., Kerr, K., Palmer, S., Remmen, H., Miller, J., Buchan, K., Chetty, M., Gomersall, L., Lester, J., Edwards, A., Morgan, F., Adams, H., Davies, H., Kornaszewska, M., Attanoos, R., Lock, S., Verjee, A., MacKenzie, M., Wilcox, M., Bell, H., Hackshaw, A., Ngai, Y., Smith, S., Gower, N., Ottensmeier, C., Chee, S., Johnson, B., Alzetani, A., Shaw, E., Lim, E., De Sousa, P., Barbosa, M. T., Bowman, A., Jordan, S., Rice, A., Raubenheimer, H., Proli, C., Cufari, M. E., Ronquillo, J. C., Kwayie, A., Bhayani, H., Hamilton, M., Bakar, Y., Mensah, N., Ambrose, L., Devaraj, A., Buderi, S., Finch, J., Azcarate, L., Chavan, H., Green, S., Mashinga, H., Nicholson, A. G., Lau, K., Sheaff, M., Schmid, P., Conibear, J., Ezhil, V., Ismail, B., Irvin-sellers, M., Prakash, V., Russell, P., Light, T., Horey, T., Danson, S., Bury, J., Edwards, J., Hill, J., Matthews, S., Kitsanta, Y., Suvarna, K., Fisher, P., Keerio, A. D., Shackcloth, M., Gosney, J., Postmus, P., Feeney, S., Asante-Siaw, J., Aerts, H. J. W. L., Dentro, S., and Dessimoz, C. (2017) Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*,
47. Chowell, D., Morris, L. G. T., Grigg, C. M., Weber, J. K., Samstein, R. M., Makarov, V., Kuo, F., Kendall, S. M., Requena, D., Riaz, N., Greenbaum, B., Carroll, J., Garon, E., Hyman, D. M., Zehir, A., Solit, D., Berger, M., Zhou, R., Rizvi, N. A., and Chan, T. A. (2018) Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science (80-.).* 359, 582–587
48. Najafimehr, H., Hajizadeh, N., Nazemalhosseini-Mojarad, E., Pourhoseingholi, M. A., Abdollahpour-Alitappeh, M., Ashtari, S., and Zali, M. R. (2020) The role of

Human leukocyte antigen class I on patient survival in Gastrointestinal cancers: a systematic review and meta- analysis. *Sci. Rep.* 10, 1–8

49. Coulie, P. G., Van den Eynde, B. J., van der Bruggen, P., and Boon, T. (2014) Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat. Rev. Cancer* 14, 135–146
50. Yao, J., Caballero, O. L., Yung, W. K. A., Weinstein, J. N., Riggins, G. J., Strausberg, R. L., and Zhao, Q. (2014) Tumor subtype-specific cancer-testis antigens as potential biomarkers and immunotherapeutic targets for cancers. *Cancer Immunol. Res.*,
51. Gjerstorff, M. F., Andersen, M. H., and Ditzel, H. J. (2015) Oncogenic cancer/testis antigens: Prime candidates for immunotherapy. *Oncotarget*,
52. Baylin, S. B., and Jones, P. A. (2011) A decade of exploring the cancer epigenome-biological and translational implications. *Nat. Rev. Cancer*,
53. Siebenkäs, C., Chiappinelli, K. B., Guzzetta, A. A., Sharma, A., Jeschke, J., Vatapalli, R., Baylin, S. B., and Ahuja, N. (2017) Inhibiting DNA methylation activates cancer testis antigens and expression of the antigen processing and presentation machinery in colon and ovarian cancer cells. *PLoS One*,
54. Lambert, P. (2009) in *Encyclopedia of Microbiology*
55. Efremova, M., Finotello, F., Rieder, D., and Trajanoski, Z. (2017) Neoantigens generated by individual mutations and their role in cancer immunity and immunotherapy. *Front. Immunol.*,
56. Maby, P., Galon, J., and Latouche, J. B. (2016) Frameshift mutations, neoantigens and tumor-specific CD8+ T cells in microsatellite unstable colorectal cancers. *Oncoimmunology*,
57. Maletzki, C., Schmidt, F., Dirks, W. G., Schmitt, M., and Linnebacher, M. (2013) Frameshift-derived neoantigens constitute immunotherapeutic targets for patients with microsatellite-instable haematological malignancies: Frameshift peptides for treating MSI+ blood cancers. *Eur. J. Cancer*,
58. Faridi, P., Li, C., Ramarathinam, S. H., Vivian, J. P., Illing, P. T., Mifsud, N. A., Ayala, R., Song, J., Gearing, L. J., Hertzog, P. J., Ternette, N., Rossjohn, J., Croft, N. P., and Purcell, A. W. (2018) A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Sci. Immunol.* 3, 1–11
59. Liepe, J., Marino, F., Sidney, J., Jeko, A., Bunting, D. E., Sette, A., Kloetzel, P. M., Stumpf, M. P. H., Heck, A. J. R., and Mishto, M. (2016) A large fraction of HLA class I ligands are proteasome-generated spliced peptides. 354, 605–610
60. Mylonas, R., Beer, I., Iseli, C., Chong, C., Pak, H., Gfeller, D., Coukos, G., Xenarios, I., Müller, M., and Bassani-Sternberg, M. (2018) Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome. *Mol Cell Proteomics Pap.*,
61. Rolfs, Z., Müller, M., Shortreed, M. R., Smith, L. M., and Bassani-Sternberg, M. (2019) Comment on "A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands." *Sci. Immunol.*, 1–3
62. Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., Justesen, S.,

Stevens, J., Lane, W. J., Eisenhaure, T., Lan Zhang, G., Clauser, K. R., Hacohen, N., Carr, S. A., Wu, C. J., and Keskin, D. B. (2020) A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* 38, 199–209

63. Vigneron, N. (2015) Human Tumor Antigens and Cancer Immunotherapy. *Biomed Res. Int.* 2015,
64. Bräunlein, E., and Krackhardt, A. M. (2017) Identification and Characterization of Neoantigens As Well As Respective Immune Responses in Cancer Patients. *Front. Immunol.* 8, 1–8
65. Gfeller, D., Bassani-Sternberg, M., Schmidt, J., and Luescher, I. F. (2016) Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology* 5,
66. Roudko, V., Greenbaum, B., and Bhardwaj, N. (2020) Computational Prediction and Validation of Tumor-Associated Neoantigens. *Front. Immunol.*,
67. Zhou, C., Zhu, C., and Liu, Q. (2019) Toward in silico Identification of Tumor Neoantigens in Immunotherapy. *Trends Mol. Med.*,
68. Justesen, S., Harndahl, M., Lamberth, K., Nielsen, L. L. B., and Buus, S. (2009) Functional recombinant MHC class II molecules and high-throughput peptide-binding assays. *Immunome Res.*,
69. Salvat, R., Moise, L., Bailey-Kellogg, C., and Griswold, K. E. (2014) A high throughput MHC II binding assay for quantitative analysis of peptide epitopes. *J. Vis. Exp.*,
70. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 46, 315–326
71. Abelin, J. G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S. P., Creech, A. L., Serrano, L. R., Nasir, G., Nasrullah, Y., McGann, C. D., Velez, D., Ting, Y. S., Poran, A., Rothenberg, D. A., Chhangawala, S., Rubinsteyn, A., Hammerbacher, J., Gaynor, R. B., Fritsch, E. F., Greshock, J., Oslund, R. C., Barthelme, D., Addona, T. A., Arieta, C. M., and Rooney, M. S. (2019) Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. *Immunity* 51, 766-779.e17
72. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015) Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation. *Mol. Cell. Proteomics* 14, 658–673
73. Caron, E., Kowalewski, D. J., Koh, C. C., Sturm, T., Schuster, H., and Aebersold, R. (2015) Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol. Cell. Proteomics*,
74. Sercarz, E. E., and Maverakis, E. (2003) MHC-guided processing: Binding of large antigen fragments. *Nat. Rev. Immunol.*,
75. Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A., and Peters, B. (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 43, D405–D412
76. Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S., and Nielsen, M.

(2013) NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*,

77. Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. (2009) Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinformatics*,
78. Nielsen, M., and Andreatta, M. (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* 8, 33
79. Bassani-Sternberg, M., and Gfeller, D. (2016) Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *J. Immunol.* 197, 2492–2499
80. Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., Ojo, N. C., Caldwell, K., Abhyankar, J., Boucher, T., Hart, M. G., Makarov, V., Montpreville, V. T. De, Mercier, O., Chan, T. A., Scagliotti, G., Bironzo, P., Novello, S., Karachaliou, N., Rosell, R., Anderson, I., Gabrail, N., Hrom, J., Limvarapuss, C., Choquette, K., Spira, A., Rousseau, R., Voong, C., Rizvi, N. A., Fadel, E., Frattini, M., Jooss, K., Skoberne, M., Francis, J., and Yelensky, R. (2018) Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol. 2018*,
81. Alvarez, B., Barra, C., Nielsen, M., and Andreatta, M. (2018) Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes. *Proteomics* 1700252, 1–10
82. Barra, C., Alvarez, B., Andreatta, M., Buus, S., and Nielsen, M. (2018) Footprints of antigen processing boost MHC class II natural ligand binding predictions. *Genome Med.*, 285767
83. Chen, B., Khodadoust, M. S., Olsson, N., Wagar, L. E., Fast, E., Liu, C. L., Muftuoglu, Y., Sworder, B. J., Diehn, M., Levy, R., Davis, M. M., Elias, J. E., Altman, R. B., and Alizadeh, A. A. (2019) Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* 37,
84. Garde, C., Ramarathinam, S. H., Jappe, E. C., Nielsen, M., Kringelum, J. V., Trolle, T., and Purcell, A. W. (2019) Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics* 71, 445–454
85. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* 199, 3360–3368
86. O'Donnell, T. J. O., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018) MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst.* 7, 129–132
87. O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020) MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.*,
88. Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W. H., Peters, B., Peters, B., Nielsen, M., and Nielsen, M. (2020) Improved Prediction of MHC II Antigen

Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. *J. Proteome Res.* 19, 2304–2315

89. Rammensee, H. G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., and Stevanović, S. (1999) SYFPEITHI: Database for MHC ligands and peptide motifs. *Immunogenetics*,
90. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G., and Gfeller, D. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* 13,
91. Andreatta, M., and Nielsen, M. (2015) Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. *Bioinformatics* 32, 511–517
92. Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*,
93. Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*,
94. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020) 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
95. Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S., and Nielsen, M. (2009) NetMHCpan, a method for MHC class i binding prediction beyond humans. *Immunogenetics* 61, 1–13
96. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., and Buus, S. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2,
97. Nielsen, M., and Lund, O. (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*,
98. Mitchell, P. J., and Tjian, R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (80-.).*,
99. Pawson, T., and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science (80-.).*,
100. Amos-Binks, A., Patulea, C., Pitre, S., Schoenrock, A., Gui, Y., Green, J. R., Golshani, A., and Dehne, F. (2011) Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences. *BMC Bioinformatics*,
101. Golovin, A., and Henrick, K. (2008) MSDmotif: Exploring protein sites and motifs. *BMC Bioinformatics*,
102. Jayaram, N., Usvyat, D., and Martin, A. C. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*,
103. Bailey, T. L., and Elkan, C. (1995) Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Mach. Learn.*,

104. Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006) MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*,

105. Gfeller, D., Butty, F., Wierzbicka, M., Verschueren, E., Vanhee, P., Huang, H., Ernst, A., Dar, N., Stagljar, I., Serrano, L., Sidhu, S. S., Bader, G. D., and Kim, P. M. (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol.* 7, 484–484

106. Kim, T., Tyndel, M. S., Huang, H., Sidhu, S. S., Bader, G. D., Gfeller, D., and Kim, P. M. (2012) MUSI: An integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res.* 40,

107. Nielsen, M., and Andreatta, M. (2017) NNAlign: A platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* 45, W344–W349

108. Andreatta, M., Lund, O., and Nielsen, M. (2013) Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* 29, 8–14

109. Andreatta, M., Alvarez, B., and Nielsen, M. (2017) GibbsCluster: Unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.*,

110. Walsh, C. T., Garneau-Tsodikova, S., and Gatto, G. J. (2005) Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chemie - Int. Ed.* 44, 7342–7372

111. Deribe, Y. L., Pawson, T., and Dikic, I. (2010) Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* 17, 666–672

112. Hunter, T. (2009) Tyrosine phosphorylation: thirty years and counting. *Curr. Opin. Cell Biol.* 21, 140–146

113. Wang, Y. C., Peterson, S. E., and Loring, J. F. (2014) Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.* 24, 143–160

114. Fukuda, S., Nishida-Fukuda, H., Nakayama, H., Inoue, H., and Higashiyama, S. (2012) Monoubiquitination of pro-amphiregulin regulates its endocytosis and ectodomain shedding. *Biochem. Biophys. Res. Commun.* 420, 315–320

115. Goth, C. K., Vakhrushev, S. Y., Joshi, H. J., Clausen, H., and Schjoldager, K. T. (2018) Fine-Tuning Limited Proteolysis: A Major Role for Regulated Site-Specific O-Glycosylation. *Trends Biochem. Sci.* 43, 269–284

116. Xu, C., and Ng, D. T. W. (2015) Glycosylation-directed quality control of protein folding. *Nat. Rev. Mol. Cell Biol.* 16, 742–752

117. Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017) The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *Int. J. Mol. Med.* 40, 271–280

118. Graves, J. D., and Krebs, E. G. (1999) in *Pharmacology and Therapeutics*

119. Ochoa, D., Jarnuczak, A. F., Gehre, M., Soucheray, M., Kleefeldt, A. A., Vieitez, C., Hill, A., Garcia-Alonso, L., Swaney, D. L., Vizcaino, J. A. A., Noh, K.-M., and Beltrao, P. (2019) The functional landscape of the human phosphoproteome. *Nat. Biotechnol.*,

120. Guo, Y., Costa, R., Ramsey, H., Starnes, T., Vance, G., Robertson, K., Kelley, M., Reinbold, R., Scholer, H., and Hromas, R. (2002) The embryonic stem cell transcription factors Oct-4 and FoxD3 interact to regulate endodermal-specific

promoter expression. *Proc. Natl. Acad. Sci. U. S. A.* 99, 3663–3667

121. Brumbaugh, J., Hou, Z., Russell, J. D., Howden, S. E., Yu, P., Ledvina, A. R., Coon, J. J., and Thomson, J. A. (2012) Phosphorylation regulates human OCT4. *Proc. Natl. Acad. Sci. U. S. A.* 109, 7162–7168

122. Lewis, T. S., Shapiro, P. S., and Ahn, N. G. (1998) Signal transduction through MAP kinase cascades. *Adv. Cancer Res.*,

123. Pearson, G., Robinson, F., Beers Gibson, T., Xu, B., Karandikar, M., Berman, K., and Cobb, M. H. (2001) Mitogen-Activated Protein (MAP) Kinase Pathways: Regulation and Physiological Functions*. *Endocr. Rev.*,

124. Cohen, P. (2002) The origins of protein phosphorylation. *Nat. Cell Biol.* 4,

125. Sharma, K., D'Souza, R. C. J., Tyanova, S., Schaab, C., Wisniewski, J. R., Cox, J., and Mann, M. (2014) Resource Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature. *Cell Rep.* 8, 1583–1594

126. Newton, A. C. (2003) Regulation of the ABC kinases by phosphorylation: Protein kinase C as a paradigm. *Biochem. J.* 370, 361–371

127. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science (80-.).* 298, 1912–1934

128. Rowland, M. A., Fontana, W., and Deeds, E. J. (2012) Crosstalk and competition in signaling networks. *Biophys. J.*,

129. Hitosugi, T., and Chen, J. (2014) Post-translational modifications and the Warburg effect. *Oncogene* 33, 4279–4285

130. Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., Beer, D. G., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Hayes, D. N., Hernandez, B., Herman, J. G., Heymach, J. V., Jurisica, I., Kucherlapati, R., Kwiatkowski, D., Ladanyi, M., Robertson, G., Schultz, N., Shen, R., Sinha, R., Sougnez, C., Tsao, M. S., Travis, W. D., Weinstein, J. N., Wigle, D. A., Wilkerson, M. D., Chu, A., Cherniack, A. D., Hadjipanayis, A., Rosenberg, M., Weisenberger, D. J., Laird, P. W., Radenbaugh, A., Ma, S., Stuart, J. M., Byers, L. A., Baylin, S. B., Govindan, R., Meyerson, M., Gabriel, S. B., Cibulskis, K., Kim, J., Stewart, C., Lichtenstein, L., Lander, E. S., Lawrence, M. S., Getz, E., Fulton, R., Fulton, L. L., McLellan, M. D., Wilson, R. K., Ye, K., Fronick, C. C., Maher, C. A., Miller, C. A., Wendl, M. C., Cabanski, C., Mardis, E., Wheeler, D., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Chuah, E., Dhalla, N., Guin, R., Hirst, C., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Varhol, R., Robertson, A. G., Wye, N., Thiessen, N., Holt, R. A., Jones, S. J. M., Marra, M. A., Imielinski, M., Onofrio, R. C., Hodis, E., Zack, T., Helman, E., Pedamallu, C. S., Mesirov, J., Saksena, G., Schumacher, S. E., Carter, S. L., Garraway, L., Beroukhim, R., Lee, S., Mahadeshwar, H. S., Pantazi, A., Protopopov, A., Ren, X., Seth, S., Song, X., Tang, J., Yang, L., Zhang, J., Chen, P. C., Parfenov, M., Xu, A. W., Santoso, N., Chin, L., Park, P. J., Hoadley, K. A., Auman, J. T., Meng, S., Shi, Y., Buda, E., Waring, S., Veluvolu, U., Tan, D., Mieczkowski, P. A., Jones, C. D., Simons, J. V., Soloway, M. G., Bodenheimer, T., Jefferys, S. R., Roach, J., Hoyle, A. P., Wu, J., Balu, S., Singh, D., Prins, J. F., Marron, J. S., Parker, J. S., Perou, C. M., Liu, J., Maglinte, D. T., Lai, P. H., Bootwalla, M. S., Van Den Berg, D. J., Triche, T., Cho,

J., DiCara, D., Heiman, D., Lin, P., Mallard, W., Voet, D., Zhang, H., Zou, L., Noble, M. S., Gehlenborg, N., Thorvaldsdottir, H., Nazaire, M. D., Robinson, J., Aksoy, B. A., Ciriello, G., Taylor, B. S., Dresdner, G., Gao, J., Gross, B., Seshan, V. E., Reva, B., Sumer, S. O., Weinhold, N., Sander, C., Ng, S., Zhu, J., Benz, C. C., Yau, C., Haussler, D., Spellman, P. T., Kimes, P. K., Broom, B. M., Wang, J., Lu, Y., Ng, P. K. S., Diao, L., Liu, W., Amos, C. I., Akbani, R., Mills, G. B., Curley, E., Paulauskis, J., Lau, K., Morris, S., Shelton, T., Mallery, D., Gardner, J., Penny, R., Saller, C., Tarvin, K., Richards, W. G., Cerfolio, R., Bryant, A., Raymond, D. P., Pennell, N. A., Farver, C., Czerwinski, C., Huelsenbeck-Dill, L., Iacocca, M., Petrelli, N., Rabeno, B., Brown, J., Bauer, T., Dolzhanskiy, C. O., Potapova, O., Rotin, D., Voronina, O., Nemirovich-Danchenko, E., Fedosenko, K. V., Gal, A., Behera, M., Ramalingam, S. S., Sica, G., Flieder, D., Boyd, J., Weaver, J. E., Kohl, B., Thinh, D. H. Q., Sandusky, G., Juhl, H., Duhig, E., Illei, P., Gabrielson, E., Shin, J., Lee, B., Rogers, K., Trusty, D., Brock, M. V., Williamson, C., Burks, E., Rieger-Christ, K., Holway, A., Sullivan, T., Asiedu, M. K., Kosari, F., Rekhtman, N., Zakowski, M., Rusch, V. W., Zippile, P., Suh, J., Pass, H., Goparaju, C., Owusu-Sarpong, Y., Bartlett, J. M. S., Kodeeswaran, S., Parfitt, J., Sekhon, H., Albert, M., Eckman, J., Myers, J. B., Morrison, C., Gaudioso, C., Borgia, J. A., Bonomi, P., Pool, M., Liptay, M. J., Moiseenko, F., Zaytseva, I., Dienemann, H., Meister, M., Schnabel, P. A., Muley, T. R., Peifer, M., Gomez-Fernandez, C., Herbert, L., Egea, S., Huang, M., Thorne, L. B., Boice, L., Salazar, A. H., Funkhouser, W. K., Rathmell, W. K., Dhir, R., Yousem, S. A., Dacic, S., Schneider, F., Siegfried, J. M., Hajek, R., Watson, M. A., McDonald, S., Meyers, B., Clarke, B., Yang, I. A., Fong, K. M., Hunter, L., Windsor, M., Bowman, R. V., Peters, S., Letovanec, I., Khan, K. Z., Jensen, M. A., Snyder, E. E., Srinivasan, D., Kahn, A. B., Baboud, J., Pot, D. A., Shaw, K. R. M., Sheth, M., Davidsen, T., Demchok, J. A., Yang, L., Wang, Z., Tarnuzzer, R., Zenklusen, J. C., Ozenberger, B. A., Sofia, H. J., and Cheney, R. (2014) Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature* 511, 543–550

131. Planchard, D., Popat, S., Kerr, K., Novello, S., Smit, E. F., Faivre-Finn, C., Mok, T. S., Reck, M., Van Schil, P. E., Hellmann, M. D., and Peters, S. (2018) Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.*,

132. Hatzivassiliou, G., Haling, J. R., Chen, H., Song, K., Price, S., Heald, R., Hewitt, J. F. M., Zak, M., Peck, A., Orr, C., Merchant, M., Hoeflich, K. P., Chan, J., Luoh, S. M., Anderson, D. J., Ludlam, M. J. C., Wiesmann, C., Ultsch, M., Friedman, L. S., Malek, S., and Belvin, M. (2013) Mechanism of MEK inhibition determines efficacy in mutant KRAS- versus BRAF-driven cancers. *Nature*,

133. Solomon, D. A., Kim, J. S., Cronin, J. C., Sibenaller, Z., Ryken, T., Rosenberg, S. A., Ressom, H., Jean, W., Bigner, D., Yan, H., Samuels, Y., and Waldman, T. (2008) Mutational inactivation of PTPRD in glioblastoma multiforme and malignant melanoma. *Cancer Res.*,

134. Veeriah, S., Brennan, C., Meng, S., Singh, B., Fagin, J. A., Solit, D. B., Paty, P. B., Rohle, D., Vivanco, I., Chmielecki, J., Pao, W., Ladanyi, M., Gerald, W. L., Liau, L., Cloughesy, T. C., Mischel, P. S., Sander, C., Taylor, B., Schultz, N., Major, J., Heguy, A., Fang, F., Mellinghoff, I. K., and Chan, T. A. (2009) The tyrosine

phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proc. Natl. Acad. Sci. U. S. A.*,

135. Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K. L., Lin, C., McLellan, M. D., Yan, P., Davies, S. R., Townsend, R. R., Skates, S. J., Wang, J., Zhang, B., Kinsinger, C. R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A. G., Fenyö, D., Ellis, M. J., and Carr, S. A. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*,
136. Andersen, M. H., Bonfill, J. E., Neisig, A., Arsequell, G., Sondergaard, I., Neefjes, J., Zeuthen, J., Elliott, T., and Haurum, J. S. (1999) Phosphorylated peptides can be transported by TAP molecules, presented by class I MHC molecules, and recognized by phosphopeptide-specific CTL. *J Immunol* 163, 3812–3818
137. Meyer, V. S., Drews, O., Günder, M., Hennenlotter, J., Rammensee, H.-G., and Stevanovic, S. (2009) Identification of natural MHC class II presented phosphopeptides and tumor-derived MHC class I phospholigands. *J. Proteome Res.* 8, 3666–3674
138. Zarling, A. L., Ficarro, S. B., White, F. M., Shabanowitz, J., Hunt, D. F., and Engelhard, V. H. (2000) Phosphorylated Peptides Are Naturally Processed and Presented by Major Histocompatibility Complex Class I Molecules In Vivo. *J. Exp. Med* 120800, 1755–1762
139. Zarling, A. L., Polefrone, J. M., Evans, A. M., Mikesh, L. M., Shabanowitz, J., Lewis, S. T., Engelhard, V. H., and Hunt, D. F. (2006) Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *Proc. Natl. Acad. Sci.* 103, 14889–14894
140. Zarling, A. L., Obeng, R. C., Desch, A. N., Pinczewski, J., Cummings, K. L., Deacon, D. H., Conaway, M., Slingluff, C. L., and Engelhard, V. H. (2014) MHC-restricted phosphopeptides from insulin receptor substrate-2 and CDC25b offer broad-based immunotherapeutic agents for cancer. *Cancer Res.* 74, 6784–6795
141. Mohammed, F., Cobbold, M., Zarling, A. L., Salim, M., Barrett-Wilt, G. A., Shabanowitz, J., Hunt, D. F., Engelhard, V. H., and Willcox, B. E. (2008) Phosphorylation-dependent interaction between antigenic peptides and MHC class I: a molecular basis for the presentation of transformed self. *Nat. Immunol.* 9, 1236–1243
142. Mohammed, F., Stones, D. H., Zarling, A. L., Willcox, C. R., Shabanowitz, J., Cummings, K. L., Hunt, D. F., Cobbold, M., Engelhard, V. H., and Willcox, B. E. (2017) The antigenic identity of human class I MHC phosphopeptides is critically dependent upon phosphorylation status. *Oncotarget* 8, 54160–54172
143. Petersen, J., Wurzbacher, S. J., Williamson, N. A., Ramarathinam, S. H., Reid, H. H., Nair, A. K. N., Zhao, A. Y., Nastovska, R., Rudge, G., Rossjohn, J., and Purcell, A. W. (2009) Phosphorylated self-peptides alter human leukocyte antigen class I-restricted antigen presentation and generate tumor-specific epitopes. *Proc. Natl. Acad. Sci.* 106, 2776–2781
144. Cobbold, M., De La Peña, H., Norris, A., Polefrone, J. M., Qian, J., English, A. M., Cummings, K. L., Penny, S., Turner, J. E., Cottine, J., Abelin, J. G., Malaker, S. A., Zarling, A. L., Huang, H.-W., Goodyear, O., Freeman, S. D., Shabanowitz, J.,

Pratt, G., Craddock, C., Williams, M. E., Hunt, D. F., and Engelhard, V. H. (2013) MHC Class I-Associated Phosphopeptides Are the Targets of Memory-like Immunity in Leukemia. *Sci. Transl. Med.* 5, 1–10

145. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., Busch, D. H., Peschel, C., Rad, R., Cox, J., Mann, M., and Krackhardt, A. M. (2016) Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* 7, 1–16
146. Engelhard, V. H., Obeng, R. C., Cummings, K. L., Petroni, G. R., Ambakhutwala, A. L., Chianese-Bullock, K. A., Smith, K. T., Lulu, A., Varhegyi, N., Smolkin, M. E., Myers, P., Mahoney, K. E., Shabanowitz, J., Buettner, N., Hall, E. H., Haden, K., Cobbold, M., Hunt, D. F., Weiss, G., Gaughan, E., and Slingluff, C. L. (2020) MHC-restricted phosphopeptide antigens: Preclinical validation and first-in-humans clinical trial in participants with high-risk melanoma. *J. Immunother. Cancer* 8,
147. Alpízar, A., Marino, F., Ramos-Fernández, A., Lombardía, M., Jeko, A., Pazos, F., Paradela, A., Santiago, C., Heck, A. J. R., and Marcilla, M. (2017) A Molecular Basis for the Presentation of Phosphorylated Peptides by HLA-B Antigens. *Mol. Cell. Proteomics* 16, 181–193
148. Mommen, G. P. M., Frese, C. K., Meiring, H. D., van Gaans-van den Brink, J., de Jong, A. P. J. M., van Els, C. A. C. M., and Heck, A. J. R. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci.* 111, 4507–4512
149. Marcilla, M., Alpízar, A., Lombardía, M., Ramos-Fernandez, A., Ramos, M., and Albar, J. P. (2014) Increased Diversity of the HLA-B40 Ligandome by the Presentation of Peptides Phosphorylated at Their Main Anchor Residue. *Mol. Cell. Proteomics* 13, 462–474
150. Lin, M.-H., Shen, K.-Y., Liu, B.-S., Chen, I.-H., Sher, Y.-P., Tseng, G.-C., Liu, S.-J., and Sung, W.-C. (2019) Immunological Evaluation of a Novel HLA-A2 Restricted Phosphopeptide of Tumor Associated Antigen, TRAP1, on Cancer Therapy. *Vaccine X*, 100017
151. Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Jayatilake, H., Gusterson, B. A., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G. J., Bigner, D. D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J. W. C., Leung, S. Y., Yuen, S. T., Weber, B. L., Seigler, H. F., Darrow, T. L., Paterson, H., Wooster, R., Stratton, M. R., and Futreal, P. A. (2002) Mutations of the BRAF gene in human cancer. *Nature* 417, 949–954
152. Depontieu, F. R., Qian, J., Zarling, A. L., McMiller, T. L., Salay, T. M., Norris, A., English, A. M., Shabanowitz, J., Engelhard, V. H., Hunt, D. F., and Topalian, S. L. (2009) Identification of tumor-associated, MHC class II-restricted phosphopeptides as targets for immunotherapy. *Proc. Natl. Acad. Sci. U. S. A.*

106, 12073–8

153. Butterfield, L. H., Comin-Anduix, B., Vujanovic, L., Lee, Y., Dissette, V. B., Yang, J. Q., Vu, H. T., Seja, E., Oseguera, D. K., Potter, D. M., Glaspy, J. A., Economou, J. S., and Ribas, A. (2008) Adenovirus MART-1-engineered autologous dendritic cell vaccine for metastatic melanoma. *J. Immunother.* 31, 294–309
154. Coulie, P. G., Brichard, V., Van Pel, A., Wölfel, T., Schneider, J., Traversari, C., Mattei, S., De Plaen, E., Lurquin, C., Szikora, J.-P., Renauld, J.-C., and Boon, T. (1994) A new gene coding for a Differentiation Antigen Recognized by Autologous Cytolytic T Lymphocytes on HLA-A2 Melanomas. *Immunogenetics* 180, 35–42
155. Kawakami, Y., Eliyahu, S., Delgado, C. H., Robbins, P. F., Rivoltini, L., Topalian, S. L., Miki, T., and Rosenberg, S. A. (1994) Cloning of the gene coding for a shared human melanoma antigen recognized by autologous T cells infiltrating into tumor. *Proc. Natl. Acad. Sci. U. S. A.* 91, 3515–3519
156. Li, Y., Depontieu, F. R., Sidney, J., Salay, T. M., Engelhard, V. H., Hunt, D. F., Sette, A., Topalian, S. L., and Mariuzza, R. A. (2010) Structural basis for the presentation of tumor-associated MHC class II-restricted phosphopeptides to CD4+T cells. *J. Mol. Biol.* 399, 596–603
157. Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017) "Hotspots" of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front. Immunol.* 8, 1–14
158. Chong, C., Marino, F., Pak, H.-S., Racle, J., Daniel, R. T., Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017) High-throughput and sensitive immunopeptidomics platform reveals profound IFNγ-mediated remodeling of the HLA ligandome. *Mol. Cell. Proteomics* 17, 533–548
159. Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H.-G., and Stevanović, S. (2017) Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. *J. Immunol.* 199, 2639–2651
160. Giam, K., Ayala-Perez, R., Illing, P. T., Schittenhelm, R. B., Croft, N. P., Purcell, A. W., and Dudek, N. L. (2015) A comprehensive analysis of peptides presented by HLA-A1. *Tissue Antigens* 85, 492–496
161. Ostrov, D. A., Grant, B. J., Pompeu, Y. A., Sidney, J., Harndahl, M., Southwood, S., Oseroff, C., Lu, S., Jakoncic, J., de Oliveira, C. A. F., Yang, L., Mei, H., Shi, L., Shabanowitz, J., English, A. M., Wriston, A., Lucas, A., Phillips, E., Mallal, S., Grey, H. M., Sette, A., Hunt, D. F., Buus, S., and Peters, B. (2012) Drug hypersensitivity caused by alteration of the MHC-presented self-peptide repertoire. *Proc. Natl. Acad. Sci.* 109, 9959–9964
162. Ramarathinam, S. H., Gras, S., Alcantara, S., Yeung, A. W. S., Mifsud, N. A., Sonza, S., Illing, P. T., Glaros, E. N., Center, R. J., Thomas, S. R., Kent, S. J., Ternette, N., Purcell, D. F. J., Rossjohn, J., and Purcell, A. W. (2018) Identification of Native and Posttranslationally Modified HLA-B*57:01-Restricted HIV Envelope Derived Epitopes Using Immunoproteomics. *Proteomics* 18, 1–11
163. Schittenhelm, R. B., Sian, T. C. C. L. K., Wilmann, P. G., Dudek, N. L., and Purcell, A. W. (2015) Revisiting the arthritogenic peptide theory: Quantitative not qualitative changes in the peptide repertoire of HLA-B27 allotypes. *Arthritis*

Rheumatol. 67, 702–713

164. Roche, P. A., and Furuta, K. (2015) The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat. Rev. Immunol.* 15, 203–216
165. Graciotti, M., Marino, F., Pak, H. S., Baumgaertner, P., Thierry, A. C., Chiffelle, J., Perez, M. A. S., Zoete, V., Harari, A., Bassani-Sternberg, M., and Kandalaft, L. E. (2020) Deciphering the mechanisms of improved immunogenicity of hypochlorous acid-treated antigens in anti-cancer dendritic cell-based vaccines. *Vaccines* 8,
166. Kranz, L. M., Diken, M., Haas, H., Kreiter, S., Loquai, C., Reuter, K. C., Meng, M., Fritz, D., Vascotto, F., Hefesha, H., Grunwitz, C., Vormehr, M., Hüsemann, Y., Selmi, A., Kuhn, A. N., Buck, J., Derhovanessian, E., Rae, R., Attig, S., Diekmann, J., Jabulowsky, R. A., Heesch, S., Hassel, J., Langguth, P., Grabbe, S., Huber, C., Türeci, Ö., and Sahin, U. (2016) Systemic RNA delivery to dendritic cells exploits antiviral defence for cancer immunotherapy. *Nature*,
167. Kreiter, S., Vormehr, M., Van De Roemer, N., Diken, M., Löwer, M., Diekmann, J., Boegel, S., Schrörs, B., Vascotto, F., Castle, J. C., Tadmor, A. D., Schoenberger, S. P., Huber, C., Türeci, O., and Sahin, U. (2015) Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature*,
168. Sahin, U., Derhovanessian, E., Miller, M., Kloke, B. P., Simon, P., Löwer, M., Bukur, V., Tadmor, A. D., Luxemburger, U., Schrörs, B., Omokoko, T., Vormehr, M., Albrecht, C., Paruzynski, A., Kuhn, A. N., Buck, J., Heesch, S., Schreeb, K. H., Müller, F., Ortseifer, I., Vogler, I., Godehardt, E., Attig, S., Rae, R., Breitkreuz, A., Tolliver, C., Suchan, M., Martic, G., Hohberger, A., Sorn, P., Diekmann, J., Ciesla, J., Waksmann, O., Brück, A. K., Witt, M., Zillgen, M., Rothermel, A., Kasemann, B., Langer, D., Bolte, S., Diken, M., Kreiter, S., Nemecek, R., Gebhardt, C., Grabbe, S., Höller, C., Utikal, J., Huber, C., Loquai, C., and Türeci, Ö. (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*,
169. Malaker, S. A., Ferracane, M. J., Depontieu, F. R., Zarling, A. L., Shabanowitz, J., Bai, D. L., Topalian, S. L., Engelhard, V. H., and Hunt, D. F. (2017) Identification and Characterization of Complex Glycosylated Peptides Presented by the MHC Class II Processing Pathway in Melanoma. *J. Proteome Res.* 16, 228–237
170. Caron, E., Aebersold, R., Banaei-Esfahani, A., Chong, C., and Bassani-Sternberg, M. (2017) A Case for a Human Immuno-Peptidome Project Consortium. *Immunity* 47, 203–208
171. Krueger, K. E., and Srivastava, S. (2006) Posttranslational protein modifications: CURRENT IMPLICATIONS FOR CANCER DETECTION, PREVENTION, AND THERAPEUTICS. *Mol. Cell. Proteomics* 5, 1799–1810
172. López-Otín, C., and Hunter, T. (2010) The regulatory crosstalk between kinases and proteases in cancer. *Nat. Rev. Cancer* 10, 278–292
173. Martín-Bernabé, A., Balcells, C., Tarragó-Celada, J., Foguet, C., Bourgoin-Voillard, S., Seve, M., and Cascante, M. (2017) The importance of post-translational modifications in systems biology approaches to identify therapeutic targets in cancer metabolism. *Curr. Opin. Syst. Biol.* 3, 161–169
174. Petersen, J., Purcell, A. W., and Rossjohn, J. (2009) Post-translationally modified T cell epitopes: Immune recognition and immunotherapy. *J. Mol. Med.* 87, 1045–

1051
175. Solleder, M., Guillaume, P., Racle, J., Michaux, J., Pak, H. S., Müller, M., Coukos, G., Bassani-Sternberg, M., and Gfeller, D. (2020) Mass spectrometry based immunopeptidomics leads to robust predictions of phosphorylated HLA class I ligands. *Mol. Cell. Proteomics* 19, 390–404
176. Ciudad, M. T., Sorvillo, N., van Alphen, F. P., Catalán, D., Meijer, A. B., Voorberg, J., and Jaraquemada, D. (2017) Analysis of the HLA-DR peptidome from human dendritic cells reveals high affinity repertoires and nonconventional pathways of peptide generation. *J. Leukoc. Biol.* 101, 15–27
177. Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., and Pandey, A. (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.* 25, 285–286
178. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372
179. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., Gerstmair, A., Faerber, F., and Kuster, B. (2014) Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587
180. Waldman, A. D., Fritz, J. M., and Lenardo, M. J. (2020) A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.*,
181. Kruger, S., Ilmer, M., Kobold, S., Cadilha, B. L., Endres, S., Ormanns, S., Schuebbe, G., Renz, B. W., D'Haese, J. G., Schloesser, H., Heinemann, V., Subklewe, M., Boeck, S., Werner, J., and Von Bergwelt-Baildon, M. (2019) Advances in cancer immunotherapy 2019 - Latest trends. *J. Exp. Clin. Cancer Res.*,
182. Rafiq, S., Hackett, C. S., and Brentjens, R. J. (2020) Engineering strategies to overcome the current roadblocks in CAR T cell therapy. *Nat. Rev. Clin. Oncol.*,
183. Havel, J. J., Chowell, D., and Chan, T. A. (2019) The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat. Rev. Cancer*,
184. Bordoli, M. R., Yum, J., Breitkopf, S. B., Thon, J. N., Italiano, J. E., Xiao, J., Worby, C., Wong, S. K., Lin, G., Edenius, M., Keller, T. L., Asara, J. M., Dixon, J. E., Yeo, C. Y., and Whitman, M. (2014) A secreted tyrosine kinase acts in the extracellular environment. *Cell*,
185. Tagliabracci, V. S., Wiley, S. E., Guo, X., Kinch, L. N., Durrant, E., Wen, J., Xiao, J., Cui, J., Nguyen, K. B., Engel, J. L., Coon, J. J., Grishin, N., Pinna, L. A., Pagliarini, D. J., and Dixon, J. E. (2015) A Single Kinase Generates the Majority of the Secreted Phosphoproteome. *Cell*,
186. Mei, S., Ayala, R., Ramarathinam, S. H., Illing, P. T., Faridi, P., Song, J., Purcell, A. W., and Croft, N. P. (2020) Immunopeptidomic Analysis Reveals That Deamidated HLA-bound Peptides Arise Predominantly from Deglycosylated Precursors. *Mol. Cell. Proteomics* 19, 1236–1247
187. Ting, Y. T., Petersen, J., Ramarathinam, S. H., Scally, S. W., Loh, K. L., Thomas, R., Suri, A., Baker, D. G., Purcell, A. W., Reid, H. H., and Rossjohn, J. (2018) The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide

binding hierarchies in rheumatoid arthritis. *J. Biol. Chem.*,

188. Poulard, C., Corbo, L., and Le Romancer, M. (2016) Protein arginine methylation/demethylation and cancer. *Oncotarget*,
189. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015) PhosphoSitePlus, 2014 : mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43, 512–520
190. Krug, K., Mertins, P., Zhang, B., Hornbeck, P., Raju, R., Ahmad, R., Szucs, M., Mundt, F., Forestier, D., Jane-Valbuena, J., Keshishian, H., Gillette, M. A., Tamayo, P., Mesirov, J. P., Jaffe, J. D., Carr, S. A., and Mani, D. R. (2019) A Curated Resource for Phosphosite-specific Signature Analysis. *Mol. Cell. Proteomics* 18, 576–593
191. Reimand, J., and Bader, G. D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*,
192. Xu, H., Wang, Y., Lin, S., Deng, W., Peng, D., Cui, Q., and Xue, Y. (2018) PTMD: A Database of Human Disease-associated Post-translational Modifications. *Genomics, Proteomics Bioinforma.*,

Appendix A – Manuscript "Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands"

Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands

Authors

Marthe Solleder, Philippe Guillaume, Julien Racle, Justine Michaux, Hui-Song Pak, Markus Müller, George Coukos, Michal Bassani-Sternberg, and David Gfeller

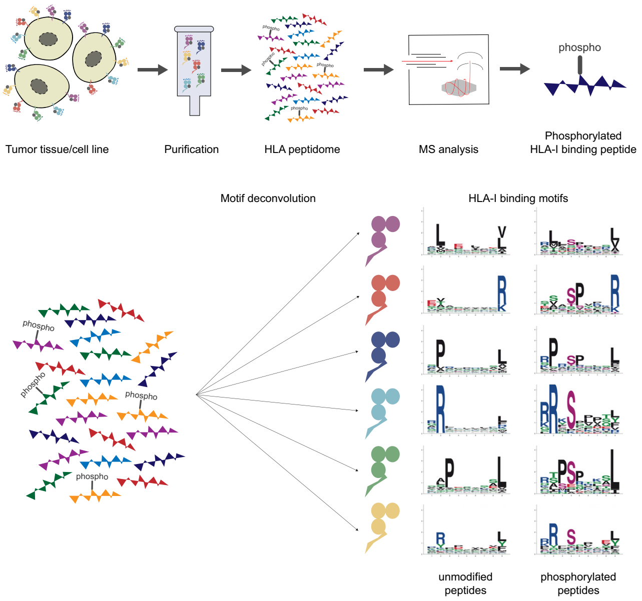
Correspondence

David.Gfeller@unil.ch;
Michal.Bassani@chuv.ch

In Brief

No HLA-I ligand predictor is available today for post-translationally modified ligands. We curated phosphorylated HLA-I ligands from immunopeptidomics studies and retrieved 2066 unique sequences. We expanded our motif deconvolution tool to identify precise binding motifs of phosphorylated HLA-I ligands, found enrichment of phosphorylated peptides among HLA-C ligands, and demonstrated a prevalent role of both HLA-I and kinase motifs on presentation of phosphorylated peptides. We further developed and validated the first predictor of interactions between HLA-I molecules and phosphorylated peptides.

Graphical Abstract



Highlights

- Curation of 2066 phosphorylated HLA class I peptides from immunopeptidomics data.
- Determination of 22 HLA class I binding motifs for phosphorylated peptides.
- Observation of a higher frequency of phosphorylated ligands binding HLA-C molecules.
- Development of a predictor of phosphorylated peptide interactions with HLA class I.

Solleder et al., 2020, Molecular & Cellular Proteomics *19*, 390–404
February 2020 © 2020 Solleder et al. Published under exclusive license by The American Society for Biochemistry and Molecular Biology, Inc.
https://doi.org/10.1074/mcp.TIR119.001641

Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands*Ⓢ

Marthe Solleder‡§, Philippe Guillaume‡, Julien Racle‡§, Justine Michaux‡¶, Hui-Song Pak‡¶, Markus Müller§, George Coukos‡¶, Michal Bassani-Sternberg‡¶‡‡, and David Gfeller‡§||

The presentation of peptides on class I human leukocyte antigen (HLA-I) molecules plays a central role in immune recognition of infected or malignant cells. In cancer, non-self HLA-I ligands can arise from many different alterations, including non-synonymous mutations, gene fusion, cancer-specific alternative mRNA splicing or aberrant post-translational modifications. Identifying HLA-I ligands remains a challenging task that requires either heavy experimental work for *in vivo* identification or optimized bioinformatics tools for accurate predictions. To date, no HLA-I ligand predictor includes post-translational modifications. To fill this gap, we curated phosphorylated HLA-I ligands from several immunopeptidomics studies (including six newly measured samples) covering 72 HLA-I alleles and retrieved a total of 2,066 unique phosphorylated peptides. We then expanded our motif deconvolution tool to identify precise binding motifs of phosphorylated HLA-I ligands. Our results reveal a clear enrichment of phosphorylated peptides among HLA-C ligands and demonstrate a prevalent role of both HLA-I motifs and kinase motifs on the presentation of phosphorylated peptides. These data further enabled us to develop and validate the first predictor of interactions between HLA-I molecules and phosphorylated peptides. ***Molecular & Cellular Proteomics 19: 390–404, 2020. DOI: 10.1074/mcp.TIR119.001641.***

Human leukocyte antigen class I (HLA-I)[1] molecules mediate cell surface presentation of peptides originating from intracellular protein degradation. Proteins are fragmented by the proteasome into short peptides. These peptides can enter the endoplasmic reticulum through the transporter associated with antigen processing (TAP) protein complex, where they are loaded onto HLA-I molecules and transported to the cell surface (1). HLA-I molecules are encoded by three genes (HLA-A, HLA-B, and HLA-C) and these genes are among the most polymorphic of the human genome, resulting in currently more than 17,000 different alleles (2). HLA-I molecules have specific binding motifs and different alleles typically bind distinct sets of peptides (3). Foreign or altered-self HLA-I ligands can be recognized by CD8+ T cells and induce an immune response to eliminate infected or malignant cells. These non-self HLA-I ligands can have various origins, such as viral proteins or genetically or post-translationally modified proteins in cancer. Identification of peptides presented HLA-I molecules is labor intensive because it relies either on challenging immunopeptidomics experiments, or predictions of HLA-I ligands followed by experimental validation. Currently, many optimized algorithms are available for predicting unmodified HLA-I ligands (4–7), but none of them include specifically post-translational modifications.

Multiple studies have identified post-translational modifications (PTMs) showing aberrant behavior in cancer (8–11), resulting for instance in abnormal cellular signaling, one of the hallmarks of cancer (12). Phosphorylation of serine, threonine, and tyrosine is one of the most frequent and best studied PTMs (13), and is carried out by different types of protein kinases, consisting mainly of serine/threonine- and tyrosine-specific protein kinases. There are over 500 known kinases in the human genome (14, 15) and different phosphorylation motifs, such as [pS/pT]P for CDK1 or MAPK1 or Rxx[pS/pT] for PKA or PKB, characterize individual kinases. It has been shown that aberrant phosphorylation can occur in cancer cells because of a deregulated balance between phosphorylation and dephosphorylation events (16), thereby altering key signaling pathways and processes within cells. A recent study estimated that phosphorylation-related single nucleotide variants are present in ~90% of tumor genomes (17) and predicted that 29% of these variants affect signaling pathways. Further, phosphosite-specific signature analysis showed to be able to identify dysregulation of phosphorylation-regulated pathways in cancer (18).

From the ‡Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland; §Swiss Institute of Bioinformatics, Lausanne, Switzerland; ¶Department of Oncology, Ludwig Institute for Cancer Research, University Hospital of Lausanne, Lausanne, Switzerland

Received June 25, 2019, and in revised form, December 6, 2019
Published, MCP Papers in Press, December 17, 2019, DOI 10.1074/mcp.TIR119.001641

Peptides with phosphorylated residues can be processed by the antigen presentation pathway, bind to HLA-I molecules, and be presented on the cell surface (19, 20, 29, 30, 21–28). Several studies reported that phosphorylated peptides could induce immune responses through T cell recognition. For instance, T cells were shown to recognize phosphorylated peptides presented on primary tumors and normal tissues and kill tumor cell lines (24, 31). Studies have reported a clear preference for the phosphorylation at position 4 on HLA-I ligands and revealed an increased presence of arginine at P1 as well as an enrichment of proline after the phosphosite caused by proline-dependent kinases (20, 22, 24–29, 32, 33). However, so far, only a handful of phosphorylated HLA-I ligands were determined for a small number of HLA-I alleles and no method is available to specifically predict their binding. As a result, predictions of HLA-I interactions with phosphorylated ligands are performed without including the modified amino acid (either by using the unmodified version of the residue or by substituting it with "X"), which is likely suboptimal because no information about phosphorylation is included in the training set of the predictors.

To fill this gap, we first measured the immunopeptidome of 6 new samples and reprocessed existing immunopeptidomics raw data for 55 other samples, in order to search for phosphorylated HLA-I ligands. We complemented these data with a small subset of HLA-I restricted phosphorylated peptides identified previously by mass spectrometry (MS), some of which came from phospho-enrichment protocols, and curated a large data set of 2,066 unique phosphorylated HLA-I ligands experimentally determined by MS. This enabled us to accurately determine phosphorylated motifs for 22 of the most frequent HLA-I alleles and revealed clear discrepancies among alleles in terms of propensity to bind phosphorylated peptides. We observed a much higher frequency of phosphorylated ligands for HLA-C alleles. We further analyzed several properties of phosphorylated HLA-I ligands and performed binding assays to validate and interpret these results. Using these data, we then developed the first predictor of phosphorylated HLA-I ligands.

EXPERIMENTAL PROCEDURES

HLA Typing—High-resolution 4-digit HLA-I typing was experimentally determined before this work and is provided for all samples in this study (supplemental Data S1). DNA was extracted from the samples for HLA typing with the DNeasy Blood & Tissue Kit (Qiagen, Germantown, Maryland), following the manufacturer's protocols. The amplification of the HLA genes was conducted with the TruSight HLA v2 Sequencing Panel kit (CareDx, Brisbane, California) according to the manufacturer's protocol. Sequencing was performed on the Illumina® MiniSeq™ System (Illumina, San Diego, California) using paired-end 2 × 150 bp protocol. The data was analyzed with the Assign TruSight HLA v2.1 software (CareDx).

Preparation of HLA Class I Peptide Samples—Several tissue samples, 3993, 4052-BA, 3989-HT, OE37–1N, OVZW-1P, and OXVD-09, were provided by the biobank of the Center of Experimental Therapies at the CHUV after informed consent of the participants was obtained following requirements of the institutional review board (Ethics Commission, CHUV). 2–5 biological replicates per tissue were processed using our previously described protocol (34). Briefly, tissues were homogenized on ice in lysis buffer with Ultra Turrax homogenizer (IKA, Staufen, Germany) for 10 s at maximum speed, and then incubated on ice for 1 h. Lysis buffer contained 0.25% sodium deoxycholate (Sigma-Aldrich, St. Louis, Missouri), 0.2 mM iodoacetamide (Sigma-Aldrich), 1 mM EDTA, 1:200 Protease Inhibitors Mixture (Sigma-Aldrich), 1 mM Phenylmethylsulfonylfluoride (Roche, Mannheim, Germany), 1% octyl-beta-D glucopyranoside (Sigma-Aldrich) in PBS. Subsequently, 20 min centrifugation for clearance (table-top centrifuge, Eppendorf Centrifuge 5430R, Schönenbuch, Switzerland) was performed at 4 °C at 14,200 rpm. Immuno-affinity purification through Protein-A Sepharose beads covalently bound to W6–32 antibodies was performed in a format of 96-well single-use micro plate with 10 μm polypropylene membranes (SeaHorse Bioscience, North Billerica, Massachusetts). The plates were then washed 4 times with 2 ml 150 mM NaCl and 20 mM Tris HCl (buffer A), 4 times with 2 ml 400 mM NaCl and 20 mM Tris Hcl, further 4 times 2 ml buffer A and final twice with with 20 mM Tris HCl, pH 8. HLA molecules and peptides were eluted with 1% trifluoroacetic acid (TFA, Merck, Darmstadt, Germany) directly into Sep-Pak tC18 100 mg Sorbent 96-well plates (Waters, Milford, Massachusetts) pre-conditioned with 80% acetonitrile (ACN) in 0.1% TFA and with 0.1% TFA only. Wells were washed twice with 0.1% TFA and then the peptides were eluted with 28% ACN in 0.1% TFA. Peptides were dried using vacuum centrifugation (Eppendorf Concentrator Plus, Schönenbuch, Switzerland) and were resuspended in a final volume of 12 μl 0.1% formic acid. 3 μl of these peptides were used for each MS run.

Mass Spectrometry Analysis of HLA Class I Peptides—HLA peptides were separated by a nanoflow HPLC (Proxeon Biosystems, Thermo Fisher Scientific, Odense, Denmark) on 50 cm long column (75 μm inner diameter) self-packed with ReproSil-Pur C18-AQ 1.9 μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) in 0.1% formic acid coupled on-line to a Q Exactive HF-X mass spectrometers (Thermo Fisher Scientific, Bremen, Germany) with a nanoelectrospray ion source (Proxeon Biosystems). HLA-I peptides were eluted with a linear gradient of 2–30% of 80% ACN and 0.1% formic acid at a flow rate of 250 nl/min over 125 min. MS spectra were acquired from m/z = 300–1,650 in the Orbitrap with a resolution of 60,000 (m/z = 200) and ion accumulation time of 80 ms. The auto gain control was set to 3e6 ions. MS/MS spectra were acquired on 10 most abundant precursor ions with a resolution of 15,000 (m/z = 200), ion accumulation time of 120 ms and an isolation window of 1.2 m/z. The auto gain control was set to 2e5 ions. Dynamic exclusion to 20 s and a normalized collision energy of 27 was used for fragmentation. The peptide match option was disabled. No fragmentation was performed in case of assigned precursor ion charge states of four and above.

Identification of HLA Class I Peptides—We employed the MaxQuant platform (35) version 1.5.5.1 to search the MS peak lists against a fasta file containing the human UniProt database containing 42,170 entries including isoforms (March 2017) and a list of 247 frequently observed contaminants. Peptides with a length between 8 and 15 amino acids were allowed. The second peptide identification option in Andromeda was enabled. The enzyme specificity was set as unspecific and FDR of 5% was required for peptides and no protein FDR was set. As a large score difference to the second best match (delta score) is important for identification of phosphorylated peptides (36), the delta score was set to a minimum of 10 for both modified and unmodified peptides (see Results and supplemental Fig. S1 for com-

[1] The abbreviations used are: HLA-I, human leukocyte antigen class I; TAP, transporter associated with antigen processing; PTM, post-translational modifications; FDR, false discovery rate.

parison among different delta score thresholds). The initial allowed mass deviation of the precursor ion was set to 6 ppm and the maximum fragment mass deviation was set to 20 ppm. Methionine oxidation (15.994915 Da), N-terminal acetylation (42.010565 Da) and phosphorylation (79.9663304 Da) on serine, threonine and tyrosine were set as variable modifications.

Experimental Design and Statistical Rationale—In addition to the six novel samples mentioned above, 38 MS samples from published immunopeptidomics studies were reanalyzed together, 209 raw files in total (6, 26, 34, 37, 38) (see supplemental Data S1 and S2). For each sample, at least two technical replicates of raw MS files were included. In a separate run, the MaxQuant platform was employed with the same parameters on 85 MS raw files of 17 monoallelic samples with five technical replicates of raw MS files for each sample (39) (supplemental Data S3), and the data was similarly filtered to obtain peptide spectrum matches with high confidence.

Curation of Immunopeptidomics HLA-I MS Data Sets—We filtered the list of identified phosphorylated HLA-I peptides listed in the MaxQuant MSMS output table by removing reverse hits and peptides matching contaminants. To maintain peptide spectrum matches with high confidence, the list was further filtered by restricting the identification score ≥70, and the localization probabilities to ≥0.75. Only unique modified and unmodified sequences were further analyzed (see supplemental Data S4 for all identified phosphorylated HLA-I peptides and supplemental Data S5 for all alleles for which phosphorylated HLA-I peptides were found).

Additionally, data from various publications (20, 21, 33, 40–43, 22–25, 27–29, 32) was added to our data set, using from each sample both, known phosphorylated as well as unmodified HLA-I binders if available. Identified phosphorylated HLA-I peptides from enrichments studies (20–22, 25, 28) (see supplemental Data S4 for details) were included in the determination of phosphorylated HLA-I binding motifs and the training of the predictor, but not in the comparison of the fraction of phosphorylated ligands for different HLA-I molecules.

HLA-I Motif Deconvolution for Identification of HLA-I Binding Motifs—To determine allelic restriction among HLA-I ligands found by MS, including phosphorylated peptides, we expanded our motif deconvolution tool MixMHCp (6, 44) to allow for additional non-standard amino acids. Briefly speaking, the motif deconvolution method infers with Expectation-Maximization algorithm K different position weight matrices that optimally model the list of peptides. In this extended version of the motif deconvolution algorithm, phosphorylated residues are treated as additional amino acids, leading to an alphabet of size 23 (*i.e.* position weight matrices of size 9 × 23, instead of 9 × 20 as described in our previous manuscript (44)). Finally, motifs were assigned to their respective HLA-I allele using the approach described in (6, 37). Briefly, binding motifs from our samples were first annotated by identifying common motifs across samples sharing the same alleles, and these motifs were further compared with those from previous studies or from IEDB (45). These results were manually checked to exclude ambiguous cases, which were excluded from our data, both in terms of unmodified and phosphorylated peptides. Of note, MixMHCp also contains a flat motif to which peptides that do not match any of the motifs inferred by the algorithm are assigned. The command-line script to run the motif deconvolution (MixMHCp2.1) can be obtained at https://github.com/GfellerLab/MixMHCp.

Visualization of HLA-I Phosphorylated Motifs—Binding motifs of HLA-I alleles were visualized by sequence logos. The sequence logos were generated by modifying the R package ggseqlogo (46) in a way to include sequences with modified amino acids. Purple letters were used to visualize phosphorylated residues in sequence logos of HLA-I binding motifs. Phosphorylated motifs for HLA-I alleles with more than 22 phosphorylated ligands are displayed in Fig. 1. The modified version of ggseqlogo to plot sequence logos including modified residues is provided at https://github.com/GfellerLab/ggseqlogo.

Analysis of Phosphorylated HLA-I Ligands—After assigning unmodified and phosphorylated peptides to alleles through motif deconvolution, peptides were pooled from all samples for each allele and merged into a unique set of peptides per allele. Phosphorylated and unmodified binding motifs for each HLA-I allele were built. The overall frequency of phosphorylated peptides per allele was analyzed by computing the fraction of phosphorylated peptides among all discovered peptides per allele from any length and for each peptide length ranging from 8 to 12 amino acids separately. To identify potential structural differences between binding regions of alleles with high and low frequency of phosphorylated peptides, alleles were split into two groups based on the median frequency in HLA-A, -B, or -C alleles, respectively. HLA-I binding sites were analyzed by (1) selecting positions of the binding regions that show interaction with the peptide in the 3D structure (28 positions in total) and (2) computing the Euclidean distances for these selected positions of the binding regions between the two groups. For each group of alleles, the block of binding site sequences was transformed into position weight matrices (M_{ij}) with $i = 1, \ldots, 20$ and $j = 1, \ldots, 28$ by calculating the frequency of each amino acid i at each position j. For each position j, the Euclidean distance between the columns of the matrices ($M^{(1)}$ and $M^{(2)}$) was computed as:

$$\left[\sum_{i=1}^{20}(M_{ij}^{(1)} - M_{ij}^{(2)})^2\right]^{1/2} \qquad \text{(Eq. 1)}$$

Sequence logos were used to visualize the ten most different positions in each comparison for binding sites of alleles with high frequency *versus* alleles with low frequency of phosphorylated peptides.

For each allele in the underlying data set, we calculated how many unmodified ligands contained phosphosites from the phosphoproteome (47). Phosphosites positioned at P4 in unmodified HLA-I 9-mers were counted for all alleles with more than 50 unmodified ligands. The frequency of known phosphosites per allele was computed as the fraction of detected phosphosites at P4 within the unmodified 9-mer HLA-I ligands. The correlation between phosphorylated HLA-I ligands per allele and the amount of detected phosphosites within unmodified HLA-I ligands was measured using the Pearson correlation coefficient.

The length distribution ranging from 8- to 12-mers was computed for phosphorylated and unmodified HLA-I ligands per allele and error bars show the variability across alleles.

The distribution of phosphorylated amino acids (pS, pT, and pY) in the human phosphoproteome was obtained from (47) and compared with the one observed in the whole phosphorylated immunopeptidome and separately for each length 8 to 12. p values were calculated by t-tests. Further, among all unique phosphorylated HLA-I ligands, we measured how often each position in any 8- to 12-mer was phosphorylated.

To test if proline enrichment exists in our data set, the proline frequency in phosphorylated and unmodified HLA-I ligands was analyzed. First, the frequency of proline occurring next to a phosphorylated residue was measured in all phosphorylated peptides per allele, for all alleles with at least 5 phosphorylated peptides. Second, the overall proline frequency in unmodified HLA-I peptides at non-anchor positions (3 to 8 for HLA-I 9-mers) was extracted allele-wise. As a third measurement, the proline frequency in the human proteome (UniProt as of October 2017) was also included as a comparative means in the analysis of proline enrichment. p values were computed by t-tests among the different groups of data.

To compute the enrichment in arginine at P1, for each allele with at least 5 phosphorylated HLA-I ligands the occurrence of arginine at P1 was calculated among all phosphorylated peptides with a phosphorylated P4 (phosphorylated serine, threonine or tyrosine), which is the most frequent phosphorylated position within 9-mers. These values were compared with (1) the frequency of arginine at P1 in unmodified HLA-I ligands with serine, threonine, or tyrosine at P4 and (2) the overall frequency of arginine in the human proteome (UniProt as of October 2017). *p* values comparing the different measurements were computed with *t*-tests.

Binding motifs of different kinases were determined with the Phospho.ELM data set (48) and sequence logos of 3 positions upstream and 3 positions downstream of the phosphosite were visualized with the modified version of ggseqlogo.

Experimental Testing of HLA-I–Phosphorylated Ligand Binding—Experimental testing of HLA-I ligands was performed as described before (6), consisting of refolding assays, followed by ELISA assays. ELISA absorbance signals were used to define binding stabilities of phosphorylated and unmodified versions of several peptides and several alleles. Two replicates per experiment were performed and negative controls correspond to experiments performed in the absence of a peptide. Measured absorbance of the binding assays were normalized by t = 0 h of the positive controls. Half-lives were computed as ln(2)/k_{off}. The background signal (i.e. measurements from the negative controls) was removed from the measured ELISA absorbance values and k_{off} was determined through fitting exponential curves to absorbance values.

Mutation of position 69 for arginine to alanine in the heavy chain of HLA-C*06:02 was done by site-directed mutagenesis by overlap extension using the polymerase chain reaction (PCR). Two PCR products are obtained from the HLA-C*06:02 BSP coding sequence using as forward primer 5′-GATATACATATGTGCTCCCACTCCATGAGG-3′ (primer A) and reverse primer containing the mutation (in bold and underlined) 5′-CACTCGGTCAGCCTGTGCCTG**<u>GGC</u>**CTTGTACTTCT-GTGTCTCCCG-3′ (primer B) and second PCR with forward primer containing the mutation (in bold and underlined) 5′-CGGGAGACACA-GAAGTACAAG**<u>GCC</u>**CAGGCACAGGCTGACCGAGTG-3′ (primer C) and reverse primer 5′-GGCCGCAAGCTTTTAGTGCCATTCGATT-TTCTGAGC-3′ (primer D). The two PCR products are mixed in a third PCR with primer A and D. The coding sequence was cloned between NdeI and HindIII sites in plasmid pET-23a. Expression of mutated R69A HLA-C*06:02 was performed by using the *Escherichia coli* strain BL21(DE3)(pLysE).

Predictor and Cross Validation—For each HLA-I allele with at least 20 phosphorylated HLA-I 9-mer ligands, position weight matrices (PWM) were built. PWMs are then used to calculate a peptide score for each peptide ($X_1, \ldots, X_L$):

$$S = \frac{1}{L}\sum_{i=1}^{L} \log\left(\frac{p_{x_i,i}}{q_{x_i}}\right) \qquad \text{(Eq. 2)}$$

The peptide score describes for a peptide with which frequency each amino acid occurs at its position in the binding motif of the allele. *L* corresponds to the length of the predicted peptide, $p_{x_i,i}$ is the PWM entry at position *i* for amino acid X_i, and q_{x_i} is a background frequency. Here, average frequencies of each amino acid within the human phosphoproteome are used as background frequencies. For each allele, peptides are ranked according to their score to identify most likely binders.

Before calculating the peptide score *S*, a pseudocount is added to the PWM, as described in (49). This is done to prevent zero occurrence of any amino acid at any position in the PWM, which may arise especially for small training data sets. The pseudocount for phosphorylated PWMs (PWMs of size 9 × 23 for 9-mers) is based on the work on the BLOSUM62 alignment score (50). The transition probabilities from the original BLOSUM62 were used for unmodified amino acids. BLOSUM62 was then expanded to include the three phosphorylated residues (phosphorylated serine, phosphorylated threonine, and phosphorylated tyrosine), based on the BLOSUM62 transition probabilities of unmodified serine, threonine and tyrosine. The phosphorylated-BLOSUM62 was extended by (1) transition probabilities from each phosphorylated amino acid to any of the three phosphorylated as well as any of the 20 unmodified amino acid, and (2) transition probabilities from any unmodified amino acid into each of the three phosphorylated residues. In more details, for phosphorylated residues *p*ϵ(s,t,*y*) and unmodified residues *U*ϵ(*A*,*C*,*D*,. . .,*Y*) BLOSUM62 was extended into phospho-BLOSUM62 in the following way: transitions *b* from any phosphorylated amino acid p_1 to any other phosphorylated amino acid p_2 were defined as

$$b_{p_1,p_2} = \frac{B_{p_1^*,p_2^*}}{B_{p_1^*,p_1^*}} b_{p_1,p_1} \qquad \text{(Eq. 3)}$$

with *p*∗ denoting the corresponding unmodified amino acid of phosphorylated amino acid *p* and *B* corresponding to transition probabilities from the original BLOSUM62. Further,

$$\sum_{p_2=1}^{3} b_{p_1,p_2} = 0.9 \qquad \text{(Eq. 4)}$$

was defined to strengthen the transition probability of a phosphorylated amino acid p_1 to stay phosphorylated and to have only little probability to transform into any unmodified residue (sum of transition probability into unmodified residues being 0.1). The transition *b* from phosphorylated residue p_1 to any unmodified amino acid *U* was set to

$$b_{p_1,U} = \frac{B_{p_1^*,U}}{10} \qquad \text{(Eq. 5)}$$

Transitions of the unmodified counterpart of p_1 to *U* from the original BLOSUM62 were used to reflect the relationship between two amino acids also in the transitions from the phosphorylated version of the amino acid into all unmodified residues, but only contributing with a reduced weight to the total transitions for a phosphorylated amino acid (factor 10 is used to reduce the row sum of the original BLOSUM62 from 1 to 0.1, considering Eq. 4). For any unmodified amino acid *U* the transition probability into a phosphorylated residue *p* was defined as

$$b_{U,p} = \frac{B_{U,p^*}}{10} \qquad \text{(Eq. 6)}$$

Reduced transition probabilities were defined to provide a low transition probability from unmodified into phosphorylated amino acids yet modeling a similar proportion for the transitions into phosphorylated versions of the amino acids like the transitions into unmodified serine, threonine, and tyrosine. For each unmodified amino acid, a row-wise normalization over the transitions to all unmodified and all phosphorylated amino acids was performed. The phosBLOSUM62 matrix is given in supplemental Data S6.

Various versions of the predictor with different training data compositions were developed to find the best prediction method. One predictor was trained only on phosphorylated peptides. A second predictor included all unmodified HLA-I ligands of the allele in addition to the phosphorylated peptides in the training set. In addition, to avoid over-fitting of the predictor, if the unmodified version of a phosphorylated peptide in the testing data was present among unmodified HLA-I ligands, it was removed from the training data set. A third predictor was trained exclusively on unmodified HLA-I binders. In this version of the predictor every phosphorylated residue in the prediction data was treated like an unmodified amino acid, because no information about phosphorylated residue occurrence was given in the PWM. Further, the performance of the different versions of the predictor is also benchmarked against NetMHCpan4.0 (5). NetMHCpan was carried out on the same prediction data replacing the phosphorylated residues by "X" and peptides were ranked according to their Rank score.

To validate the performance of the above-described predictors with varying training data, a 5-fold cross validation was performed on all alleles with more than 20 phosphorylated HLA-I peptides of length 9 and run 100 times. For each allele the set of phosphorylated peptides was randomly divided into five groups, four were used as training data for the predictor and the remaining one was used as positive testing data. In addition, four times the amount of positive peptides was added as negative data to the testing data. Negative peptides were randomly selected from a pool of all known phosphosites of the human proteome, excluding phosphosites at anchor positions 2 and 9. The Area Under the ROC Curve (AUC), AUC0.1 as well as precision of the top 20% of predicted phosphorylated peptides (corresponding to recall as four fifth of the prediction data was negative data) were used to measure the performance of the different prediction models (51) and shown as a mean over all 100 runs averaged over all alleles and for each allele separately. *p* values comparing the results of the cross validation were computed using Wilcoxon signed-rank test.

The code of the predictor of phosphorylated HLA-I ligands is available at https://github.com/GfellerLab/PhosMHCpred.

RESULTS

Identification of Phosphorylated HLA-I Ligands Uncovers Phosphorylated HLA-I Binding Motifs—We collected data from six new samples and curated 55 publicly available immunopeptidomics studies (6, 26, 34, 37–39) comprising both pooled and mono-allelic data sets (see supplemental Data S1, S2, and S3). None of these MS studies were performed with phospho-enrichment protocols. Raw MS data were reprocessed allowing for phosphorylation on serine, threonine and tyrosine as variable modifications (see Experimental Procedures). To gain sensitivity while maintaining peptide spectrum matches with relatively higher confidence that is typically achieved with the more conservative false discovery rate (FDR) of 1%, we applied FDR of 5% and considered peptides identified with Andromeda search engine score ≥70, score difference to the second best peptide spectrum match (delta score) ≥10 and localization probability ≥0.75. This resulted in 2,190 unique phosphorylated peptides in total for all 61 samples. Supplemental Fig. S1*A* shows the distribution of Andromeda search engine peptide spectrum match score and delta score for phosphorylated peptides with different localization probabilities. To determine allelic restriction, we expanded our motif deconvolution algorithm MixMHCp (6, 44) to consider both the phosphorylated and unmodified peptides in each sample (see "Experimental procedures"). MixMHCp removes potentially wrongly identified peptides that do not match the inferred motifs by assigning them to a so-called flat motif (6). 1,841 unique phosphorylated peptides (84.1%) were assigned to HLA-I motifs following this deconvolution step, the remaining phosphorylated peptides were assigned to the flat motif and excluded from downstream analyses. 31% of the peptides identified with delta score <20 were assigned to the flat motif compared with only ~12% of the relatively more reliable identifications of peptides with delta score ≥20 (see supplemental Fig. S1*B*). We then compared different properties (*i.e.* peptide length, position of the phosphosite and frequency of the different phosphorylated residues) among phosphorylated HLA-I ligands with delta score ≥10 or ≥20. Peptides assigned to HLA-I motifs by MixMHCp displayed very similar properties across different choices of delta score thresholds (supplemental Fig. S1*C*). Reversely, peptides assigned to the flat motif showed a dramatically different behavior in the aforementioned properties (supplemental Fig. S1*C*). Together with the fact that these peptides were enriched in low delta scores, this suggests that the filtering with MixMHCp efficiently removes several wrongly identified peptides and can act as additional filter for phosphorylated immunopeptidomics data. Similar properties were observed between phosphorylated peptides assigned to HLA-I motifs when using FDRs of 5% or 1% (supplemental Fig. S1*D*). Finally, binding motifs of phosphorylated peptides were robust to different choices of thresholds on delta scores and FDRs (see supplemental Fig. S2).

We further curated phosphorylated HLA-I ligands with known allelic restriction reported in earlier studies (20, 21, 33, 40–43, 22–25, 27–29, 32), including phosphorylated peptides from five samples analyzed with phospho-enrichment protocols (20–22, 25, 28). In the final step, we restricted HLA-I peptides to length 8–12 for further analysis. Altogether, a total number of 2,066 unique phosphorylated peptide sequences were retrieved, representing 2,585 unique HLA-I-phosphorylated peptide interactions with 72 different HLA-I alleles (20 HLA-A, 30 HLA-B, 21 HLA-C alleles and 1 HLA-G allele, see supplemental data 4 for all identified phosphorylated peptides). 740 of the 2,585 (28.63%) unique HLA-I-phosphorylated peptide interactions had been reported in previous studies. Phosphorylation occurred mostly once per peptide (97.77%) and very few multiple phosphorylated peptides were found (2.23% double phosphorylated peptides, see also supplemental Fig. S3*A*). Among the 2,066 phosphosites, 800 (38.7%) were not observed in phosphosite databases (dbPAF, phosphoELM, and phosphositePLUS (52–54)), which is in line with what had been previously reported (26). Comparison of binding motifs of all phosphorylated peptides with binding motifs excluding unknown phosphosites showed that these were very similar (see supplemental Fig. S2, column 1 and 4). 171 unique phosphorylated HLA-I ligands (8.3%) were only found in the five phosphorylation-enriched samples from previous studies (20–22, 25, 28) and not in any of the other unenriched samples included in our work (see supplemental Data S5 for details). This demonstrates that many phosphorylated HLA-I ligands can be detected in samples that are not specifically enriched in phosphorylation residues, as already shown in previous studies (26). ~30% of the unique interactions between HLA-I alleles and phosphorylated peptides were also detected in their unmodified version. A GO enrichment analysis of source proteins of all HLA-I ligands showed very similar proteins of origin for phosphorylated and unmodified peptides (supplemental Fig. S3*B*).

22 HLA-I alleles had more than 20 unique phosphorylated 9-mer ligands and their phosphorylated and unmodified binding motifs are shown in Fig. 1. These 22 alleles were later used to develop the predictor for HLA-I-phosphorylated peptide interactions and have a population frequency of 91.76% on average worldwide (99.41% in Europe).

Phosphorylated Peptides are Enriched Among HLA-C Ligands—We analyzed the fraction of phosphorylated HLA-I ligands for alleles derived from each gene (HLA-A, HLA-B, and HLA-C). Phosphorylation-enriched samples (20–22, 25, 28) (see supplemental Data S4 and S5 for detailed information) were not considered in this analysis to prevent biases in the estimation of the frequency of phosphorylated HLA-I ligands. For all peptide lengths combined, we observed variability across different alleles and a significant enrichment of phosphorylated peptides among HLA-C alleles compared with HLA-A and HLA-B (Fig. 2*A* and see supplemental Fig. S4*A* for separate lengths). To show that this could not be ascribed to the peptides from the monoalleleic HLA-C samples (39), we performed the same analysis without these samples, and obtained similar results (supplemental Fig. S4*B*, left panel). Further, we could see that these results were also robust when only analyzing phosphorylated peptides identified with delta score ≥20 (supplemental Fig. S4*B*, right panel). To investigate why specific alleles, and especially HLA-C alleles, show a higher fraction of phosphorylated peptides in our data set, we measured the binding stability differences (*i.e.* half-life ratio) between phosphorylated and unmodified peptides for several HLA-A, -B, and -C alleles (HLA-A*01:01, HLA-A*25:01, HLA-B*07:02, HLA-B*18:01, HLA-C*06:02, and HLA-C*07:02) with both high and low phosphorylated peptide frequency in immunopeptidomics data (arrows in Fig. 2*A*). Overall, the results indicate significantly higher half-life ratio for HLA-C alleles compared with HLA-A or HLA-B alleles (Fig. 2*B*). However, the selected HLA-A and HLA-B alleles showed opposite binding preference compared with what would have been expected from the frequency of phosphorylated peptides in the immunopeptidome (Fig. 2*B*, compare HLA-A*01:01 and HLA-A*25:01, HLA-B*07:02, and HLA-B*18:01).

HLA-C molecules are characterized by the specific presence of R at position 69 (A/T in HLA-A or -B), which may interact with phosphorylated residues (supplemental Fig. S4*C*). However, R69A mutation did not show a sharper decrease in binding affinity for phosphorylated ligands (supplemental Fig. S4*D*), suggesting that this residue is not responsible for the preference of HLA-C alleles for phosphorylated peptides. Moreover, comparison of binding sites between alleles with high and low fractions of phosphorylated ligands did not suggest clear differences that could favor the binding of phosphorylated peptides (supplemental Fig. S4*E*).

The lack of correlation between stability measurements and fraction of phosphorylated peptides in MS data within alleles from the same gene and the lack of molecular features explaining the variability observed across alleles (Fig. 2*A*–2*B* and supplemental Fig. S4*A*–S4*E*) led us to hypothesize that this variability may be related to a better compatibility of specific HLA-I motifs with phosphorylation motifs. To explore this hypothesis, we analyzed the unmodified ligands of each allele in our data set by checking for the occurrence of human phosphosites from the phosphoproteome (47). The results showed that the frequency of unmodified ligands containing known phosphosites at P4 is on average higher in HLA-C than in HLA-A or HLA-B alleles (Fig. 2*C*). We could further detect a significant positive correlation between the amount of phosphosites of the human phosphoproteome found within unmodified ligands and the frequency of phosphorylated ligands detected per allele (supplemental Fig. S4*F*). These results support the hypothesis that HLA-I alleles that preferentially bind phosphorylated peptides (especially HLA-C alleles) have motifs that are better suited to bind peptides coming from known phosphosites. (Fig. 2*A* and supplemental Fig. S4*A*). Overall, our analysis shows large variability in the fraction of phosphorylated ligands across alleles and suggest that some of this variability comes from a better compatibility between the HLA-I motifs of specific alleles and phosphorylation motifs.

When we compared the fraction of phosphorylated peptides for different lengths, we could observe that longer peptides are enriched with phosphosites (Fig. 2*D* and supplemental Fig. S4*A*). Further, phosphorylated residues (phosphorylated serine, threonine, and tyrosine) were observed at similar frequencies as in the human phosphoproteome (47) (Fig. 2*E*), and binding assays indicated no difference in binding stability for different phosphorylated residues (Fig. 2*F*).

Phosphorylated HLA-I Ligands Show a Preference for Phosphosites at P4 Which Does Not Only Result From Higher Binding Stability—Fig. 3*A* shows the distribution of phosphorylated positions for peptides of lengths 8 to 12 in our phosphorylated immunopeptidome. Like what has been shown before (24, 26, 28, 32, 33), we could detect a clear preference for phosphorylation at P4 in all lengths and an additional preference for P6 in 10- and 12-mer phosphorylated HLA-I ligands. To explore the biochemical reason for this preference at P4, the binding of 9-mer peptides with phosphorylated serine at non-anchor positions 3 to 8 was tested. As expected, HLA-A*02:01 and HLA-B*07:02 binding peptides with a phosphorylated P4 bind more stably compared with peptides with phosphorylated positions 3 and 5 to 7. This trend was also observed for a peptide that was originally found in the MS data with a phosphorylated P3 (third peptide of HLA-B*07:02 in the panel). Further and less expectedly, the results demonstrated that phosphorylation at P8 also shows an increased binding compared with phosphorylated positions 3 and 5 to 7 (Fig. 3*B*), especially for ligands tested for HLA-B*07:02.

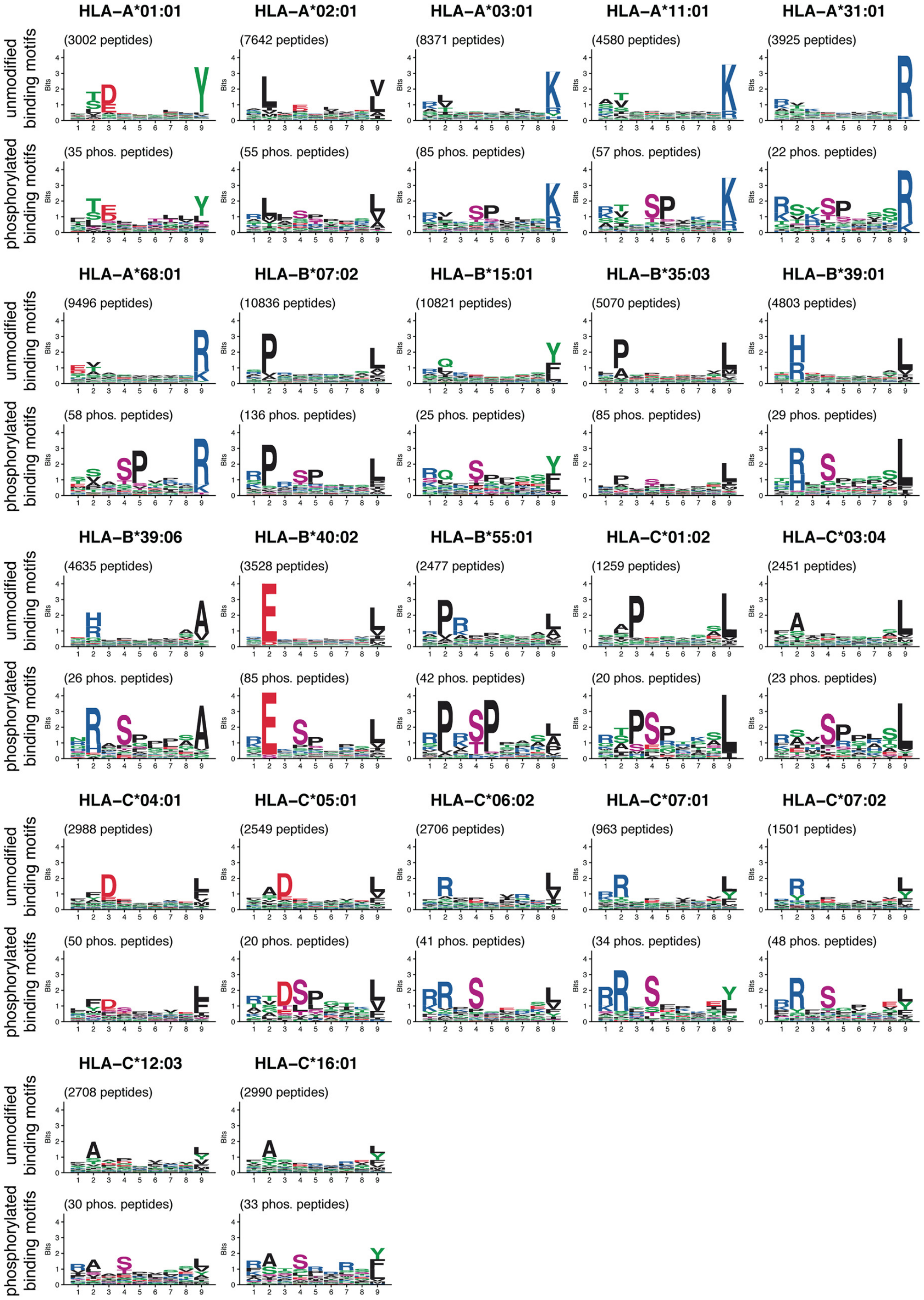
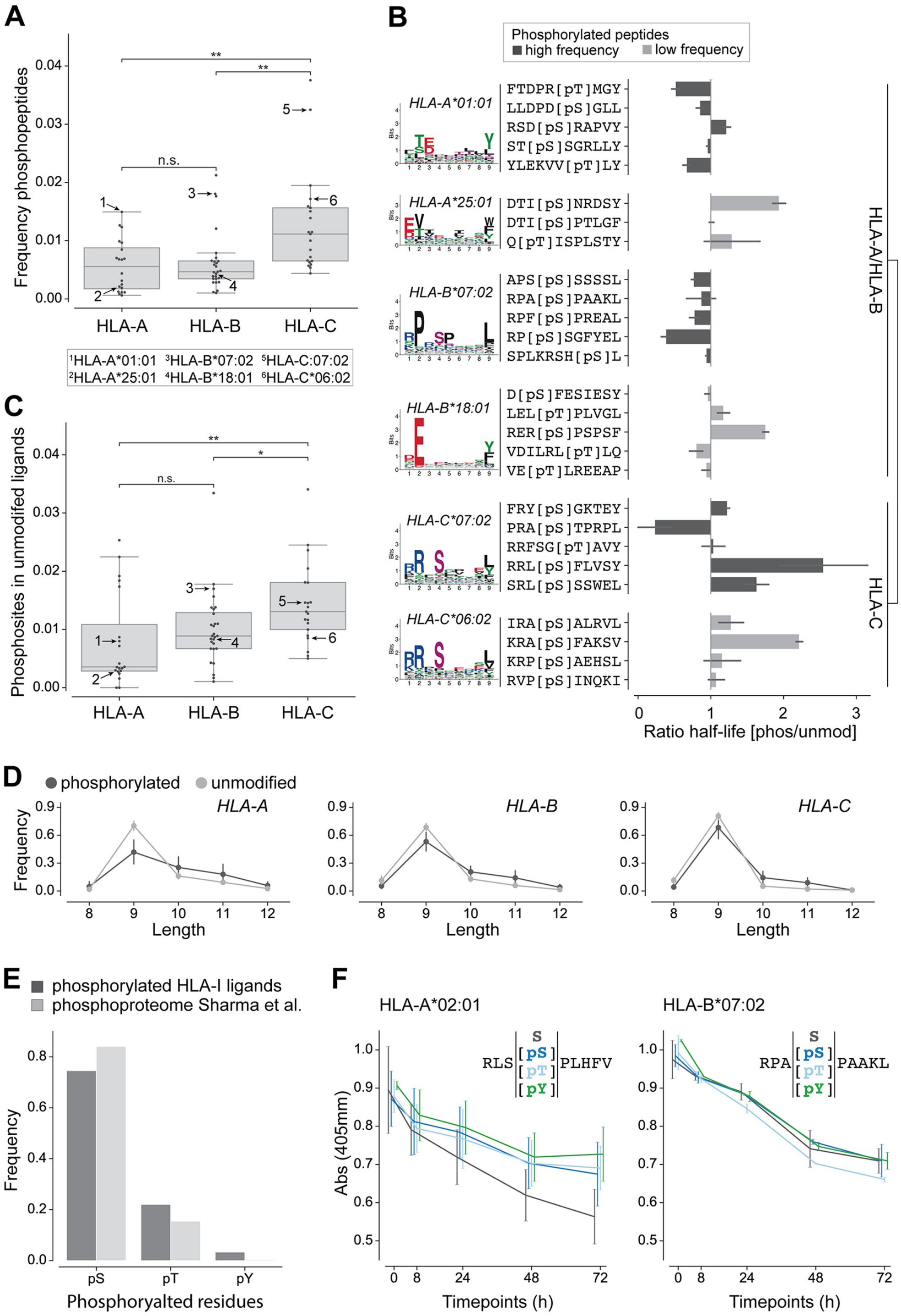


FIG. 1. **Overview of 9-mer HLA-I binding motifs of unmodified (top) and phosphorylated (bottom) ligands for HLA-I alleles with at least 20 phosphorylated ligands (9-mers) determined in this work.** Phosphorylated residues are shown in purple.


A
Frequency phosphopeptides
0.04
0.03
0.02
0.01
0.00
**
**
n.s.
1
2
3
4
5
6
HLA-A
HLA-B
HLA-C
¹HLA-A*01:01 ³HLA-B*07:02 ⁵HLA-C:07:02
²HLA-A*25:01 ⁴HLA-B*18:01 ⁶HLA-C*06:02
B
Phosphorylated peptides
■ high frequency ■ low frequency
HLA-A*01:01
FTDPR[pT]MGY
LLDPD[pS]GLL
RSD[pS]RAPVY
ST[pS]SGRLLY
YLEKVV[pT]LY
HLA-A*25:01
DTI[pS]NRDSY
DTI[pS]PTLGF
Q[pT]ISPLSTY
HLA-B*07:02
APS[pS]SSSSL
RPA[pS]PAAKL
RPF[pS]PREAL
RP[pS]SGFYEL
SPLKRSH[pS]L
HLA-B*18:01
D[pS]FESIESY
LEL[pT]PLVGL
RER[pS]PSPSF
VDILRL[pT]LQ
VE[pT]LREEAP
HLA-C*07:02
FRY[pS]GKTEY
PRA[pS]TPRPL
RRFSG[pT]AVY
RRL[pS]FLVSY
SRL[pS]SSWEL
HLA-C*06:02
IRA[pS]ALRVL
KRA[pS]FAKSV
KRP[pS]AEHSL
RVP[pS]INQKI
HLA-A/HLA-B
HLA-C
p < 0.05
0
1
2
3
Ratio half-life [phos/unmod]
C
Phosphosites in unmodifed ligands
0.04
0.03
0.02
0.01
0.00
**
*
n.s.
HLA-A
HLA-B
HLA-C
D
● phosphorylated ● unmodified
HLA-A
HLA-B
HLA-C
Frequency
0.9
0.6
0.3
0.0
8
9
10
11
12
Length
E
■ phosphorylated HLA-I ligands
■ phosphoproteome Sharma et al.
Frequency
0.8
0.6
0.4
0.2
0.0
pS
pT
pY
Phosphoryalted residues
F
HLA-A*02:01
HLA-B*07:02
RLS[S/pS/pT/pY]PLHFV
RPA[S/pS/pT/pY]PAAKL
Abs (405mm)
1.0
0.9
0.8
0.7
0.6
0.5
0
8
24
48
72
Timepoints (h)

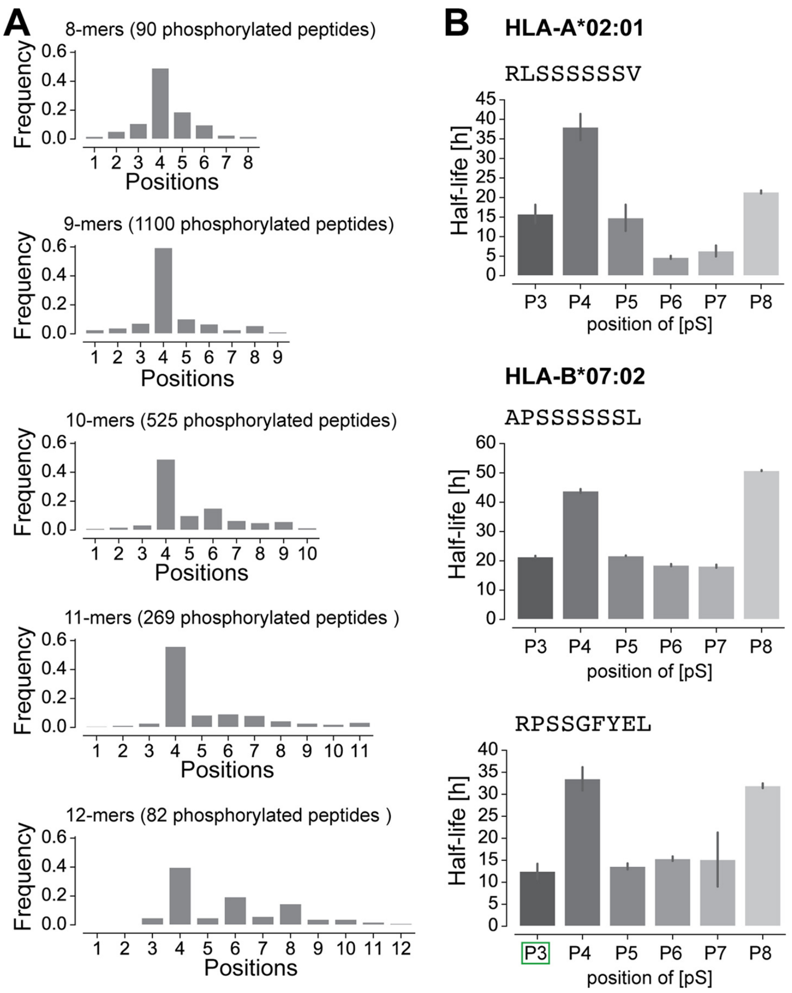
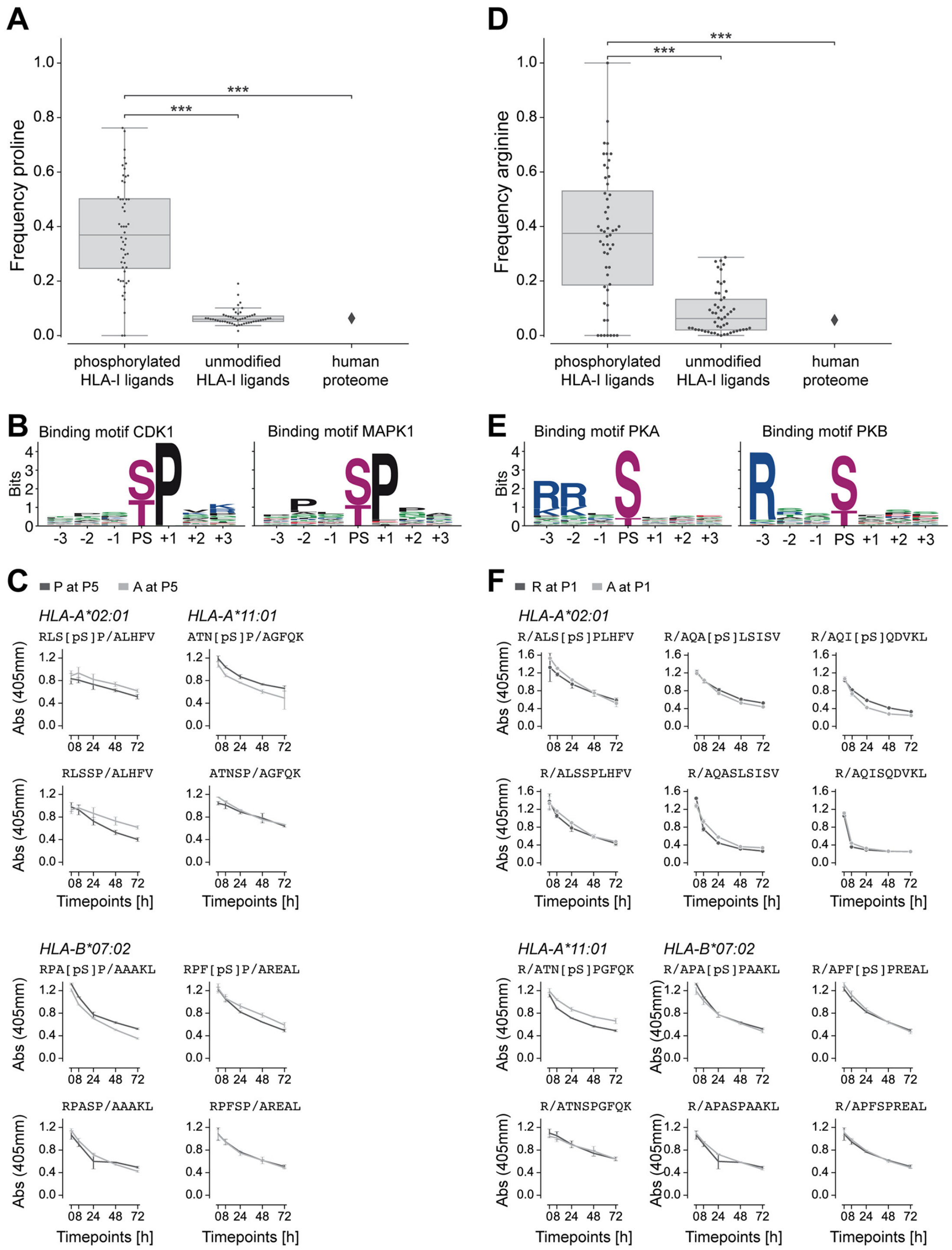


FIG. 3. **Phosphorylated positions in HLA-I ligands.** *A,* Distribution of the position of phosphorylated residues in phosphorylated HLA-I ligands of lengths 8 to 12. *B*, Half-lives of HLA-I ligands for peptides with positions 3 to 8 substituted by phosphorylated serine. Green squares mark the position of the phosphosite (phosphorylated serine) for peptides found in MS data. Lack of green square indicates one unmodified peptide observed in MS data (APSSSSSSL) or one synthetic peptide (RLSSSSSSV) used in this *in vitro* assay.

Proline Adjacent to the Phosphorylated Residue and Arginine at P1 are a Result of Kinase Motifs—Previous studies reported proline enrichment next to phosphorylated residues in phosphorylated HLA-I ligands as a consequence of the [pS/pT]P phosphorylation motif (20, 24–26, 28, 33). Our analysis with a much larger allelic coverage confirmed these re-

FIG. 2. **Analysis of phosphorylated peptides across HLA-I alleles.** *A*, Frequency of phosphorylated peptides per HLA-A, -B, and -C alleles for peptides of any length. Numbers in the plot indicate alleles tested in panel *B*. *B*, Ratio of half-lives between the phosphorylated (pS/pT) and the unmodified (S/T) peptides for several alleles. The colors of the bars correspond to alleles with high and low frequency of phosphorylated peptides in *A*. For HLA-A*01:01, HLA-B*07:02, HLA-C*06:02 and HLA-C*07:02 phosphorylated HLA-I binding motifs are shown, for HLA-A*25:01 and HLA-B*18:01 binding motifs of unmodified HLA-I ligands are given because too few phosphorylated peptides were observed in MS data for these alleles. *C*, Fraction of unmodified HLA-I 9-mer ligands containing a phosphosite at P4 for HLA-A, -B, and -C alleles. Arrows indicate the same alleles as in panel *A*. *D*, Length distribution of phosphorylated and unmodified ligands of HLA-A, HLA-B, and HLA-C alleles. *E*, Frequency of the different phosphorylated residues within phosphorylated HLA-I ligands of length 8 to 12 and within the human phosphoproteome (47). *F*, Dissociation assay (absorbance from ELISA) for unmodified and phosphorylated peptides (with phosphorylated serine, phosphorylated threonine, and phosphorylated tyrosine). (*, $p \leq 0.05$; **, $p \leq 0.01$.)


A
Frequency proline
1.0
0.8
0.6
0.4
0.2
0.0
***
***
phosphorylated HLA-I ligands
unmodified HLA-I ligands
human proteome
D
Frequency arginine
1.0
0.8
0.6
0.4
0.2
0.0
***
***
phosphorylated HLA-I ligands
unmodified HLA-I ligands
human proteome
B
Binding motif CDK1
Binding motif MAPK1
Bits
4
3
2
1
0
-3 -2 -1 PS +1 +2 +3
E
Binding motif PKA
Binding motif PKB
Bits
-3 -2 -1 PS +1 +2 +3
C
P at P5
A at P5
HLA-A*02:01
HLA-A*11:01
RLS[pS]P/ALHFV
ATN[pS]P/AGFQK
RLSSP/ALHFV
ATNSP/AGFQK
HLA-B*07:02
RPA[pS]P/AAAKL
RPF[pS]P/AREAL
RPASP/AAAKL
RPFSP/AREAL
Abs (405mm)
1.2
0.8
0.4
0.0
08 24 48 72
Timepoints [h]
F
R at P1
A at P1
HLA-A*02:01
R/ALS[pS]PLHFV
R/AQA[pS]LSISV
R/AQI[pS]QDVKL
R/ALSSPLHFV
R/AQASLSISV
R/AQISQDVKL
HLA-A*11:01
HLA-B*07:02
R/ATN[pS]PGFQK
R/APA[pS]PAAKL
R/APF[pS]PREAL
R/ATNSPGFQK
R/APASPAAKL
R/APFSPREAL
Abs (405mm)
1.6
1.2
0.8
0.4
0.0
08 24 48 72
Timepoints [h]

sults. In particular, we observed a significantly higher frequency of proline next to phosphorylated residues in HLA-I ligands compared with unmodified HLA-I ligands or to the human proteome (see Experimental Procedures and Fig. 4*A* and supplemental Fig. S5*A*). To support the hypothesis that the higher frequency of proline next to phosphorylated residues in HLA-I ligands reflects kinase phosphorylation motifs (Fig. 4*B*), we performed binding assays for three alleles (HLA-A*02:01, HLA-A*11:01, HLA-B*07:02) and four peptides with or without a phosphorylated residue at P4 and with or without proline at P5. The results of these binding assays show that proline or alanine at P5 did not change the binding stability, both for phosphorylated and unmodified peptides, consistent with the hypothesis that the proline enrichment is mainly because of kinase phosphorylation motifs (Fig. 4*C*).

Previous studies have shown that some phosphorylated HLA-I ligands show a preference for basic amino acids at P1 (20, 22, 24, 26, 28, 32, 33). For several alleles, we could see the same trend in our data when comparing phosphorylated and unmodified binding motifs (Fig. 1, *e.g.* HLA-B07:02 or HLA-C06:02). Comparisons between the frequency of arginine at P1 in phosphorylated HLA-I ligands with the frequency of arginine at P1 in unmodified HLA-I ligands and in the human proteome confirmed that the enrichment is statistically significant (see Experimental Procedures and Fig. 4*D* and supplemental Fig. S5*B*). We then asked whether this is because of enhanced binding of phosphorylated HLA-I ligands in the presence of R at P1 or to a signature of the known Rxx[pS/pT] phosphorylation motifs of protein kinases (see examples for PKA and PKB in Fig. 4*E*). To this end, we measured the binding stability of six different peptides with or without phosphorylated serine at P4 and with or without arginine at P1. For multiple alleles and peptides, the peptides with R at P1 have a similar binding compared with the peptides with A at P1, for both the phosphorylated and unmodified versions (Fig. 4*F*), suggesting that the preference for R at P1 does not affect the binding of phosphorylated ligands, but rather results from the phosphorylation motifs of specific kinases. Supplemental Table S1 shows phosphorylated HLA-I ligands that correspond to known phosphosites from the phosphoELM database (53) and were identified to be phosphorylated by kinases CDK1, PKA, and PKB.

Training Predictors on HLA-I Phospho-peptidomes Improves Predictions of Phosphorylated HLA-I Ligands—We then used our large curation of eluted phosphorylated peptides to train a predictor of phosphorylated HLA-I ligands. In particular, we explored different alternatives consisting of training the predictor (1) combining both the phosphorylated and non-phosphorylated peptides, (2) considering only unmodified peptides or (3) considering only phosphorylated peptides (see Experimental Procedures). For validation, we focused on the 9-mer peptides and performed a 5-fold cross validation on the 22 alleles with at least 20 phosphorylated peptides. Phosphorylated ligands per allele were divided into training and testing data set and negative peptides were added by randomly selecting peptides from the human phosphosite reference database (see Experimental Procedures). Fig. 5*A*–5*C* show the area under the receiver operating characteristics (ROC) curve (AUC), AUC0.1, and the precision for the top 20% of predicted phosphorylated peptides, respectively, of the cross validations for each version of training data for each allele separately (see supplemental Fig. S6*A*–S6*C* for average values over all 22 alleles). The results indicate that training the predictor with a combination of phosphorylated and unmodified HLA-I ligands performs best (1st bar, Fig. 5*A*–5*C* and supplemental Fig. S6*A*–S6*C*). When comparing the AUC values of the predictor trained with a combined data set (1st bar) and the predictor trained only with unmodified peptides (2nd bar), we can see that training only on unmodified HLA-I ligands and ignoring modification on a residue is also not as good as training on a combined data set for the prediction of HLA-I-phosphorylated peptide interactions. Finally, we observed that our predictor trained on a combined data set of phosphorylated and unmodified HLA-I ligands outperforms the state-of-the-art predictor NetMHCpan4.0 (4th bar in Fig. 5*A*–5*C* and supplemental Fig. S6*A*–S6*C*) (phosphorylated residues were substituted with "X" because NetMHCpan cannot take phosphorylated amino acids as input). This further confirms that including phosphorylated residues within the training data set of HLA-I ligand predictors improves their accuracy for predicting phosphorylated peptides.

To test the robustness of our predictor to the presence of some level of wrongly identified phosphorylated peptides, we included 5% phosphorylated decoy peptides in the training data. We did not observe significant changes in predictions, which demonstrates that our predictor can tolerate some level of contaminants or wrongly identified peptides (supplemental Fig. S6*D*). To exclude potential batch effects in our data, we trained the predictor on the newly curated data and tested it on previously reported HLA-I-restricted phosphorylated pep-

FIG. 4. **Proline and arginine enrichment in phosphorylated HLA-I ligands.** *A*, Frequency of proline next to phosphorylated residues in phosphorylated HLA-I ligands, proline at non-anchor positions in unmodified HLA-I ligands, and proline frequency in the human proteome. *B*, Kinase binding motifs for kinases CDK1 and MAPK1, three positions up- and downstream of the phosphosite (PS). *C*, Dissociation of peptides with proline or alanine next to phosphorylated serine (top) and next to unmodified serine in unmodified versions of the peptides (bottom). *D*, Frequency of arginine at P1 in phosphorylated HLA-I ligands, in unmodified HLA-I ligands, and in the human proteome. *E*, Kinase binding motifs for kinases PKA and PKB, three positions up- and downstream of the phosphosite (PS). *F*, Dissociation of peptides with arginine at P1 compared with peptides with alanine at P1 for both the phosphorylated (top) and unmodified (bottom) versions of the peptides. (***, $p \leq 0.001$)

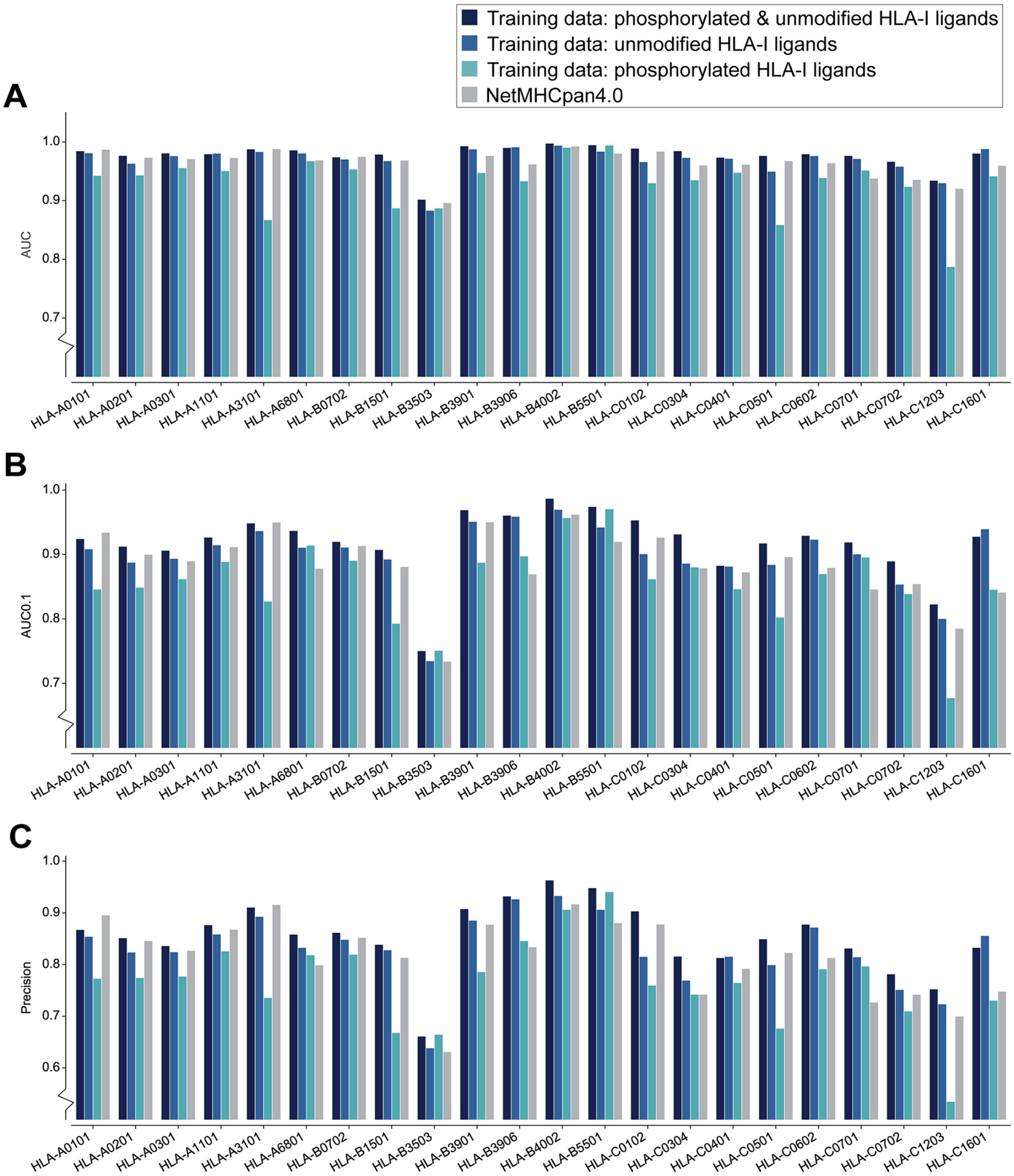


FIG. 5. **Cross validation of the predictor for each HLA-I allele with more than 20 phosphorylated 9-mer peptides.** *A*, AUC values for phosphorylated HLA-I 9-mer peptides when trained on both phosphorylated and unmodified ligands (1st bar), trained only on unmodified ligands (2nd bar, treating phosphorylated residues as their unmodified counterpart), or when trained only on phosphorylated HLA-I ligands (3rd bar). For comparison, AUC values are also shown when using NetMHCpan4.0 and replacing phosphorylated residues by "X" in the input (4th bar). *B*, Results of the 5-fold cross validation measured by AUC0.1. *C*, Precision measured for the top 20% of the predicted peptides (equivalent to recall of the prediction data).

tides. This was done for two alleles (HLA-A*02:01 and HLA-B*07:02) with enough previously reported HLA-I restricted phosphorylated peptides. Supplemental Fig. S6*E* shows that the results from these predictions are like the case where the training and testing data are selected from the whole pool of peptides (*i.e.* without distinguishing our newly curated data from previously reported data). A similar analysis was performed for ligands from phospho-enriched samples. Supplemental Fig. S6*F* shows that prediction of enriched samples trained on data from non-enriched samples performs equally well than randomly selected training and testing data of the same size. Finally, a saturation analysis with different number of phosphorylated ligands in the training data was performed for HLA-A*02:01 (supplemental Fig. S6*G*), showing that the choice of 20 phosphorylated peptides is already providing good prediction accuracy.

DISCUSSION

Aberrant phosphorylation is frequent in malignant cells. However, prediction of the presentation of phosphorylated peptides on HLA-I molecules has been poorly explored, mainly because of the lack of training data. Here, we curated phosphorylated HLA-I ligands across many immunopeptidomics studies to investigate molecular properties of interactions between phosphorylated peptides and HLA-I molecules and developed the first predictor for HLA-I interactions with phosphorylated peptides.

Our unsupervised approach to assign allelic restriction and infer binding motifs based on motif deconvolution (6, 37, 44) is especially appropriate for phosphorylated peptides because it does not require a priori information on their interactions with HLA-I alleles. In addition, it enabled us to use relatively permissive thresholds and subsequently filter potentially wrongly identified peptides that did not match the inferred motifs. As expected, the resulting phosphorylated motifs show similarity with those derived from unmodified peptides, especially at anchor positions (second and last positions for most HLA-I alleles).

Our results outlined a clear preference for phosphorylated peptides to bind to HLA-C alleles compared with HLA-A and -B alleles (Fig. 2*A*). Yet, differences in the binding site, such as R69 in the HLA heavy chain, do not appear to determine the preference for phosphorylated peptides of HLA-C alleles. Previous work (28) identified the interaction between R62 in HLA-B*40 and the phosphate moiety of the phosphorylated HLA-I ligand to support the binding of the phosphorylated peptide to this allele. All HLA-C but also most HLA-B alleles in our data set, including those with low fractions of phosphorylated ligands, contain arginine at position 62 (supplemental Fig. S4*G*). This suggests that, at least for the alleles studied in this work, arginine at position 62 does not necessarily favor the binding of phosphorylated HLA-I ligands. The analysis of unmodified ligands showed a higher fraction of human phosphosites in HLA-C ligands, suggesting that phosphosites fit the binding motifs of HLA-C alleles on average better than those of HLA-A and HLA-B alleles (Fig. 2*C*). This hypothesis is further supported by the correlation between the number of detected phosphosites per allele and the number of phosphorylated HLA-I ligands (supplemental Fig. S4*F*). This could explain, at least partly, the higher fraction of phosphorylated HLA-C ligands observed in immunopeptidomics data (Fig. 2*A*).

Our results demonstrated a clear preference for phosphorylated residues at P4 in phosphorylated HLA-I ligands, confirming previous observations (26, 28, 32). However, binding assays for HLA-B*07:02 pointed out that peptides with phosphorylation at P8 show a similar binding stability as peptides with phosphorylation at P4 (Fig. 3*B*). 9-mer binding motifs of alleles shown in Fig. 1 indicate that proline at P9 is not favorable for binding. This can explain why peptides with phosphorylation at P8 are less often observed in MS data, because the common kinase motif with proline next to the phosphorylated residue is not compatible with the binding motifs of HLA-I alleles. This suggests that the clear specificity of phosphorylation at P4 is because of (1) better binding of these phosphorylated peptides and (2) incompatibility of kinase motifs for phosphorylated peptides with phosphorylation at P8.

Consistent with what has been shown in previous studies (20, 22, 24, 26, 28, 29, 32, 33), we observed a clear preference for arginine at P1 for several alleles (Fig. 4*D* and supplemental Fig. S5*B*). Basic residues at P1 were observed to interact with the negatively charged phosphorylated residue at P4 and were suggested to improve the general stability of the phosphorylated peptide with the HLA-I molecule through this intramolecular bond (28, 29). However, when we tested the binding of phosphorylated and unmodified peptides with both arginine or alanine at P1, we observed very similar binding stability, suggesting that arginine does not specifically strengthen the binding of phosphorylated peptides in these alleles. Of course, we cannot exclude that the intramolecular bridge may be present and enhance binding of phosphorylated peptides to other alleles. Moreover, some alleles show preference for arginine at P1 for both phosphorylated and unmodified peptides, which could explain why arginine at P1 was reported to enhance binding of phosphorylated peptides (28, 29). Yet, the analysis of different kinase binding motifs indicates that many serine/threonine-kinases have a binding motif with arginine three positions upstream of the phosphorylated residue (Fig. 4*E*). This, together with the results of binding assays in Fig. 4*F*, provides a more likely explanation for the enrichment in arginine at P1 observed in Fig. 1 and Fig. 4*D*. Of note, arginine at P1 is not observed in all alleles (*e.g.* HLA-A*68:01 where P1 serves as an additional anchor residue, see Fig. 1).

Our results demonstrated that a combined training data set of phosphorylated and unmodified peptides outperforms training that is based solely on unmodified peptides. Hence, we can conclude that without phosphorylated peptides in the training data the predictor lacks important information about the phosphorylated residue (especially the preference for the phosphorylated residue at P4).

Overall, our analysis of phosphorylated immunopeptidomes shows that the presentation of phosphorylated peptides on HLA-I molecules is governed by a combination of HLA-I binding motifs (specificity mainly at P2 and PΩ), intrinsic HLA-I binding properties of phosphorylated peptides (specificity at P4) and kinase motifs (specificity at P1 and $P_{phospho}+1$). Our ability to integrate these different features into a robust predictor of phosphorylated HLA-I ligands explains the improvement over existing tools and provides a rationale for training on both unmodified and phosphorylated HLA-I ligands. In this work, we have leveraged existing immunopeptidomics data for the identification of a large collection of phosphorylated HLA-I ligands, mostly originating from un-enriched samples. To develop the predictor for as many alleles as possible, we applied relatively permissive filters and thresholds for interpretation the MS immunopeptidomic data set because motif deconvolution can filter many potentially wrongly identified peptides (6) (see also a related strategy for general immunopeptidomics experiments (55)) and our computational approach can handle some level of false positives. This contrasts with other applications of MS based immunopeptidomics studies aimed at directly identifying novel epitopes, where high-confidence peptide identification is crucial. We anticipate that this predictor will facilitate the identification of phosphorylated T cell epitopes for researchers that do not have access to high-quality but expensive MS immunopeptidomics technology and will foster future research on their role in immune recognition of infected or malignant cells.

Acknowledgments—We thank Stefan Stevanovic and Moreno Di Marco for sharing with us the raw MS data from (39). We are thankful to Julien Schmidt and the Protein and Peptide Chemistry Facility from UNIL for synthesizing the peptides.

DATA AVAILABILITY

The mass spectrometry immunopeptidomics raw files, the MaxQuant version used in this study as well as all MaxQuant output files and tables have been deposited to the ProteomeXchange Consortium via the PRIDE (56) partner repository with the data set identifier PXD013831 (https://www.ebi.ac.uk/pride/archive/projects/PXD013831).

* The authors declare that they have no conflicts of interest with the contents of this article.

S This article contains supplemental Figures, Data, and Table.

‖ To whom correspondence may be addressed. E-mail: David.Gfeller@unil.ch.

‡‡ To whom correspondence may be addressed. E-mail: Michal.Bassani@chuv.ch.

Author contributions: M.S., G.C., M.B.-S., and D.G. designed research; M.S., P.G., J.R., J.M., H.-S.P., M.B.-S., and D.G. performed research; M.S., P.G., and D.G. contributed new reagents/analytic tools; M.S., J.R., M.M., M.B.-S., and D.G. analyzed data; M.S., M.B.-S., and D.G. wrote the paper.

REFERENCES

1. Neefjes, J., Jongsma, M. L. M., Paul, P., and Bakke, O. (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11,** 823–836
2. Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2015) The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* **43,** D423–D431
3. Gfeller, D., and Bassani-sternberg, M. (2018) Predicting antigen presentation — what could we learn from a million peptides? *Front. Immunol.* **9,** 1–17
4. Andreatta, M., and Nielsen, M. (2015) Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. *Bioinformatics* **32,** 511–517
5. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199,** 3360–3368
6. Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R. T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018) The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol.* **202,** 1–12
7. Donnell, T. J. O., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018) MHCflurry: Open-source class I MHC binding affinity prediction. *Cell Syst.* **7,** 129–132
8. Cho, Y., Kang, H. G., Kim, S., Lee, S., Jee, S., Ahn, S. G., Kang, M. J., Song, J. S., Chung, J., Yi, E. C., and Chun, K.-H. (2018) Post-translational modification of OCT4 in breast cancer tumorigenesis. *Cell Death Differ.* **25,** 1781–1795
9. Jarrold, J., and Davies, C. C. (2019) PRMTs and arginine methylation: cancer's best-kept secret? *Trends Mol. Med.* **25,** 1–16
10. Krueger, K. E., and Srivastava, S. (2006) Posttranslational protein modifications. *Mol. Cell. Proteomics* **5,** 1799–1810
11. Archer, T. C., Ehrenberger, T., Mundt, F., Mesirov, J. P., Pomeroy, S. L., and Fraenkel, E. (2018) Proteomics, post-translational modifications, and integrative analyses reveal molecular heterogeneity within medulloblastoma subgroups. *Cancer Cell* **34,** 396–410
12. Hanahan, D., and Weinberg, R. A. (2000) The hallmarks of cancer. *Cell* **100,** 57–70
13. Humphrey, S. J., James, D. E., and Mann, M. (2015) Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **26,** 676–687
14. Hunter, T. (1995) Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80,** 225–236
15. Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017) The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *Int. J. Mol. Med.* **40,** 271–280
16. Blume-Jensen, P., and Hunter, T. (2001) Oncogenic kinase signalling. *Nature* **2,** 355–366
17. Reimand, J., Wagih, O., and Bader, G. D. (2013) The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* **3,** 1–9
18. Krug, K., Mertins, P., Zhang, B., Hornbeck, P., Raju, R., Ahmad, R., Szucs, M., Mundt, F., Forestier, D., Jane-Valbuena, J., Keshishian, H., Gillette, M. A., Tamayo, P., Mesirov, J. P., Jaffe, J. D., Carr, S. A., and Mani, D. R. (2019) A curated resource for phosphosite-specific signature analysis. *Mol. Cell. Proteomics* **18,** 576–593
19. Andersen, M. H., Bonfill, J. E., Neisig, A., Arsequell, G., Sondergaard, I., Valencia, G., Neefjes, J., Zeuthen, J., Elliott, T., and Haurum, J. S. (1999) Phosphorylated peptides can be transported by TAP molecules, presented by class I MHC molecules, and recognized by phosphopeptide-specific CTL. *J. Immunol.* **163,** 3812–3818
20. Zarling, A. L., Ficarro, S. B., White, F. M., Shabanowitz, J., Hunt, D. F., and Engelhard, V. H. (2000) Phosphorylated peptides are naturally processed and presented by major histocompatibility complex class I molecules in vivo. *J. Exp. Med.* **192,** 1755–1762
21. Zarling, A. L., Polefrone, J. M., Evans, A. M., Mikesh, L. M., Shabanowitz, J., Lewis, S. T., Engelhard, V. H., and Hunt, D. F. (2006) Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *Proc. Natl. Acad. Sci.* **103,** 14889–14894
22. Meyer, V. S., Drews, O., Günder, M., Hennenlotter, J., Rammensee, H. G., and Stevanovic, S. (2009) Identification of natural MHC class II presented phosphopeptides and tumor-derived MHC class I phospholigands. *J. Proteome Res.* **8,** 3666–3674
23. Petersen, J., Wurzbacher, S. J., Williamson, N. A., Ramarathinam, S. H., Reid, H. H., Nair, A. K. N., Zhao, A. Y., Nastovska, R., Rudge, G., Rossjohn, J., and Purcell, A. W. (2009) Phosphorylated self-peptides

alter human leukocyte antigen class I-restricted antigen presentation and generate tumor-specific epitopes. *Proc. Natl. Acad. Sci.* **106,** 2776–2781

24. Cobbold, M., De La Pena, H., Norris, A., Polefrone, J. M., Qian, J., English, A. M., Cummings, K. L., Penny, S., Turner, J. E., Cottine, J., Abelin, J. G., Malaker, S. A., Zarling, A. L., Huang, H.-W., Goodyear, O., Freeman, S. D., Shabanowitz, J., Pratt, G., Craddock, C., Williams, M. E., Hunt, D. F., and Engelhard, V. H. (2013) MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia. *Sci. Transl. Med.* **5,** 1–10
25. Marcilla, M., Alpízar, A., Lombardía, M., Ramos-Fernandez, A., Ramos, M., and Albar, J. P. (2014) Increased diversity of the HLA-B40 ligandome by the presentation of peptides phosphorylated at their main anchor residue. *Mol. Cell. Proteomics* **13,** 462–474
26. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R. H, Busch, D., Peschel, C., Rad, R., Cox, J., Mann, M., and Krackhardt, A. M. (2016) Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7,** 1–16
27. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46,** 315–326
28. Alpízar, A., Marino, F., Ramos-Fernández, A., Lombardía, M., Jeko, A., Pazos, F., Paradela, A., Santiago, C., Heck, A. J. R., and Marcilla, M. (2017) A molecular basis for the presentation of phosphorylated peptides by HLA-B antigens. *Mol. Cell. Proteomics* **16,** 181–193
29. Mohammed, F., Stones, D. H., Zarling, A. L., Willcox, C. R., Shabanowitz, J., Cummings, K. L., Hunt, D. F., Cobbold, M., Engelhard, V. H., and Willcox, B. E. (2017) The antigenic identity of human class I MHC phosphopeptides is critically dependent upon phosphorylation status. *Oncotarget* **8,** 54160–54172
30. Olsson, N., Schultz, L. M., Zhang, L., and Khodadoust, M. S. (2018) T-cell immunopeptidomes reveal cell subtype surface markers derived from intracellular proteins. *Proteomics* **18,** e1700410
31. Lin, M., Shen, K., Liu, B., Chen, I., Sher, Y., Tseng, G., Liu, S., and Sung, W. (2019) Immunological evaluation of a novel HLA-A2 restricted phosphopeptide of tumor associated antigen, TRAP1, on cancer therapy. *Vaccine X*, **100017**
32. Mohammed, F., Cobbold, M., Zarling, A. L., Salim, M., Barrett-Wilt, G. A., Shabanowitz, J., Hunt, D. F., Engelhard, V. H., and Willcox, B. E. (2008) Phosphorylation-dependent interaction between antigenic peptides and MHC class I: a molecular basis for the presentation of transformed self. *Nat. Immunol.* **9,** 1236–1243
33. Mommen, G. P. M., Frese, C. K., Meiring, H. D., van Gaans-van den Brink, J., de Jong, A. P. J. M., van Els, C. A. C. M., and Heck, A. J. R. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci.* **111,** 4507–4512
34. Chong, C., Marino, F., Pak, H.-S., Racle, J., Daniel, R. T., Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017) High-throughput and sensitive immunopeptidomics platform reveals profound IFNγ-mediated remodeling of the HLA ligandome. *Mol. Cell. Proteomics* **17,** 533–548
35. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
36. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011) Confident phosphorylation site localization using the mascot delta score. *Mol. Cell. Proteomics* **10,** 1–12
37. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G., and Gfeller, D. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* **13,** e1005725
38. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015) Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14,** 658–673
39. Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H.-G., and Stevanović, S. (2017) Unveiling the peptide motifs of HLA-C and HLA-G from naturally presented peptides and generation of binding prediction matrices. *J. Immunol.* **199,** 2639–2651
40. Ostrov, D. A., Grant, B. J., Pompeu, Y. A., Sidney, J., Harndahl, M., Southwood, S., Oseroff, C., Lu, S., Jakoncic, J., de Oliveira, C. A. F., Yang, L., Mei, H., Shi, L., Shabanowitz, J., English, A. M., Wriston, A., Lucas, A., Phillips, E., Mallal, S., Grey, H. M., Sette, A., Hunt, D. F., Buus, S., and Peters, B. (2012) Drug hypersensitivity caused by alteration of the MHC-presented self-peptide repertoire. *Proc. Natl. Acad. Sci.* **109,** 9959–9964
41. Schittenhelm, R. B., Sian T. C. C. L. K, Wilmann, P. G., Dudek, N. L., and Purcell, A. W. (2015) Revisiting the arthritogenic peptide theory: quantitative not qualitative changes in the peptide repertoire of HLA-B27 allotypes. *Arthritis Rheumatol.* **67,** 702–713
42. Giam, K., Ayala-Perez, R., Illing, P. T., Schittenhelm, R. B., Croft, N. P., Purcell, A. W., and Dudek, N. L. (2015) A comprehensive analysis of peptides presented by HLA-A1. *Tissue Antigens* **85,** 492–496
43. Ramarathinam, S. H., Gras, S., Alcantara, S., Yeung, A. W. S., Mifsud, N. A., Sonza, S., Illing, P. T., Glaros, E. N., Center, R. J., Thomas, S. R., Kent, S. J., Ternette, N., Purcell, D. F. J., Rossjohn, J., and Purcell, A. W. (2018) Identification of Native and Posttranslationally Modified HLA-B*57:01-Restricted HIV Envelope Derived Epitopes Using Immunoproteomics. *Proteomics* **18,** 1–11
44. Bassani-Sternberg, M., and Gfeller, D. (2016) Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide–HLA interactions. *J. Immunol.* **197,** 2492–2499
45. Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A., and Peters, B. (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43,** D405–D412
46. Wagih, O. (2017) Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **33,** 3645–3647
47. Sharma, K., D'Souza, R. C. J., Tyanova, S., Schaab, C., Wisniewski, J. R., Cox, J., and Mann, M. (2014) Resource ultradeep human phosphoproteome reveals a distinct regulatory nature. *Cell Rep.* **8,** 1583–1594
48. Diella, F., Gould, C. M., Chica, C., Via, A., and Gibson, T. J. (2008) Phospho.ELM: A database of phosphorylation sites - Update 2008. *Nucleic Acids Res.* **36,** 240–244
49. Nielsen, M., Lundegaard, C., Worning, P., Sylvester Hvid, C., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20,** 1388–1397
50. Henikoff, S., and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89,** 10915–10919
51. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12,** 2825–2830
52. Ullah, S., Lin, S., Xu, Y., Deng, W., Ma, L., Zhang, Y., Liu, Z., and Xue, Y. (2016) dbPAF: An integrative database of protein phosphorylation in animals and fungi. *Sci. Rep.* **6,** 1–9
53. Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. (2011) Phospho.ELM: a database of phosphorylation sites — update 2011. *Nucleic Acids Res.* **39,** 261–267
54. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** 512–520
55. Andreatta, M., Nicastri, A., Peng, X., Hancock, G., Dorrell, L., Ternette, N., and Nielsen, M. (2019) MS-rescue: a computational pipeline to increase the quality and yield of immunopeptidomics experiments. *Proteomics* **19,** e1800357
56. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A. F., Ternent, T., Brazma, A., and Vizcaíno, J. A. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47,** 442–450

Supplemental Information

Supplementary Figures

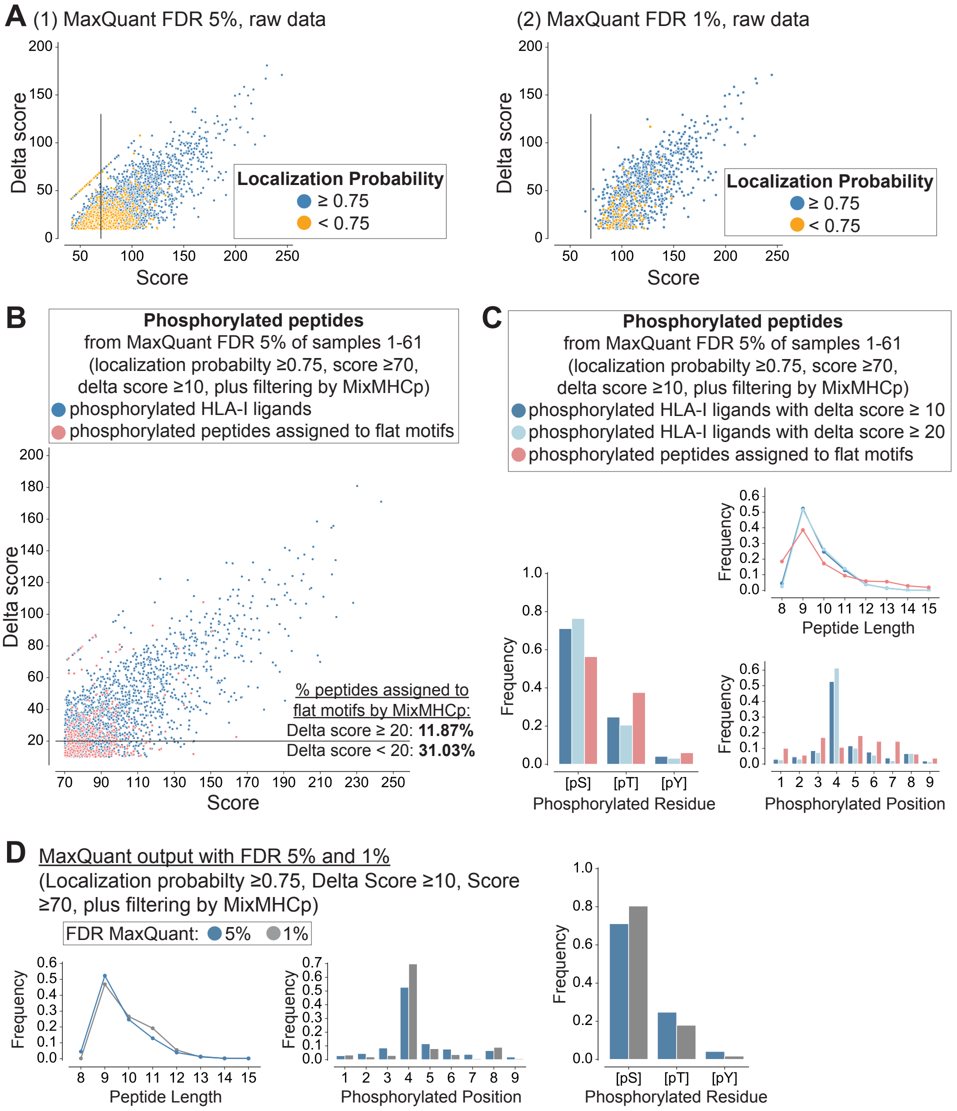


Figure S1: (A) Distribution of Andromeda search engine scores vs. score differences to the second best peptide spectrum match (delta scores) for MaxQuant search with FDR of 5% (left) and with FDR of 1% (right). Blue dots mark phosphorylated peptides with localization probability of ≥0.75, orange dots mark phosphorylated peptides with localization probability <0.75. **(B)** Distribution of Andromeda search engine scores vs. delta scores for all peptides assigned to HLA-I alleles (blue dots) and the flat motif (red dots) in the motif deconvolution. **(C)** Distribution of phosphorylated residues ([pS/pT/pY]), peptide lengths,

and phosphorylated positions in 9-mers for phosphorylated peptides assigned to HLA-I alleles with delta scores ≥10 (blue) and delta scores ≥20 (cyan), and for peptides assigned to the flat motif by MixMHCp (red). **(D)** Distribution of phosphorylated residues ([pS/pT/pY]), peptide lengths, and phosphorylated positions in 9-mers for phosphorylated peptides assigned to HLA-I alleles with FDR of 5% (blue) and FDR of 1% (grey).

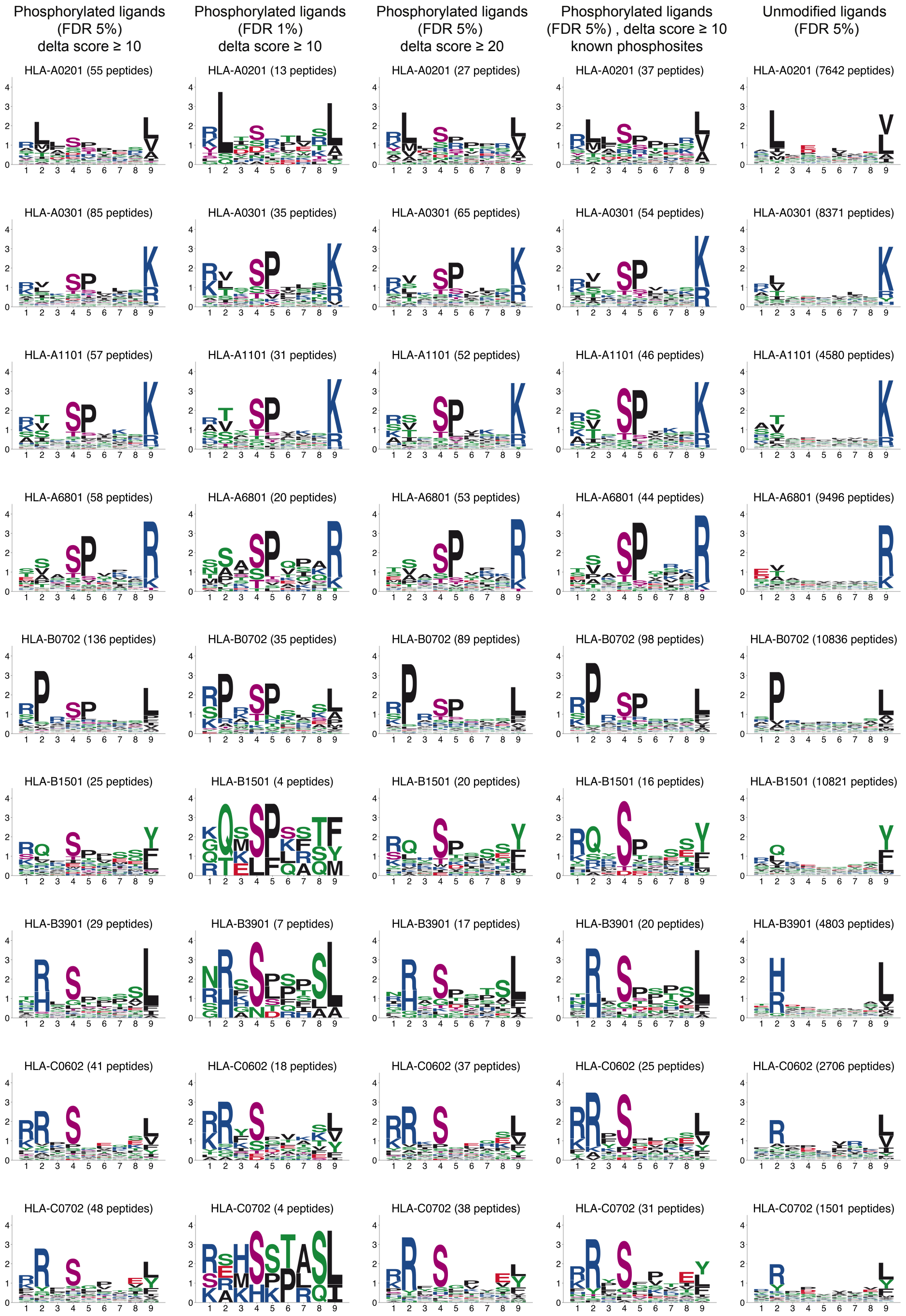


Figure S2: Comparison of binding motifs for different choices of parameters used to search the MS data. Column 1 shows the motifs based on phosphorylated peptides with FDR of 5% (same data as in Figure 1). Column 2 shows the motifs based on phosphorylated peptides with FDR of 1%. Column 3 shows the motifs based on phosphorylated peptides with delta

score ≥20. Column 4 shows the motifs based only on phosphorylated peptides containing known phosphosites. Column 5 shows the motifs based on unmodified peptides.

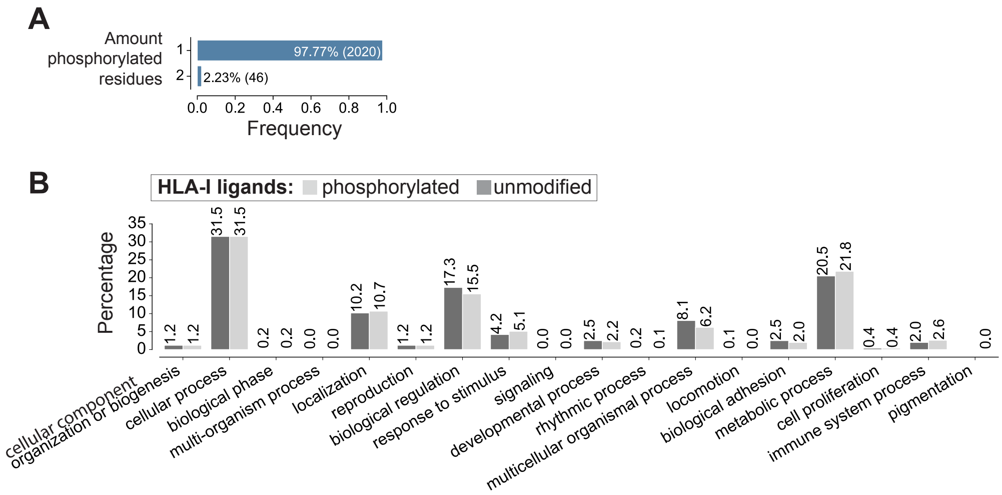


Figure S3: (A) Amount of phosphorylated residues per HLA-I ligand. **(B)** GO enrichment analysis with biological process classification for the source proteins of all phosphorylated and unmodified HLA-I ligands (performed with Panther tool [1]).

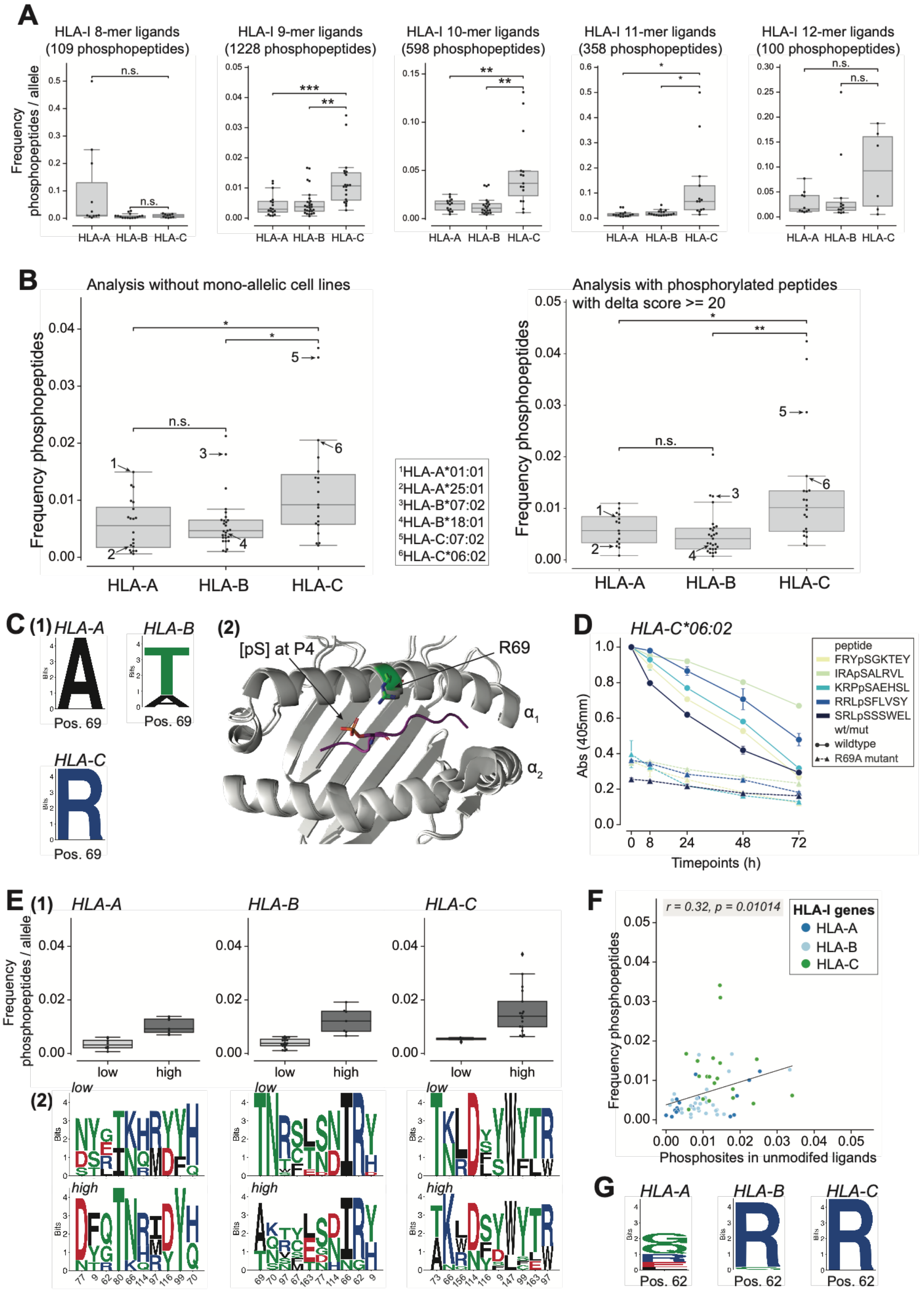


Figure S4: Analysis of phosphorylated peptides per HLA-I allele. **(A)** Frequency of detected phosphorylated peptides per allele for different lengths. **(B)** Frequency of phosphorylated peptides per HLA-A, -B, and -C alleles for peptides of any length without monoallelic HLA-C samples (left panel) and for HLA-I ligands with delta score ≥20 (right

panel). Arrows indicate same alleles as in Fig. 2A and C and correspond to alleles tested in Fig. 2E. (**C)** Sequence logos of position 69 in HLA-A, -B, and -C alleles (1). Crystal structure of HLA-C*06:02 (PDB code: 5w67) superimposed to HLA-A*02:01 (PDB code: 4nnx) with phosphorylated ligand RQA[pS]LSISV. R69 of HLA-C*06:02 is shown in green (2). **(D)** Dissociation assays of phosphorylated peptides with HLA-C*06:02 wt (R at P69) and mutated (R69A) alleles. **(E)** Frequency of phosphorylated peptides per allele grouped into high and low alleles. Sequence logos of the ten most different positions of the HLA-I binding site (measured by Euclidean distance between the groups with high and low alleles) for high and low alleles of HLA-A, HLA-B and HLA-C separately. **(F)** Correlation of phosphosites from the human proteome found in unmodified ligands and the frequency of phosphorylated HLA-I ligands per allele. **(G)** Sequence logos of position 62 in HLA-A, HLA-B, and HLA-C alleles. (*: p <= 0.05; **: p<= 0.01; ***: p <= 0.001)

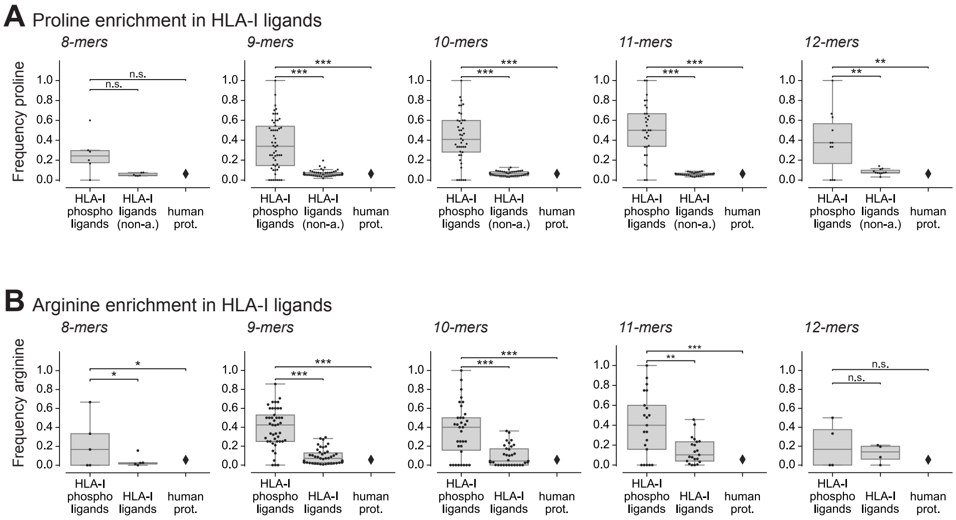


Figure S5: Analysis of arginine and proline enrichment for 8- to 12-mers. **(A)** Frequency of proline next to phosphorylated serine in phosphorylated HLA-I peptides (1), proline frequency at non-anchor positions in unmodified HLA-I ligands (2), and proline frequency in the human proteome (3). **(B)** Frequency of arginine at P1 in phosphorylated HLA-I ligands (1), at P1 in unmodified HLA-I ligands (2) and overall arginine frequency in the human proteome (3) for different lengths. (*: p <= 0.05; **: p <= 0.01; ***: p <= 0.001)

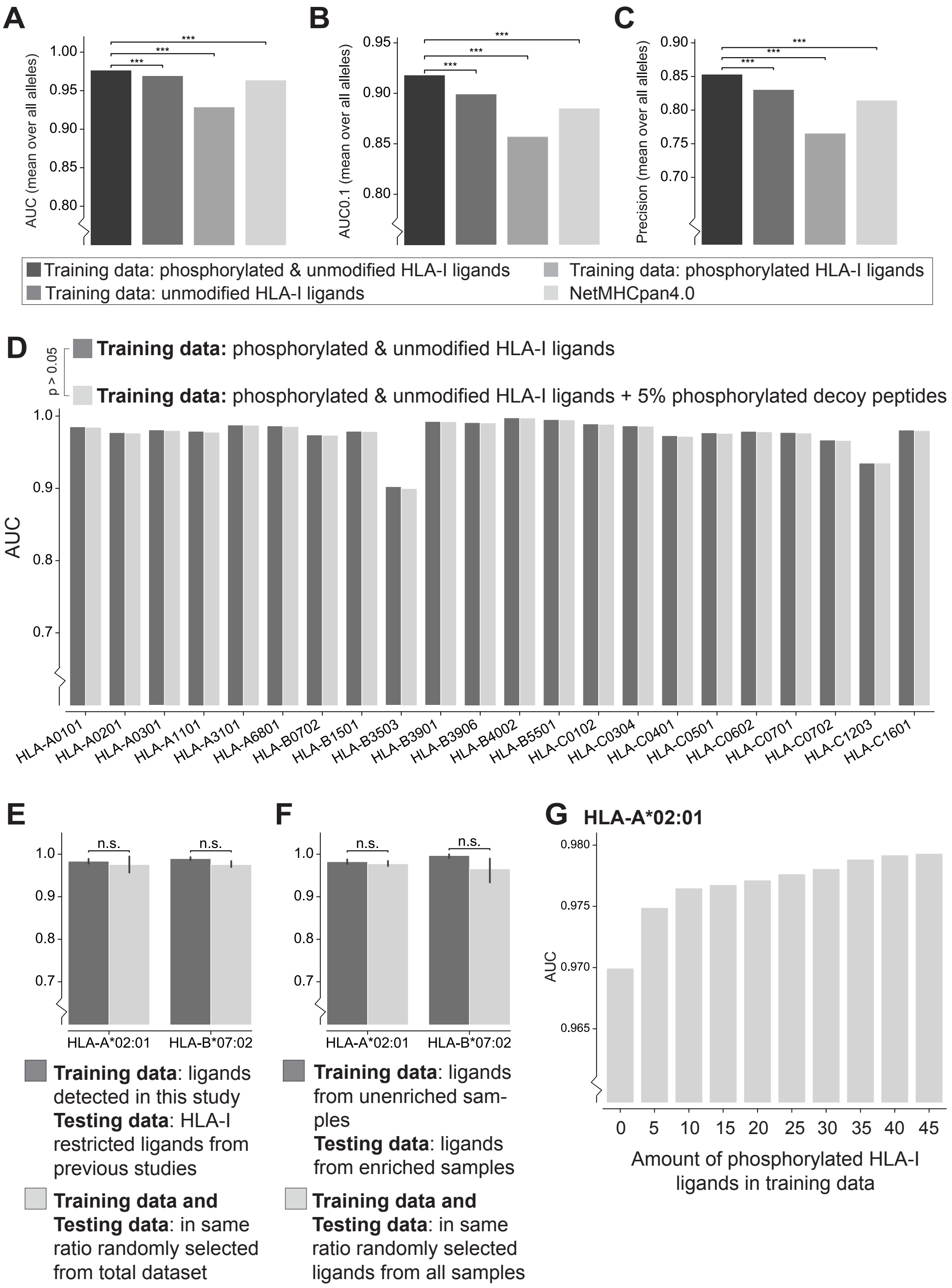


Figure S6: 5-fold cross validation of the predictor trained with different training datasets as mean over all alleles shown as **(A)** AUC values, **(B)** AUC0.1, **(C)** Precision of the top 20% of predicted peptides. Training of the predictor was performed with both, phosphorylated and unmodified ligands (1st bar), with only unmodified ligands (2nd bar), and with phosphorylated HLA-I ligands (3rd bar). 4th bar shows AUC values using NetMHCpan4.0

(***: p <= 0.001). **(D)** The predictor was trained on phosphorylated and unmodified HLA-I ligands (left bars); in a second run the predictor was trained on phosphorylated and unmodified HLA-I ligands with additional 5% phosphorylated decoy peptides (right bars). Results are shown as AUC values. **(E)** Predicting previously published phosphorylated HLA-I ligands for HLA-A*02:01 and HLA-B*07:02. The predictor is trained on phosphorylated HLA-I ligands detect by MS in this study. Testing data are HLA-I restricted ligands from previous studies (left bars). As comparison, predictions with random division of training and testing dataset from the combined set of peptides are performed (right bars). **(F)** Prediction of HLA-A*02:01 and HLA-B*07:02. The predictor is trained on phosphorylated HLA-I ligands only found in non-enriched samples and tested on ligands found in enriched samples (left bars). As comparison, predictions with random division of training and testing dataset from the combined set of peptides are performed (right bars). **(G)** Prediction of 10 randomly selected HLA-A*02:01 ligands performed with training data of increasing size.

Supplementary Tables

Kinase	Phosphorylated Peptide	Allele
CDK1	VLL**[pS]**PVPEL	HLA-A*02:01
CDK1	LQL**[pS]**PLKGLSL	HLA-A*02:06, HLA-B*55:01
CDK1	ITT**[pS]**PITVRK	HLA-A*11:01
CDK1	EVP**[pT]**PKRPR	HLA-A*68:01
CDK1	YAS**[pS]**PGGVYATR	HLA-A*68:01
CDK1	RPI**[pT]**PPRNSA	HLA-B*07:02, HLA-B*55:01
CDK1	SPK**[pS]**PTAAL	HLA-B*07:02, HLA-B*55:01, HLA-C*03:32
CDK1	SPRTPV**[pS]**PVKF	HLA-B*07:02
CDK1	SPR**[pT]**PVSPVKF	HLA-B*07:02
PKA	EPKRR**[pS]**ARL	HLA-B*07:02
PKA	RPRSL**[pS]**SPTV	HLA-B*07:02
PKA	RPRSL**[pS]**SPTVTL	HLA-B*07:02
PKA	RRK**[pS]**HEAEV	HLA-C*06:02
PKB	RAH**[pS]**SPASL	HLA-B*07:02, HLA-B*35:03, HLA-C*01:02, HLA-C*03:03, HLA-C*03:04, HLA-C*03:32, HLA-C*12:03
PKB	RHK**[pS]**DSISL	HLA-B*39:01

Table S1: All phosphorylated HLA-I ligands from this study that are known to be phosphorylated by CDK1 or PKA/B from the phosphosite database phosphoELM [3].

References

1. Mi H, Muruganujan A, Ebert D, Huang X, Thomas D. PANTHER version 14 : more genomes , a new PANTHER GO-slim and improvements in enrichment analysis tools. Oxford University Press; 2019;47: 419–426. doi:10.1093/nar/gky1038

2. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot : exploring protein sequences for globularity and disorder. Nucleic Acids Res. 2003;31: 3701–3708. doi:10.1093/nar/gkg519

3. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, et al. Phospho . ELM : a database of phosphorylation sites — update 2011. Nucleic Acids Res. 2011;39: 261–267. doi:10.1093/nar/gkq1104