Year : 2020

An Integrated Immunopeptidomics and Proteogenomics Framework to Discover Non-Canonical Targets for Cancer Immunotherapy

Chong Chloe

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'Oncologie Fondamentale

An Integrated Immunopeptidomics and Proteogenomics Framework to Discover Non-Canonical Targets for Cancer Immunotherapy

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Chloe CHONG

Master en Biologie (spécialité Biochimie)
Eidgenössische Technische Hochschule (ETH) Zürich, Suisse

Jury

Prof. Fabio Martinon	Président
Prof. George Coukos	Directeur de thèse
Dr. Michal Bassani-Sternberg	Co-directrice de thèse
Prof. Yardena Samuels	Experte
Dr. Marco Gerlinger	Expert
Dr. Peter van Veelen	Expert

Lausanne, 2020

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'Oncologie Fondamentale

An Integrated Immunopeptidomics and Proteogenomics Framework to Discover Non-Canonical Targets for Cancer Immunotherapy

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Chloe CHONG

Master en Biologie (spécialité Biochimie)
Eidgenössische Technische Hochschule (ETH) Zürich, Suisse

Jury

Prof. Fabio Martinon	Président
Prof. George Coukos	Directeur de thèse
Dr. Michal Bassani-Sternberg	Co-directrice de thèse
Prof. Yardena Samuels	Experte
Dr. Marco Gerlinger	Expert
Dr. Peter van Veelen	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.	Fabio	**Martinon**
Directeur·trice de thèse	Monsieur	Prof.	George	**Coukos**
Co-directeur·trice	Madame	Dre	Michal	**Bassani-Sternberg**
Expert·e·s	Madame	Prof.	Yardena	**Samuels**
	Monsieur	Dr	Marco	**Gerlinger**
	Monsieur	Dr	Peter	**van Veelen**

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

Madame Chloe Chong

Master of science ETH in Biology, ETHZ, Zürich

intitulée

An integrated immunopeptidomics and proteogenomics framework to discover non-canonical targets for cancer immunotherapy

Lausanne, le 27 mars 2020

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Niko GELDNER
Directeur de l'Ecole Doctorale

In memory of my grandparents,

who always believed in me.

TABLE OF CONTENTS

RESUME

Un élément essentiel de l'immunothérapie appliquée au cancer est l'identification de peptides liant les antigènes des leucocytes humains (HLA) et capables d'induire une puissante réponse T anti-tumorale. La spectrométrie de masse (MS) constitue actuellement la seule méthode non-biaisée permettant une analyse détaillée du panel d'antigènes susceptibles d'être présentés aux lymphocytes T *in vivo*. L'utilisation de cette méthode en clinique requiert toutefois des améliorations significatives de la méthodologie utilisée lors de l'identification des peptides HLA.

Un consortium multidisciplinaire de chercheurs a récemment mis en lumière les problèmes actuellement liés à l'utilisation de la MS en immunopeptidomique, soulignant le besoin de développer de nouvelles méthodes et mettant en évidence le défi que représente la standardisation de l'immuno-purification des molécules HLA. La première partie de cette thèse vise à optimiser les méthodes expérimentales permettant l'extraction des peptides apprêtés aux HLA. L'optimisation de la méthodologie de base a permis des améliorations notables en terme de débit, de reproductibilité, de sensibilité et a permis une purification séquentielle des molécules de HLA de classe I de classe II ainsi que de leurs peptides, à partir de lignées cellulaires ou de tissus. En comparaison avec les méthodes existantes, ce protocole comprend moins d'étapes et permet de limiter la manipulation des échantillons ainsi que le temps de purification. Cette méthode, pour les peptides HLA extraits, a permis d'obtenir des taux de reproductibilité et de sensibilité sans précédents (corrélations de Pearson jusqu'à 0,98 et 0,97 pour les HLA de classe I et de classe II, respectivement). De plus, la faisabilité d'études comparatives robustes a été démontrée à partir d'une lignée cellulaire de cancer de l'ovaire, traitée à l'interféron gamma. En effet, cette nouvelle méthode a mis en évidence des changements quantitatifs et qualitatifs du catalogue de peptides présentés aux HLA. Les résultats obtenus ont mis en avant une augmentation de la présentation de longs ligands chymotryptiques de classe I. Ce phénomène est probablement lié à la modulation de la machinerie de traitement et de présentation des antigènes. Dans cette première partie de thèse, nous avons développé une méthodologie robuste et rationalisée, facilitant la purification des HLA et pouvant être appliquée en recherche fondamentale et translationnelle.

Bien que les néoantigènes représentent une cible attractive, des études récentes ont mis en évidence l'existence des antigènes non canoniques. Ces antigènes tumoraux, bien que non mutés, sont aussi spécifiques aux cellules cancéreuses et semblent jouer un rôle important dans l'immunité anti-tumorale. La seconde partie de cette thèse a pour objectif le développement d'une méthodologie d'analyse permettant l'identification ainsi que la validation de ces antigènes particuliers. Les antigènes non canoniques sont d'origine présumée non codante et ne sont, par conséquent, que rarement inclus dans les bases de données des séquences de protéines de référence. De ce fait, ils ne sont généralement pas pris en compte lors des recherches de MS utilisant de telles bases de données. Afin de palier ce problème et de permettre leur identification par MS, le séquençage de l'exome entier, le séquençage de l'ARN sur une population de cellules et sur des cellules uniques, ainsi que le profilage des ribosomes ont été intégrés aux données d'immunopeptidomique. Ainsi, NewAnce, un programme informatique permettant de combiner les données de deux outils de recherche MS en tandem, a été développé afin de calculer le taux d'antigènes non canoniques identifiés comme faux positifs. L'utilisation de NewAnce sur des lignées cellulaires provenant de patients atteints de mélanomes ainsi que sur des biopsies de cancer du poumon a permis l'identification précise de centaines de peptides HLA non classiques, spécifiques aux cellules tumorales et communs à plusieurs patients. Le niveau de confirmation des peptides non canoniques a ensuite été testé à l'aide d'une approche de MS ciblée. Les peptides résultant de ces analyses ont été minutieusement validés pour un des échantillons de mélanome disponibles. De plus, le profilage des ribosomes a révélé que les nouveaux cadres de lecture ouverts, desquels résultent certains de ces peptides non classiques, sont activement traduits. L'évaluation de l'immunogenicité de ces peptides a été évaluée avec des cellules immunitaires autologues et a révélé un épitope immunogène non canonique, provenant d'un cadre de lecture ouvert alternatif du gène ABCB5, un marqueur des cellules souches du mélanome.

De manière globale, les résultats obtenus au cours de cette thèse soulignent la possibilité d'inclure ce type d'analyse de proteogénomique dans un protocole d'identification de néoantigènes existant. Cela permettrait d'inclure et prioriser des antigènes tumoraux non classiques et de proposer aux patients en impasse thérapeutique des immunothérapies anti-tumorales personnalisées.

SUMMARY

A central factor to the development of cancer immunotherapy is the identification of clinically relevant human leukocyte antigen (HLA)-bound peptides that elicit potent anti-tumor T cell responses. Mass spectrometry (MS) is the only unbiased technique that captures the *in vivo* presented HLA repertoire. However, significant improvements in MS-based HLA peptide discovery methodologies are necessary to enable the smooth transition to the clinic.

Recently, a consortium of multidisciplinary researchers presented current issues in clinical MS-based immunopeptidomics, highlighting method development and standardization challenges in HLA immunoaffinity purification. The first part of this thesis addresses improvements to the experimental method for HLA peptide extraction. The approach was optimized with several new developments, facilitating high-throughput, reproducible, scalable, and sensitive sequential immunoaffinity purification of HLA class I and class II peptides from cell lines and tissue samples. The method showed increased speed, and reduced sample handling when compared to previous methods. Unprecedented depth and high reproducibility were achieved for the obtained HLA peptides (Pearson correlations up to 0.98 and 0.97 for HLA class I and HLA class II, respectively). Additionally, the feasibility of performing robust comparative studies was demonstrated on an ovarian cancer cell line treated with interferon gamma. Both quantitative and qualitative changes were detected in the cancer HLA repertoire upon treatment. Specifically, a yet unreported and interesting phenomenon was the upregulated presentation of longer and chymotryptic-like HLA class I ligands, likely related to the modulation of the antigen processing and presentation machinery. Taken together, a robust and streamlined framework was built that facilitates peptide purification and its application in basic and translational research.

Furthermore, recent studies have shed light that, along with the highly attractive mutated neoantigens, other non-mutated, yet tumor-specific, non-canonical antigens may also play an important role in anti-tumor immunity. Non-canonical antigens are of presumed non-coding origin and not commonly included in protein reference databases, and are therefore typically disregarded in database-dependent MS searches. The second part of this thesis develops an analytical workflow enabling the confident identification and validation of non-canonical tumor antigens. For this purpose, whole exome sequencing, bulk and single-cell RNA sequencing and ribosome profiling were integrated with MS-based immunopeptidomics for personalized non-canonical HLA peptide discovery. A computational module called NewAnce was designed, which combines the results of two tandem MS search tools and implements group-specific false discovery rate calculations to control the error specifically for the non-canonical peptide group. When applied to patient-derived melanoma cell lines and paired lung cancer and normal tissues, NewAnce resulted in the accurate identification of hundreds of shared and tumor-specific non-canonical HLA peptides. Next, the level of non-canonical peptide confirmation was tested in a targeted MS-based approach, and selected non-canonical peptides were extensively validated for one melanoma sample. Furthermore, the novel open reading frames that generate a selection of these non-canonical peptides were found to be actively translated by ribosome profiling. Importantly, these peptides were assessed with autologous immune cells and a non-canonical immunogenic epitope was discovered from an alternative open reading frame of melanoma stem cell marker gene ABCB5.

This thesis concludes by highlighting the possibility of incorporating the proteogenomics pipeline into existing neoantigen discovery engines in order to prioritize tumor-specific non-canonical peptides for cancer immunotherapy.

RESUME GRAND PUBLIC

Maladie très hétérogène et multifactorielle, le cancer représente à ce jour la seconde cause de décès dans le monde. Bien que le système immunitaire soit capable de reconnaître puis d'éliminer les cellules cancéreuses, ces dernières peuvent à leur tour s'adapter et accumuler des mutations leur permettant d'échapper à cette reconnaissance.

L'immunothérapie anti-tumorale démontre le rôle clé de l'immunité dans l'éradication des tumeurs. Cependant, ces thérapies prometteuses ne sont efficaces que chez une petite proportion des patients traités. Une étape majeure dans l'établissement d'une réponse immunitaire anti-tumorale est la reconnaissance d'antigènes associés aux tumeurs. Des études récentes ont montré que les antigènes tumoraux issus de régions non-codantes du génome (antigènes non-canoniques) peuvent jouer un rôle clé dans l'induction de réponses immunitaires. Ainsi, l'identification de ces antigènes tumoraux particuliers permettrait de guider le développement d'immunothérapies anti-cancéreuses personnalisées telles que la vaccination ou encore le transfert adoptif de lymphocytes T reconnaissant ces cibles. La spectrométrie de masse (MS) est une technique non biaisée permettant l'identification et l'analyse du répertoire des antigènes présentés *in vivo*. Cependant, cette technique nécessite d'être optimisée et standardisée afin d'être utilisée en clinique.

Ainsi, la première partie de ces travaux de thèse a été dédiée à l'optimisation expérimentale de cette méthode à partir d'échantillons de tissus et de lignées cellulaires. En comparaison avec les protocoles standards, cette technique permet une couverture plus complète, rapide et reproductible du répertoire de peptides apprêtés aux HLA.

La seconde partie de cette thèse a été consacrée au développement d'une méthode permettant l'identification d'antigènes tumoraux non-canoniques via le séquençage d'ARN cellulaire, ribosomique et l'utilisation de notre méthode d'immunopeptidomique optimisée. Afin de contrôler l'identification de faux positifs, nous avons élaboré un nouveau module computationnel. Ce module a permis l'identification de plusieurs centaines de peptides-HLA non-canoniques, partagés et spécifiques au mélanome et au cancer du poumon. Le séquençage des ARN ribosomiques a mis en évidence la traduction de nouveaux cadre ouverts de lecture desquels sont traduits de nouveaux peptides non-canoniques. Cette technique nous a permis de mettre en évidence un épitope immunogène issu du gène ABCB5, un marqueur de cellules souches cancéreuses préalablement identifié dans le mélanome.

De manière globale, ces travaux de thèse, alliant immunopeptidomique et protéogénomique, ont permis la mise au point d'une méthode expérimentale permettant une meilleure identification d'antigènes tumoraux. Nous espérons que ces résultats amélioreront l'identification et la priorisation de cibles pertinentes pour l'immunothérapie anti-cancéreuse en clinique.

ABBREVIATIONS

ABCB5	ATP-binding cassette sub-family B member 5
ACT	Adoptive cell therapy
APCs	Antigen presenting cells
b2m	Beta-2-microglobulin
CARs	Chimeric antigen receptors
ccRCC	Clear cell renal cell carcinoma
CID	Collision induced dissociation
CLIP	Class II-associated Ii peptide
CTLA4	Cytotoxic T-lymphocyte-associated protein 4
CTNNB1	Catenin beta 1
DAC	Decitabine
DCs	Dendritic cells
DDA	Data dependent acquisition
DIA	Data independent acquisition
DRiPs	Defective ribosomal products
ECD	Electron capture dissociation
ETD	Electron transfer dissociation
ELISpot	Enzyme linked immunospot assay
ER	Endoplasmic reticulum
ERV	Endogenous retrovirus
ESI	Electrospray ionization
FAIMS	High field asymmetric waveform ion mobility spectrometry
FDR	False discovery rate
GTEx	The genotype-tissue expression (project)
HCD	Higher-energy collisional dissociation
hERV	Human endogenous retrovirus
HI	Hydrophobicity index
HLA	Human leukocyte antigen

HLA-I	HLA class I
HLA-II	HLA class II
HLAIp	HLA class I peptides
HLAIIp	HLA class II peptides
HPLC	High performance liquid chromatography
IEDB	Immune epitope database
IFNγ	Interferon gamma
ipMSDB	Immunopeptidomics mass spectrometry database
LC-MS	Liquid chromatography – mass spectrometry
lncRNAs	Long non-coding RNAs
LINEs	Long interspersed nuclear elements
m/z	Mass-to-charge ratio
MAE	Mild acid elution
MALDI	Matrix-assisted laser desorption/ionization
MHC	Major histocompatibility complex
MITF	Microphthalmia-associated transcription factor
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NewAnce	A new analytical approach for non-canonical element identification
NGS	Next generation sequencing
noncHLAp	Non-canonical HLA peptides
noncHLAIp	Non-canonical HLA-I peptides
ORF	Open reading frame
PBMC	Peripheral blood mononuclear cell
PDAC	Pancreatic ductal adenocarcinoma
PD-1	Programmed cell death protein 1
PD-L1	Programmed cell death protein ligand 1
PLC	Peptide loading complex
PTM	Post-translational modification
PRM	Parallel reaction monitoring
protHLAIp	Proteome-derived HLA-I peptides

Ribo-Seq	Ribosome profiling
RNA-Seq	RNA sequencing
RT	Retention time
scRNA-Seq	Single cell RNA sequencing
SRM	Selected reaction monitoring
TAA	Tumor-associated antigen
TAP	Transporter associated with antigen processing
TCGA	The cancer genome atlas
TCR	T cell receptor
TECs	Thymic epithelial cells
TE	Transposable element
TILs	Tumor infiltrating lymphocytes
TIMS	Trapped ion mobility spectrometry
TOF	Time-of-flight
TYR	Tyrosinase
TYRP1	Tyrosinase-related protein 1
UniprotKB	Uniprot knowledgebase
WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1

Introduction

Cancer Immunity

The Antigen Presentation System

Tumor Antigens

Immunopeptidomics For Antigen Discovery

Mass Spectrometry For Peptide Identification

Proteogenomics

Chapter 1 INTRODUCTION

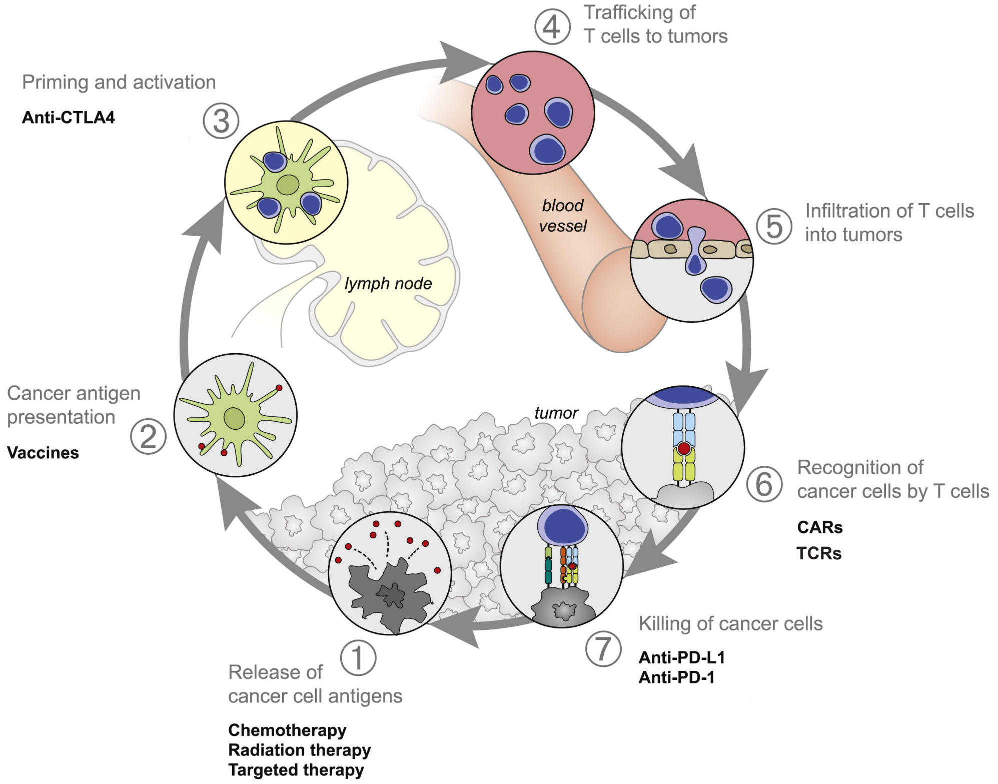
The presented work aims to advance the field of cancer immunotherapy through the integration of both large-scale and multidisciplinary approaches. As such, this introduction provides the required background into the various related fields. Briefly, the relationship between cancer and the immune system, as well as existing cancer immunotherapies, are discussed. Following this, the antigen processing and presentation machinery, and the different types of presented antigens, are outlined. Subsequently, a separate class of antigens, the non-canonical antigens, are covered in detail. Several routinely applied methods for antigen identification and tumor immunogenicity assessment are described. Thereafter, MS-based immunopeptidomics is introduced as an advanced tool for large-scale antigen discovery, followed by the fundamentals of the required MS techniques. Finally, to unravel the breadth of non-canonical epitopes, proteogenomics approaches are outlined, along with their statistical and computational challenges that hinder their robust identification. When combined together, the integration of the multidisciplinary topics covered in this introduction have the potential to advance cancer immunotherapy at a systems-level.

1.1 Cancer Immunity

Undoubtedly, the immune system plays a significant role during cancer development [1-3]. Due to genetic and molecular alterations that occur during cancer progression, the immune system has the capability to recognize cancer cells as non-self and foreign. In this manner, the immune system surveils the body for pre-cancerous cells and should eradicate these before they can properly become transformed. However, sadly, the immune system does not fully protect most individuals from the growth of cancerous cells. A plethora of factors ultimately dictate whether an effective anti-tumor immunity can be initiated and maintained, either spontaneously or through therapeutic intervention. This is described in the cancer immunity cycle, which depicts a series of steps that must be engaged in order to allow the effective elimination of cancer cells [4] **(Figure 1)**.

1.1.1 The Cancer Immunity Cycle

In an ideal and simplified self-propagating cancer immunity cycle, dying cancer cells release antigens that can be taken up by dendritic cells (DCs) at the cancer site. These DCs travel back to the lymphoid organs, where they act as professional antigen presenting cells (APCs) to prime and activate T cells through engagement with cognate T cell receptors (TCRs). In the right context, and upon other stimulatory signals, activated T cells can traffic to the tumor, infiltrate into the tumor bed and interact specifically with cancer cells. Lastly, the T effector cells kill the cancer cells, leading to more tumor antigens being released and the propagation of the cycle. However, in every step, there is delicate balance between stimulatory and inhibitory factors that ultimately dictates whether the cycle progresses, or halts altogether. Critically, tumors have evolved multiple mechanisms to evade destruction by the immune system [5]. The cycle may stop due to tumor antigens not being readily accessible, for example, through their downregulated expression on the cell surface, or the impairment of antigen uptake by immune cells. Alternatively, this stalling can also be caused by the incomplete activation and homing of T cells to the tumor. Furthermore, many factors in the tumor microenvironment are highly immunosuppressive, such as through the upregulation of inhibitory signals, and the recruitment of suppressor immune cells [6, 7]. Overall, the objective of cancer immunotherapy is to trigger the propagation and amplification of a halted cycle, ultimately resulting in the efficient elimination of cancer cells.



④ Trafficking of T cells to tumors
Priming and activation
Anti-CTLA4
③
⑤ Infiltration of T cells into tumors
blood vessel
lymph node
Cancer antigen presentation
②
Vaccines
tumor
⑥ Recognition of cancer cells by T cells
CARs
TCRs
Killing of cancer cells
⑦
Anti-PD-L1
Anti-PD-1
①
Release of cancer cell antigens
Chemotherapy
Radiation therapy
Targeted therapy

*Figure 1 – **The cancer immunity cycle** is depicted above, and specific therapies that target particular steps are shown in bold. For example, in Step 1, the release of tumor antigens from the cancer cells can be induced by chemo-, radiation-, or targeted therapies (such as epigenetic modulators). Vaccines act in Step 2 to boost antigen presentation and stimulate T cell responses. In Step 3, Anti-CTLA4 antibodies leads to priming and activation of T cells in the lymph nodes to enable their trafficking into the tumor. In Step 6, the recognition of cancer cells by T cells is enhanced by introducing genetically-modified T cells directed towards the tumor. Lastly, in Step 7, anti-PD-L1 or PD-1 antibodies lift immunosuppression in the tumor bed and activate T cells to kill cancer cells. Adapted from Chen and Mellman, Immunity, 2013 [4].*

1.1.2 Cancer Immunotherapy

Promising immunotherapies exist for both hematological and solid cancer malignancies, and many of them have received considerable attention due to their undisputed potential in re-invigorating a patient's immune system to target cancer [8-10]. It is therefore unsurprising that "cancer immunotherapy" was regarded "Breakthrough of the Year" by the Science journal in 2013 [11]. To this end, with approximately 2,600 clinical trials listed on www.clinicaltrials.gov as of November 2019, cancer immunotherapy is a recent, yet established area in the fight against cancer. The existing approaches to cancer immunotherapy can be broken down into therapies that involve the blockade of immune checkpoints, adoptive cell therapy (ACT), and the use of cancer vaccines, as well as their combinational approaches **(Figure 1)**.

Immune Checkpoint Blockade Therapy

In an effective immune system, multiple inhibitory pathways dampen the activation of immune cells to limit detrimental autoimmune reactions to normal tissues and organs. These checkpoints are often over-activated in the context of cancer, and targeted therapies focus on blocking these checkpoints to induce a successful T cell-based immune response [12-15]. For example, the binding of ligand Cytotoxic T-lymphocyte-associated Protein 4 (CTLA4) on T cells to B7.1/B7.2 on APCs constitutes a T cell inhibitory signal in the peripheral lymphoid tissues. Anti-CTLA4 monoclonal antibodies block this interaction, thereby allowing T cell activation and subsequent expansion. Another important checkpoint is the inhibitory programmed cell death protein (PD)-1/ PD Ligand 1 (PD-L1) axis. The receptor PD-1 can be found on activated effector cells, whereas the PD-L1 resides on tumor cells and other immune cells. When engaged, this axis blocks T cells' ability to produce and secrete cytotoxic modulators. As such, anti-PD-1 and anti-PD-L1 antibodies have been developed to intervene and prevent this inhibitory signal. Targeting the PD-1/PD-L1 inhibitory signaling pathway was demonstrated to be less toxic than the CTLA4/B7 axis, likely due to the specific stimulation of T cells that reside in the tumor bed [16].

The profound potential of this therapy has been shown in many clinical trials, which report durable responses [8, 9, 15]. Importantly, immune checkpoint blockade therapy is approved for several solid cancer types, such as melanoma, non-small cell lung cancer, and renal cell cancer. However, although highly promising in these malignancies, many patients don't respond, or respond incompletely to these inhibitors.

Adoptive Cell Therapy

In ACT, specific cells are infused into the patient to help the body combat a variety of diseases, often T cells in the case of cancer [17, 18]. T cells can be harvested in two ways: either from the patient's own blood, or directly from their solid tumor which results in tumor infiltrating lymphocytes (TILs) [19]. These cells are grown to large amounts (>1 × 10^8), can be selected for specific traits, such as antigen specificity and molecular state, and then re-infused back into the patient to confer anti-tumor immunity. ACT has shown remarkable tumor regression in melanoma patients, especially when using TILs.

Extending this approach, T cells can be genetically modified, for example through the introduction of antigen-specific TCRs, or chimeric antigen receptors (CARs) [20]. TCR-transduced T cells recognize specifically a tumor antigen in association with HLA molecules on the cancer cell [21, 22]. A CAR, on the other hand, is constituted of a chimeric receptor that is directed against a non-HLA restricted surface tumor antigen, along with an intracellular T cell signaling domain. In this manner, the genetically engineered T cells are deployed towards the tumor when re-infused. TCRs can be engineered to target a variety of intra- and extra-cellular processed cancer-specific antigens, whereas CARs are limited to those surface proteins that are inherently cancer specific. The use of CAR T cells has been shown to have successful outcomes when treating hematological malignancies. In comparison, therapies using TCR-transduced T cells have been effective in treating solid tumors, and are currently being evaluated in multiple ongoing clinical trials. These trials mostly target metastatic melanoma, with over half considering cancer-testis antigens such as NY-ESO-1 [21].

Generally, current limitations surrounding the applicability of genetically modified T cells include the existing toxicity issues, the sustainability of the cells *in vivo*, and the delivery system to optimally find, invade and survive in the solid tumor. However, strategies to both control and enhance the activity and specificity of CAR T cells in the tumor microenvironment are being extensively researched [23]. For example, by incorporating chemically disruptable heterodimers into a CAR, the safety of CAR T cell therapy can be improved as their activity can be halted through small molecule disrupters [24]. In addition, systems can be designed that allow controlled and enhanced CAR T cell function, such as through the inducible activation of co-stimulatory molecules [25]. Furthermore, CAR T cells can be genetically engineered to locally secrete immune modulatory factors that could enhance both T cell expansion and anti-tumor effects in the tumor microenvironment [26-29]. Lastly, the advent of utilizing allogeneic TCR-transduced or CAR T cells is particularly intriguing, offering the potential of generic treatments, as well as circumventing the laborious processing of autologous T cells [30]. For this purpose, the CRISPR/Cas9 system has proven to be a valuable technology in the genetic re-engineering of allogeneic T cells. This tool can be applied to disrupt the genes encoding endogenous TCRs and HLA molecules, which reduces the rejection risk of these cells in the host immune system [30].

Cancer Vaccines

The recent establishment that, following immune checkpoint inhibition, the recognition of specific immunogenic peptides presented on a tumor cell triggers T cell activation has led to the renewed interest in cancer vaccines [31, 32]. The goal of cancer vaccines is to allow the presentation of tumor antigens on HLA molecules for recognition by immune cells, thus driving an anti-tumor immune response. These vaccines can be based on either peptides, DNA or RNA, cells or virus vectors.

Peptide vaccines, as an example, are safe for administration to patients, however, the peptides used need to be personalized to the patient due to inherent HLA restrictions, and adjuvants are required for efficient T cell priming through co-stimulation. The selected peptides should be of non-self and foreign origin, and are ideally not presented in any other normal cells in the body. The use of long peptides containing multiple HLA class I (HLA-I) and HLA class II (HLA-II) ligands could allow stimulation of both CD8$^+$ and CD4$^+$ T cells, and superior T cell priming through their uptake and processing by DCs. Such multi-peptide targeting would ideally overcome tumor escape mechanisms, and are regarded as an attractive approach to tackle the heterogeneity imposed by the tumor [33].

In a cellular approach, lysed whole tumor product can be used as a vaccination to boost antigen levels and exposure of immunogenic antigens to T cells [34]. Here, the advantage is that antigens do not need to be pre-defined, and rather result in the presentation of a broad range of epitopes to the immune system. Often, this leads to antigen spreading and the expansion of other subsets of T cells that work synergistically against the tumor. Consequently, cancer vaccines are envisioned to be particularly potent in multimodality treatment options, such as in combination with checkpoint blockade therapy [35-37].

Combinational Approaches for Immunotherapy

Combinations of immunomodulatory treatments tackling different steps of the cancer immunity cycle are being tested in order to improve patient outcome [9, 38]. In particular, many modes of therapies are being combined with PD-1/PD-L1 inhibition, including CTLA4 inhibition, peptide vaccinations and CAR T cells. Additionally, concomitant chemo- or radiotherapy with PD-1/PD-L1 targeting has shown promising results in tumor regression in small cell lung, head and neck squamous cell, and breast cancer [39].

Furthermore, recent studies have shown that epigenetic modulators, in combination with checkpoint blockade therapy, can prime tumors towards an efficient immune attack [40-42]. For example, DNA methylation processes are often exploited by cancer, either to activate oncogenes through de-methylation or to silence tumor suppressor genes through hyper-methylation of promoter regions. DNA methyltransferase inhibitors, a class of epigenetic modulators, trigger gene re-expression by intervening with the methylation process on gene promoter regions [43, 44]. Specifically, Decitabine (DAC), a 5-aza-2'-deoxycytidine, is an epigenetic drug that elicits anti-tumor activity by inducing the widespread re-expression of genes previously silenced through DNA methylation in the cancer. Recent genomic studies have additionally shown that DAC can lead to the transcription of thousands of non-canonical transcription start sites [45]. Further, two separate reports have indicated that DAC induces the expression of endogenous retroviral (ERV) elements [46, 47]. Together, these findings demonstrate that applying DAC can revive the tumor to re-express potentially immunogenic features that contribute towards re-shaping T cell responses in the tumor.

1.1.3 Current Focus In Cancer Immunotherapy

Based on the amount of data gathered across the last years, the most successful immunotherapies are in the area of checkpoint inhibitors (often in combinational approaches) and CAR T cells [9, 10, 48]. Although shown to be very promising, immunotherapy is not successful in all patients. Strikingly, only approximately 13% of patients with various cancer types respond to checkpoint blockade therapy [49]. Predictive biomarkers that are being investigated include the expression of PD-L1 in the tumor, microsatellite instability, as well as tumor mutational burden [50, 51]. Especially the latter has received widespread consideration as it was found to be linked to tumor immunogenicity and can positively predict the efficacy of anti-PD-1 treatment [52, 53]. However, difficulties remain in defining more accurate predictive biomarkers for tumor classification, patient outcome and tumor resistance, as well as, ultimately, to determine the most efficient therapeutic strategy. Consequently, significant research is being undertaken to develop robust biomarkers for cancer immunotherapy.

A complementary aim of cancer immunotherapy strategies is to converge towards a completely personalized approach, where only certain antigen-specific T cells are activated against tumor cells. For this purpose, significant research is focused on identifying and selecting the most immunogenic antigens that would allow the targeting of the tumor cells with high specificity, envisioned to greatly diminish toxicity and on-target off-tumor effects [54]. In particular, there is a need for further research into boosting the immune response by simultaneously targeting multiple immunogenic antigens, and lifting tumor immunosuppression [33]. Furthermore, tumor heterogeneity poses a threat to successful cancer immunotherapy, and the accurate identification of tumor antigens that can be expressed on cancer stem cells, or a clonal population, could help unleash T cell specific responses to attack the cancer at its core [55, 56].

1.2 The Antigen Presentation System

The presentation of HLA binding peptides on the cell surface plays a central role in any T cell-based immune response. Therefore, the repertoire of presented antigens must represent the surrounding extra- and intracellular processes at any given time in order to properly convey a threat to the immune system. This endeavor is performed via two distinct systems, both described in this section: the HLA-I **(Figure 2)** and HLA-II antigen presentation pathways [57, 58]. This is followed by discussing deviations to the classical pathways, as well as the impact of antigen presentation in the context of cancer.

1.2.1 HLA-I And -II Antigen Presentation

The two types of HLA molecules, class I and class II, bind antigens for presentation to immune cells. HLA-I molecules are ubiquitously expressed in nearly all nucleated cells. In contrast, HLA-II molecules are found on professional APCs, including B cells, monocytes, macrophages and DCs, as well as in other cell types upon inflammatory stimuli. The modes of presentation differ between the classes: HLA-I molecules present processed peptide fragments from inside the cells, while HLA-II molecules present exogenously-derived antigens. Further, peptide-bound HLA-I and -II complexes are presented to CD8$^+$ cytotoxic T cells, or CD4$^+$ helper T cells, respectively.

There are three classical HLA-I genes (HLA-A, HLA-B, and HLA-C in humans) and six HLA-II genes (HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1). These genes are extremely polymorphic, with more than 19,000 different HLA-I and 7,000 HLA-II alleles reported to date:
http://hla.alleles.org/nomenclature/stats.html.

The HLA-I System

All humans express up to six different HLA-I alleles. Their bound peptides are mostly of amino acid length 8-12, with an average of 9 amino acids. The majority of HLA polymorphisms are found in the peptide-binding groove of the HLA-I molecule. This leads to constraints which result in the binding of peptides that harbor specific residues at amino acid position 2, and at the C-terminus. These anchor residues differ between HLAs, and therefore capture the specificity of the HLA allele through particular peptide binding motifs **(Figure 2B)** [59-61]. The immense diversity of HLA binding specificities result in very different HLA repertoires underlined by an individual's haplotype. By allowing for effective and varied sampling of peptide fragments, the likelihood for the binding and presentation of non-self peptide fragments is increased. This is shown by the fact that heterozygosity in HLA alleles can confer greater protection towards some pathogens and in cancer, in comparison to homozygosity [62-65]. That being said, the inheritance of specific HLA-I alleles has also been linked to susceptibility to autoimmune diseases and infections [66-68].

The classical HLA-I antigen presentation pathway **(Figure 2A)** starts with the degradation of nuclear and cytosolic self- and non-self-proteins by the ubiquitin-proteasome system, which is composed of a 20S core catalytic unit and two 19S caps [69]. The catalytic core usually degrades proteins into peptides of 3 to 22 amino acids, and specifies the C-terminus of peptide fragments [70]. The immunoproteasomes, another type of proteasomes, are constitutively expressed in immune cells, or induced upon stress and inflammation in non-immune cells [71, 72]. Due to molecular changes in the catalytic subunits, the immunoproteasomes harbor altered peptide cleavage preferences and have higher levels of activation in comparison to the constitutive proteasome [73, 74]. It has been shown that immunoproteasomes, upon inflammatory signals, are more suited to process large substrate pools, leading to the faster and greater presentation of peptides on the cell surface, and may ultimately change the repertoire of antigens presented [74-76].

Peptides processed by the proteasome are thereafter subjected to a variety of aminopeptidases in the cytosol [57]. Peptides with sufficient length (typically 9-16 amino acids) that have survived the proteasome and the plethora of aminopeptidases are brought to the endoplasmic reticulum (ER) via the transporter associated with antigen processing (TAP). These peptides can be further trimmed at the N-terminus by ER aminopeptidases ERAP1 and ERAP2. HLA-I molecules, consisting of a heavy chain and the light chain, specifically beta-2-microglobulin (b2m), are assembled in the ER and are associated with calreticulin ERp57, protein disulphide isomerase PDI, the chaperone tapasin and TAP. Together, they form the peptide loading complex (PLC). Once a peptide with an appropriate length and sufficient affinity has bound to the HLA molecule, the complex is stable and the chaperones are released. If a peptide fails to associate with the HLA molecule, it returns to the cytosol for degradation. A complete HLA-peptide complex is transported from the ER to the plasma membrane for presentation. This entire process is limited by many factors, ranging from the half-life of proteins and peptides, to the concentration of proteasomes, and cytosolic and ER associated peptidases. Ultimately, more than 99% of the intracellularly generated peptides do not survive for presentation to the immune system [77].

The majority of presented antigens are generally linked to the level of source protein expression, however, studies have shown that the link with protein translation offers a more accurate representation [78]. As shown in the case of defective ribosomal products (DRiPs), translation of numerous mRNAs can be tightly coupled to protein degradation and their presentation [79-83]. These occur through several mechanisms. For example, translation errors can lead to truncated or misfolded proteins which are rapidly degraded and potentially presented. Additionally, proteins may harbor degradation signals, such as disordered regions, facilitating their immediate degradation and funneling through the antigen presentation pathway. Lastly, translation products with pre-termination codons are guided through the nonsense-mediated decay pathway, thereby further contributing to the antigenic repertoire.

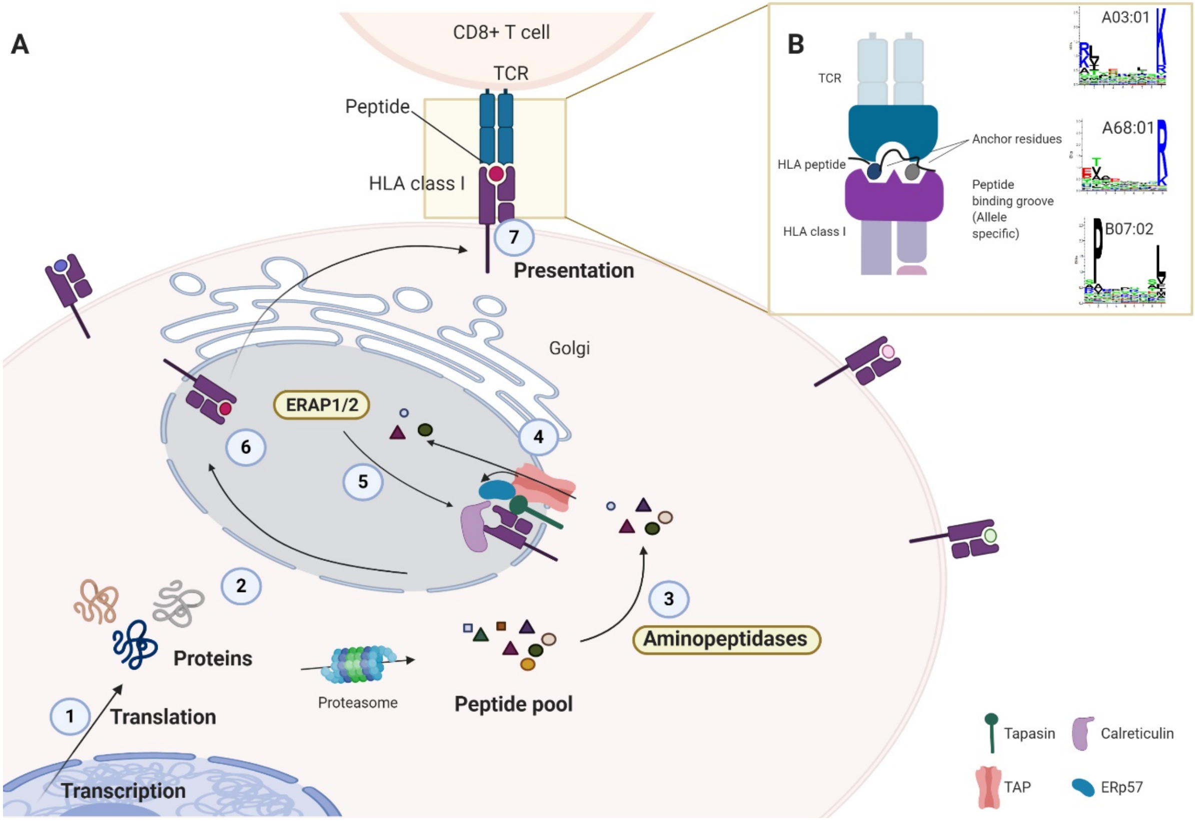


Figure 2 – The HLA-I processing and presentation pathway. ***(A)*** *Intracellular peptide generation for HLA-I presentation begins with (1) the transcription of genomic regions and their translation into proteins. (2) Proteins are eventually degraded by the proteasomes and only a small fraction of the peptides escape complete degradation, and they are further (3) trimmed by a multitude of aminopeptidases in the cytosol, resulting in differing lengths of peptides. (4) Peptides of specific lengths can be funneled through the TAP, which is part of the PLC. (5) These peptides may associate with the HLA-I molecule either directly, through the PLC, or after further trimming by the ERAP1/2 and other ER resident aminopeptidases. (6) A HLA-I complex bound to a peptide is stable, released from the PLC, and transported through the Golgi to the plasma membrane for (7) interaction with CD8$^+$ T cells.* ***(B)*** *In the top-right box, the interaction of the HLA-I peptide-bound complex with the TCR is magnified. It depicts the binding of a HLA-I peptide with the anchor residues (P2 and C-terminus) that are specific to the different peptide-binding grooves of HLA-I molecules. Three examples of HLA allele-specific motifs are shown. Inspired by Neefjes et al., Nature Reviews Immunology, 2011 [57].*

The HLA-II System

Peptides that are bound to HLA-II molecules vary from those associated with HLA-I molecules. The average peptide fragment length is 15 amino acids, and therefore generally longer than HLA class I peptides (HLAIp) [84]. Their peptide binding grooves are more flexible, with specificities at the 9-mer core, while accommodating longer peptides through extensions outside of the binding pocket. The origin of the associated peptides additionally differ from that of HLA-I, as these are sampled from outside the cell (i.e. from extracellular proteins), as well as from self-proteins that have been degraded via the endosomal pathway. HLA-II molecules (consisting of alpha and beta chains) together with its stabilizing invariant chain Ii, are assembled in the ER and brought to the endosomal pathway. In the endosomes, cathepsins S and L digest Ii, leaving a class II-associated Ii peptide (CLIP) in the binding groove. CLIPs are later exchanged by higher affinity HLAIIp that have been degraded in the endosomal pathway, aided by the HLA-DM chaperone. Thereafter, the completed HLA-II antigen complex is transferred by vesicular transport, or through tubules, to the plasma membrane for presentation.

Cross-Presentation

As with every dogma, there are exceptions to the general rule. In the case of antigen presentation by HLA-I molecules, a crucial part of immune surveillance is formed by the cross-presentation found in phagocytes, for example, DCs. Cross-presentation occurs when phagocytes sample and process extracellular antigens and present them via the HLA-I pathway [58, 85-87]. The peptides are presented on HLA-I molecules and prime CD8$^+$ T cells in the lymph nodes. The pathway that enables exogenous antigens to be processed and loaded onto HLA-I molecules is complex, and is still being unraveled. Current research shows that it may involve the movement of antigens from the phagosome to the cytosol, and their processing by the proteasome [88, 89]. On the other hand, antigens could also be degraded by lysosomal proteases and loaded onto recycling HLA-I molecules, similar to the HLA-II pathway [90]. Interestingly, the antigens that are presented on phagocytes to T cells can differ from those antigens that are ultimately found in the tumor bed. Whether, and how, this truly shapes antigen-specific T cell-based anti-tumor responses is still being extensively researched.

1.2.2 Antigen Presentation In Cancer

Antigen presentation is a crucial mechanism in alerting the immune system to infected cells [91, 92]. However, pathogenic conditions, such as viral infections, and cancer can manipulate the cellular antigen presentation and processing machinery in multiple ways, mostly at the genetic and epigenetic level, and evade recognition by the immune system.

Tumor cells can modulate the antigen presentation machinery through the mutation, silencing or loss of HLA genes, and the introduction of defects in the proteasomal and aminopeptidase components [92]. Any perturbations along the antigen presentation pathway leading to lower, altered or a lack of antigen presentation makes a successful immune escape more likely. For example, many tumors downregulate the expression of HLA-I on their cell surfaces. This is linked to a higher rate of tumor progression, a decreased number of T cells found in the tumor bed, and poor patient survival [91, 93-95]. Furthermore, the downregulation of either ERAP1/2 or TAP1/2 has been shown to reduce the antigen repertoire on the surface of cells, and has been directly linked to tumor progression [92]. Lastly, in order to avoid immune recognition, tumor cells can adapt their antigenic repertoire dynamically, a feature which can be further shaped with the application of certain drugs [96].

Due to the potential of cancer induced antigenic changes, it is vital to deduce the range of antigens presented by each tumor. The field of cancer immunotherapy is ultimately fueled by the knowledge of these antigens, and they aid in the development of peptide vaccination strategies and antigen-specific T cell-based therapies.

1.3 Tumor Antigens

The knowledge surrounding the origin, classification and characterization of tumor antigens is key to the development of efficient immuno-therapeutic approaches, and discussed in this section. Specifically, mutated neoantigens, which currently represent the greatest opportunity in immunotherapy, will be described in more detail. A separate class of antigens, typically presumed to be non-protein-coding, will then be thoroughly outlined. Finally, this is followed by the approaches that are routinely utilized for the identification and prioritization of tumor antigens for clinical applications.

1.3.1 Canonical Tumor-Associated Antigens

There are six broad classes of canonical tumor antigens: differentiation, overexpressed, cancer-testis/germline, post-translationally modified (PTMs), viral, and tumor-specific [97-101]. The properties for each of these antigen classes are discussed below and summarized in **Table 1**.

DIFFERENTIATION ANTIGENS

Differentiation antigens are derived from proteins typically expressed during cellular differentiation, such as in melanosome biogenesis. In the case of melanoma, these differentiation antigens have been found to be re-expressed, presented, and tumor-associated. Examples of differentiation antigens include peptide sequences from MART1 and gp100, which were both discovered by MS-based approaches [102]. Their clinical application has shown modest success so far, in part due to central tolerance mechanisms eliminating high affinity and self-reactive T cell clones [103].

Table 1 *- Summary of the different canonical tumor-associated antigen types. The advantages and disadvantages for use in cancer immunotherapy of each tumor antigen type are outlined, along with some antigen examples and their incidence in different cancer types. Information was derived from multiple sources: Schmidt and Lill, Journal of Proteomics, 2019, [99], Smith et al., Nature Reviews Cancer, 2019 [100] and Ilyas and Yang, Journal of Immunology, 2015 [98].*

Canonical tumor-associated antigen type	Pros	Cons	Antigen examples	Cancer examples
Differentiation	• Potential for off-the-shelf therapy	• Potentially expressed in normal tissues • Lower immunogenicity as self-antigen	• PMEL • TYRP1 • MART1 • gp100	• Melanoma
Overexpressed	• Potential for high peptide abundance • Potential for off-the-shelf therapy	• Expressed in normal tissues • Lower immunogenicity as self-antigen	• HER2 • Mesothelin • EGFR • hTERT	• Melanoma
Cancer-Testis	• High prevalence across tumor types and patients • Potential for off-the-shelf therapy	• Potentially expressed in normal tissues • Lower immunogenicity as self-antigen	• MAGE • PRAME • NY-ESO-1	• Melanoma
Post-translational modification	• Potential of tumor-specific deregulated pathway	• Potentially expressed in normal tissues • Lower immunogenicity as self-antigen	• pNCOA-Phosphopeptide [104]	• Leukemia

Overexpressed Antigens

Overexpressed antigens come from proteins that are normally found in a healthy state, such as EGFR, hTERT, p53, and carbonic anhydrase IX. However, their overexpression can serve as markers for tumorigenic cells [98, 105]. Their downregulation is often not possible, as cells require their expression to survive, therefore, they can generally serve as good target candidates for immunotherapy. However, there are often undesired on-target off-tumor and toxicity issues when targeting these antigens, as they are inherently self-proteins [106, 107]. These drawbacks also apply to the class of tumor differentiation antigens mentioned above.

Cancer-testis antigens

The first cancer-testis antigens were discovered by Boon and colleagues, and encompass the MAGE families of cancer-testis derived proteins [108-110]. To date, hundreds of cancer-testis antigen families have been discovered, albeit their function remains largely unknown. These proteins are expressed only in testes, which do not produce HLA molecules and therefore their antigens are never presented on the surface of these cells. In the case of the tumor, re-expression of these genes can result in their presentation, and consequently their detection by the immune system. Such cancer-testis antigens have been found in several malignancies, ranging from lung, breast, ovarian, colon cancer, multiple myeloma and melanoma [111]. Given its widespread expression in tumors, their immunotherapeutic potential is being tested in several clinical trials, especially for the well-studied MAGEs in melanoma [112-114].

Post-Translational Modifications

PTMs on proteins are crucial for dictating protein-protein interactions and downstream signaling processes, and the dysregulation of these pathways represents a hallmark of cancer pathogenesis. MS uniquely allows the comprehensive screening of global PTM protein signatures, including phosphorylation, glycosylation, acylation, and ubiquitination [115-117]. In particular, aberrant phosphorylation has been implicated in oncogenesis, and can survive the antigen processing and presentation machinery to generate cancer-specific phosphorylated HLAIp [118, 119]. Importantly, these phosphorylated HLAIp have been shown to elicit immune responses in primary leukemia samples [104]. Similarly, glycopeptides, such as O-linked β-N-acetylglucosamine peptides, have also been reported to be immunogenic in leukemia [120]. Collectively, PTM peptides further expand the scope of available antigens to explore for cancer immunotherapy.

OncoViral antigens

Oncoviral antigens are those found in virus-associated cancers, such as in head and neck cancer caused by human papilloma virus HPV-16, cervical and anal cancers caused by HPV-18, and hepatocellular cancer caused by hepatitis B and C [121-123]. These proteins and their resulting peptides are truly “non-self”, therefore, they represent very promising tumor-specific antigen targets **(Table 2)**. However, this benefit is only applicable to a fraction of cancers that are associated with the incidence of viral infection.

Tumor-specific antigens

Lastly, tumor-specific antigens are a category that are exclusively found in the tumor and arise from somatically acquired genomic alterations, and are not found in any healthy tissues **(Table 2)**. They include antigens derived from nonsynonymous single nucleotide variants, frameshifts through nucleotide insertions or deletions, and gene fusions [101]. Such tumor-specific neoantigens show the most promising potential for use in targeted immunotherapy due to their absolute tumor specificity, and are described in more detail in Section 1.3.2 below.

1.3.2 Mutated Neoantigens

The first discoveries of mutated cancer peptides were found in the genes MUM1 and beta-catenin [124, 125]. They were confirmed to be presented on HLA molecules, and recognized by T cells. With the advent of next generation sequencing (NGS) technology in 2005, the possibility to map the tumor mutational landscape and to search for immunogenic mutated antigens both increased markedly, especially for nonsynonymous single-nucleotide variant antigens [126, 127]. Landmark publications included the use of NGS and advanced T cell assays, have showed the significance of mutated antigens in both mouse models and, shortly after, in cancer patients using similar techniques [128-131]. Neoantigen-specific T cell responses were retrospectively observed in TIL products of melanoma and also across other cancer malignancies. Moreover, such responses can be positively correlated with the efficacy of immune checkpoint blockade and TIL therapy, providing further evidence that mutated neoantigens play an important role in efficient tumor control [52, 101, 131, 132]. The potential number of mutated neoantigens that are presented in the tumor can be extrapolated from the tumor mutational burden [133-135]. This has been used as a marker to predict tumor immunogenicity and measure the success of checkpoint blockade treatment. Unfortunately, it was shown that neoantigen burden is a weak marker when considered alone. This is due to only a minute fraction of mutations leading to actionable antigens, while the contribution from alternative factors, such as other types of antigens, or immunosuppressive mechanisms, likely play an equally important role [136].

Table 2 – *Summary of the different canonical tumor-specific antigen types. Abbreviations are as follows: HCC: Hepatocellular carcinoma, MSI: Microsatellite instability, RCC: Renal cell carcinoma, CML: Chronic myelogenous leukemia. Information was derived from multiple sources: Schmidt and Lill, Journal of Proteomics, 2019, [99], Smith et al., Nature Reviews Cancer, 2019 [100] and Ilyas and Yang, Journal of Immunology, 2015 [98].*

Canonical tumor-specific antigen type	Pros	Cons	Antigen examples	Cancer examples
Oncoviral	• Likely shared between patients	• Only in virally-infected tumors	• HPV E6/E7 • EBV • MCC	• Head and neck cancer • Cervical/anal cancer • HCC
Mutations (single nucleotide variants)	• Private to the patient • Potential targeting of shared driver/pathogenic mutations	• Usually specific to individual tumor • "Off-the-shelf therapy" limited to a few frequent mutations, reaching a small fraction of patients	• KRAS (truncal) • TP53 (truncal)	• Melanoma • Glioblastoma • Lung cancer • Bladder cancer
Insertion/deletion frameshift	• Potentially many targets per mutation • Private to the patient • Potential targeting of shared driver/pathogenic mutations	• Lower prevalence	• TGFBR2	• MSI high tumors • RCC
Fusion protein	• Potential driver gene • Likely shared between patients	• Usually specific to a tumor type	• BCR-ABL	• CML

Other underlying cancer-associated genomic changes can result in sequence insertions and deletions, or in the generation of fusion genes [100]. Specifically, insertion and deletions that trigger frameshifts in the tumor genome leading to novel open reading frames (ORFs) that harbor an array of entirely unique peptide sequences not found in normal tissue [137-141]. For example, this occurs in microsatellite instability high tumors, such as renal cell carcinomas, where there are mutations in DNA mismatch repair proteins. Clinical relevance has been shown from a shared neoantigen that resulted from a frameshift mutation of the gene TGFBR2 [100, 142]. On the other hand, gene fusions, such as the BCR-ABL in chronic myelogenous leukemia, give rise to altered proteins and novel antigens that cover the breakpoint, and can be shared across patients [143]. In general, only modest clinical efficacy has been shown with fusion peptides in various malignancies. However, due to the importance of fusion genes in tumor progression, especially in the case of driver genes, there is ongoing research in the hope to generate universal off-the-shelf treatments [144].

While there had been huge strides in identifying and validating neoantigens and their induced T cell responses, the number of mutated epitopes that were found to be immunogenic in patients are still disappointingly low [101]. They are also mostly unique to the patient's cancer and therefore, fully personalized approaches are required. Furthermore, substantial research focus is currently limited to highly somatically mutated cancers, such as melanoma and lung cancer. Therefore, the search for immunogenic tumor antigens beyond the canonical remains necessary to increase the range of targetable epitopes [145].

1.3.3 Non-Canonical Antigens

Non-canonical antigens are those derived from regions outside of the canonical proteome-derived space, and could expand the options available for cancer immunotherapy [145]. The potential of non-canonical sources of immunogenic tumor antigens is being increasingly recognized, primarily thanks to traditional targeted and reductionist approaches. Importantly, these alternative sequences may be shared across cancers, and could offer generalized treatment to a larger cohort of patients.

The historical timeline of non-canonical peptide research stretches over the last 30 years. In 1989, the Boon group first hypothesized that existence of non-proteome-derived immunogenic epitopes [146]. Remarkably, the first non-canonical antigen derived from an intronic region of MUM1 was also the first mutated neoantigen to be identified [124]. Following this, Malarkannan et al., in 1995, identified a non-ATG ORF that resulted in the identification of a non-canonical major histocompatibility (MHC) class I peptide bound to the Kb MHC molecule [147]. The groups of Boon and Rosenberg later described the existence of intronic and alternative ORF-derived sequences found on HLA molecules in melanomas [148-151]. Reverse strand mRNA transcription can also give rise to antigens as discovered in 1999 [152], and, in 2008, the group of Stevanovic utilized MS to determine a non-canonical vascular endothelial growth factor T cell epitope from an alternative start codon [153]. Many publications have since reported non-canonical peptides derived from presumed untranslated regions of mRNAs, such as through post-transcriptional events, (long) non-coding RNAs, pseudogenes and transposable elements (TEs), and which, importantly, can be recognized by T cells **(Table 3)**. Below, a broad overview of these findings is provided.

***Table 3** - Summary of the different non-canonical antigen types. Abbreviations are as follows: RCC: Renal cell carcinoma, ccRCC: Clear cell renal cell carcinoma. Information was derived from multiple sources, including: Schmidt and Lill, Journal of Proteomics, 2019, [99], Smith et al., Nature Reviews Cancer, 2019 [100] and Ilyas and Yang, Journal of Immunology, 2015 [98].*

Non-canonical antigen type	Pros	Cons	Antigen examples	Cancer examples
Proteasome-generated spliced	• Potentially many targets per type • Could be shared between patients	• Difficulties in validation of translation products • Potentially expressed in normal tissue	• gp100 [154] • FGF-5 [155]	• Melanoma • RCC
Transposable element			• HERV-E	• ccRCC • Low grade glioma • Testicular cancer
Alternative ORF			• gp75 • NY-ESO-1 • VEGF • M-CSF • TRP1 • BING4	• Melanoma • RCC • Kidney tumor
(long) ncRNAs			• MELOE [156]	• Melanoma
Reverse strand			• RU2 [152]	• Kidney tumor
Intronic			• GnT-V [148] • TRP-2 [157] • gp100 [150] • MUM1 [124]	• Melanoma
Pseudogene			• NA88-A [158]	• Melanoma

Post-transcriptional events, such as alternative splicing, intronic retention, non-canonical translation initiation, and codon read-through, can result in the generation of non-canonical antigens. For example, an antigen derived from a splice variant from the gene WT1 has been previously reported in leukemias, lung cancer, and kidney cancer, however, it has not yet been clinically validated [100, 159]. In large-scale analyses using The Cancer Genome Atlas (TCGA) dataset, splice variants were found to be enriched in tumors, and potentially expand the pool of non-canonical antigens to be explored [160].

In the family of (long) non-coding RNAs (lncRNAs), pseudogenes and reverse transcriptions, some were shown to have translation potential and can generate peptides that stimulate T cell responses [152, 158, 161]. In a study involving melanoma patients, antigens were found from the long "non-coding" RNA meloe [156, 162, 163]. Meloe is typically transcribed in a tissue specific manner in the melanocytes, however, the researchers found that translation of this lncRNA occurred in melanoma cells. Evidence was also provided of T cell responses against these non-canonical antigens in melanoma patients and healthy individuals.

Lastly, TEs make up 60% of the genome and were once classified as junk DNA [164]. Of that, ERV elements make up 8% of the genome. They are essential contributors in evolution through DNA insertions, leading to gene mutations, transcriptional modulation, dispersion of regulatory sequences and genomic recombination. Although TEs are usually silenced via epigenetic modulation, these mechanisms might be dysregulated in the context of cancer, or can be induced with epigenetic modulators. Of the retrotransposons, long interspersed nuclear elements (LINEs) have been found to impact cancer biology, as have human ERVs (hERVs) in the context of for example melanoma, ovarian and prostate cancer [100, 165-167]. The dysregulation of TEs can lead to double stranded DNA sensing and inflammatory responses, and their peptide products have been shown to lead to B and T cell activation. In clear cell renal cell carcinoma (ccRCC), a T cell clone targeting a hERV peptide was found and described in two separate publications [168, 169]. Currently, the clinical studies for this hERV peptide is ongoing in ccRCC for usage in ACT.

1.3.4 Methodologies For Antigen Discovery And Validation

The discovery and validation of tumor antigens is currently achieved in multiple ways. Traditionally, antigen discovery techniques have utilized antibodies and patients' T cells for identification, and more recently, routine reverse immunology and MS-based approaches [106, 170]. Regardless of the technique, the selection and prioritization of any antigen of interest can be performed by HLA binding prediction tools, and should be thoroughly evaluated for immunogenicity in downstream analyses.

Computational prediction of (neo) antigens

Many HLA binding prediction tools exist that are trained on experimental HLA binding affinity data, collected in the Immune Epitope Database (IEDB) [171]. More recently, these tools, especially for HLA-I, have started to incorporate cleavage specificities dictated by the proteasome [172], or take into account TAP transport efficiency [173], immunogenicity scores, or eluted ligand information [174-177]. Especially the latter has shown to greatly improve epitope prediction. The reliability of the tools related to HLA-I prediction have far outreached those of HLA-II, due to the greater availability of training data for HLA-I, as well as more distinct binding constraints. However, recent advances in the HLA-II binding predictors have been reported with improved accuracy [178-180], achieved through the use of MS-based eluted ligand information, along with the application of neural networks and sophisticated algorithms.

With the growing interest in pinpointing targetable antigens, HLA binding prediction tools are being used to predict and prioritize those that might bind to the patient HLA [127, 175, 176, 181-183]. Specifically, for mutated neoantigens, the nonsynonymous somatic mutations can be identified through whole exome sequencing (WES) and RNA-Seq, performed on the tumor and healthy matched counterpart (such as peripheral blood mononuclear cells; PBMCs). From the generated information on somatic mutations, the prediction tools provide a list of potential neoantigens that encompass the mutation, ordered by their predicted affinities (or ranks) to bind to respective HLA-I or –II allotypes. By leveraging information on gene expression, RNA-Seq data can further help prioritize antigens that are likely to be presented.

Almost all antigen discovery pipelines rely heavily on HLA binding predictions, which define the antigens that are evaluated for tumor control. Overall, immunogenicity screening techniques are extensive, and, therefore, the further development of *in silico* tools for antigen prioritization is crucial to narrow down targets for cellular validation. However, across the field, there is currently no standardized approach to perform this process, and some drawbacks in HLA prediction tools exist. For example, training data is limited for some rarer alleles, which could negatively impact the predictions for those alleles substantially [182]. Furthermore, binding predictions typically do not take into account all integrative parts of HLA processing [184]. The most comprehensive method to achieve an accurate *in vivo* representation of HLA repertoire is by MS (see Section 1.4).

Evaluation Of Peptide Immunogenicity

Regardless of whether HLA binding prediction tools, MS analyses, or a combination of both are used, the resulting peptides should be screened for immunogenicity, a key success indicator for clinical interventions [185]. A variety of *in vitro* screening techniques exist to evaluate the immunogenicity of a (neo) antigen. Traditionally, researchers worked extensively in screening CD8$^+$ T cells for recognition of target cells transfected with tumor cDNA library pools and HLA restriction elements. This technique led to the identification of the first mutated neoantigen by Coulie et al [124]. Since the dawn of NGS, researchers are facing new challenges, with large numbers of peptides that need to be interrogated for antigen reactivity with limited sample material.

Addressing this challenge, one very common and invaluable high-throughput method is to screen for certain responses in (autologous) immune cells by enzyme-linked immunospot (ELISpot) assays [186, 187]. ELISpot assay plate surfaces are coated with cytokine specific monoclonal antibodies, incubated with cells and stimulant, and the secretion of specific cytokines are captured. Unspecific interactions are washed away, before a cytokine specific detection antibody fused to an enzyme conjugate is added. Lastly, the addition of the substrate allows for spot visualization. In this manner, the frequency of reactive cells can be measured quantitatively, and with high sensitivity, by counting the spots formed. Of the variety of molecules that can be measured, interferon gamma (IFNγ) is by far the most common, especially when interrogating activated CD4$^+$ and CD8$^+$ T cells upon antigen stimuli. Some protocols may require *in vitro* expansion and stimulation of immune cells in order to amplify low frequency responses to detectable levels compatible with ELISpot. Spot counting can be performed manually, but throughput and reproducibility has been significantly increased with the development of semi- or fully automated plate readers [188]. This straightforward approach led to the successful identification of immunogenic neoantigens in melanoma, non-small cell lung, and ovarian cancer [189-192].

In parallel, antigen specificity can be interrogated by screening T cells with peptide-HLA multimers [185, 193, 194]. These assays are used to detect, quantify and isolate T cells that bind to a certain antigen of interest. Multimers come in a variety of subtypes, such as in the form of tetramers, pentamers and dextramers. These consist of HLA molecules each bound to a specific peptide, with which antigen-specific T cells can be extracted from a pool of diverse T cell specificities using fluorochrome-based techniques. This multimer assay relies on the *a priori* knowledge of the minimally processed epitope, often determined through HLA binding prediction algorithms. To accommodate multiple peptide screenings in one sample, fluorochrome-based combinatorial staining techniques have been optimized in TILs [195]. Alternatively, to allow for high-throughput screening of over a thousand of peptide specificities within a single sample, DNA barcoded peptide-HLA multimers have been designed and exploited for T cell recognition profiling [196]. While the exploration of peptide-HLA-II multimers currently lags behind HLA-I, this gap may be closed with the growing development of HLA-II binding prediction tools.

In contrast, unbiased evaluation of mutated epitopes can be performed without the need for HLA binding prediction algorithms [170, 185, 197-199]. For each nonsynonymous mutation detected, a gene fragment is designed, where the mutation is flanked by their original sequences, and up to twenty-four mutations in minigenes can be accommodated within an ORF. These genes are *in vitro* transcribed to RNA, transfected into APCs, and then used to screen for immunogenicity via T cell responses. Any responses are subsequently deconvoluted to pinpoint the minimal epitope. The advantages of this technique are that prior knowledge of

HLA restriction, and the minimal epitope, is not needed, and epitopes presented have been naturally generated through the cellular antigen processing and presentation pathways.

In parallel to existing methodologies for antigen discovery, MS-based techniques uniquely allow the unbiased exploration of exact epitopes that are naturally presented *in vivo* on tumor cells. Thus, this intensive area of research offers a complementary approach to perform high-throughput antigen discovery.

1.4 Immunopeptidomics For Antigen Discovery

With the rise of MS-based technologies, the field of immunopeptidomics has emerged that aims to map the thousands of peptides presented on cells' transmembrane HLA molecules. Due to the strong link between HLA genomic regions and the incidence of immune diseases, the importance of charting the HLA peptide repertoire, termed the immunopeptidome, is widely accepted [200]. Critically, both the qualitative, and quantitative traits of the immunopeptidome have been associated with a range of disorders. In this section, a brief background of MS-based immunopeptidomics studies are outlined, followed by the existing methodologies for the enrichment of HLA peptides and the associated challenges.

1.4.1 Immunopeptidomics Background

Nearly three decades ago, the field of immunopeptidomics was established with seminal studies conducted by the groups of Hans-Georg Rammensee and Donald Hunt [61, 201]. Importantly, the presence of different HLA binding motifs was observed. Thereafter, the first mutated MHC-I peptides were sequenced by MS in 1997 by the group of Hunt [202]. Further, sample-specific collective representation of tumor HLA peptides *in vivo* was generated using immunoaffinity purification and liquid chromatography coupled with MS [203, 204].

Since then, MS-based instrumentations have advanced rapidly, allowing the sensitive evaluation and sequencing of tens of thousands of peptide sequences from a given sample **(Figure 3)**. The widespread research and publications in the field of MS-based immunopeptidomics have been enabled by the development of both the mass analyzer Orbitrap, discussed in Section 1.5.3, and sophisticated statistical and computational tools [200, 205]. These advances have led to the MS-based identification of many presented TAAs [102, 203, 206-208]. Importantly, the identification of mutant epitopes in melanoma [189, 190, 209] and glioblastoma [210] has been achieved using customized protein sequence databases incorporating patient-specific mutation information. Furthermore, significant research has been focused on the identification of HLA peptides that harbor PTMs [104, 116, 211], or are of non-canonical origin [145].

Recently, MS-based immunopeptidomics data is being capitalized to train HLA binding predictors, and thus improve antigen prioritization strategies for both HLA-I and HLA-II [175, 178]. Similarly, leveraging immunopeptidomics data, HLA binding predictors have been developed for peptides with PTMs [212]. Thus, the growing field of MS-based antigen discovery is an exciting prospect for advancing personalized immunotherapy [200, 205, 213, 214].

1.4.2 HLA Peptide Extraction From Biological Samples

A prerequisite for MS-based immunopeptidomics is the robust extraction of HLA peptides from biological samples [215]. There are two main methods that are commonly used: mild acid elution (MAE) and immunoaffinity purification [205, 216].

MAE is a straightforward technique that only requires the treatment of the sample with acid, and leads to the direct release of peptides from the cell surface. However, this method is largely unspecific and not feasible to perform with tissue samples. Due to this unspecific "stripping" of cells, the contribution of contaminating peptides is high. An advantage of this approach is that cells remain intact and therefore allows for studies that follow HLA presentation kinetics.

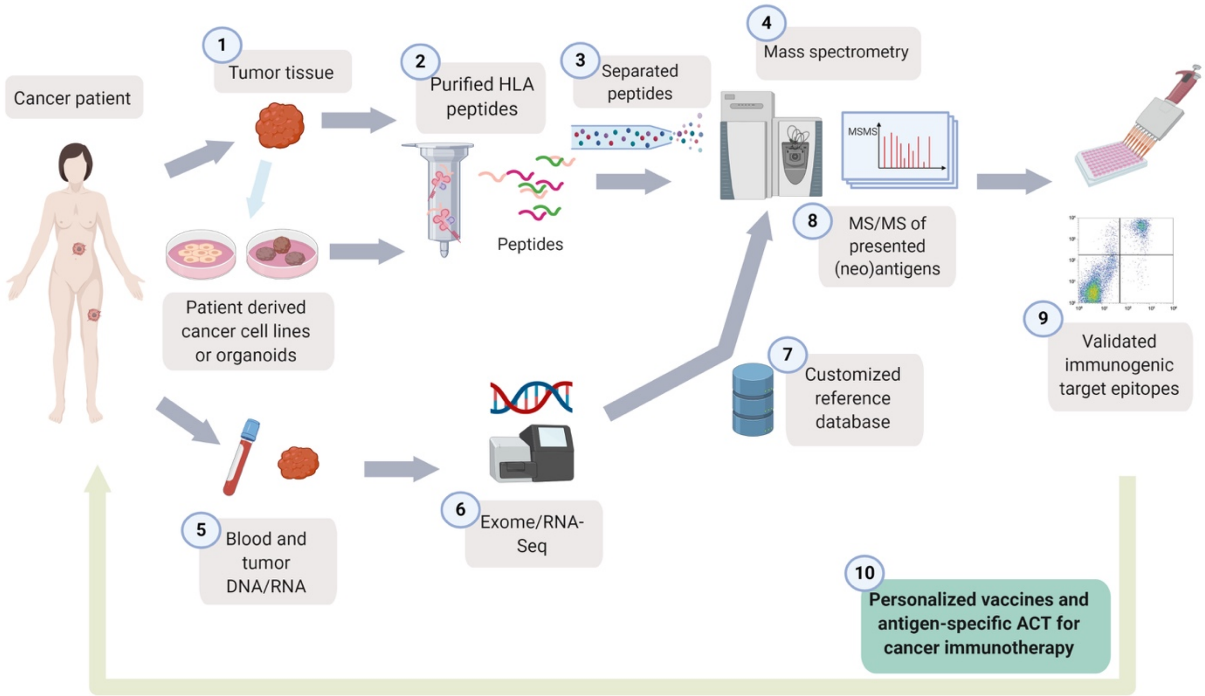

Cancer patient
1
Tumor tissue
Patient derived cancer cell lines or organoids
2
Purified HLA peptides
Peptides
3
Separated peptides
4
Mass spectrometry
MSMS
8
MS/MS of presented (neo)antigens
9
Validated immunogenic target epitopes
7
Customized reference database
5
Blood and tumor DNA/RNA
6
Exome/RNA-Seq
10
Personalized vaccines and antigen-specific ACT for cancer immunotherapy

Figure 3 –MS-based immunopeptidomics for cancer immunotherapy. *(1) Tumor tissue is harvested from a cancer patient, and processed to extract HLA peptides. Alternatively, tumor derived cell lines or organoids can be established and used for immunopeptidomics. (2) HLA immunoaffinity purification is performed, traditionally using chromatography columns, and peptides are separated from HLA complexes through reverse phase C18 extraction. (3) Purified peptides are further separated in a HPLC and (4) directly injected into the mass spectrometer to acquire MS/MS spectra. In parallel to this process, (5) tumor and healthy matched DNA and RNA are extracted, (6) sequencing performed, and (7) the results analyzed to generate protein reference databases personalized to the patient. (8) This is used to interpret MS data and can lead to the identification of TAAs, along with mutated, or other types of non-canonical antigens. (9) The antigens-of-interest are validated for immunogenicity downstream, and (10) if proven to be relevant, may direct vaccine strategies and antigen-specific adoptive T cell therapy for the cancer patient.*

Immunoaffinity purification, on the other hand, employs anti HLA antibodies to capture specific HLA-peptide complexes. Aside from the commonly used anti pan-HLA-I and pan-HLA-II antibodies, a variety of allele-specific antibodies exist that give flexibility for investigating a range of biological questions. After cell or tissue lysates are incubated with the antibodies that have been crosslinked to beads, the bound HLA peptides are dissociated with acid denaturation. Peptides are separated from the heavy chains and in the case of HLA-I, b2m, via molecular weight cut-off spin filters and C18 reverse phase extraction. Depending on the complexity of the peptide mixture, further fractionations can be performed prior to injection into the mass spectrometer. While immunoaffinity purification is typically chosen over the two methods, a disadvantage is that the sample is lost during the HLA extraction process, in contrast to MAE.

Importantly, researchers utilizing these sample preparation methods face a range of challenges, including uncertainty in peptide yield quantity, the need for large amounts of initial sample material, as well as low throughput and low reproducibility issues [215]. Not only do existing processes need to be improved, but the limitations of the current approaches need to be thoroughly evaluated. For example, the uncertainty in peptide yield has been tested by the group of van Veelen [217]. Experimental losses during peptide purification were measured via isotopically labelled peptide-MHC monomers spiked into the cell lysate, prior to immunoaffinity purification. The results showed that immunoaffinity purification is accompanied by loss of up to 99% of the original HLA complexes. Furthermore, an important limitation of pre-fractionation procedures is the creation of extraction bias, which should be taken into account when drawing conclusions on the immunopeptidome [218]. Due to these known issues, quantitative analyses of the immunopeptidome are lagging behind qualitative studies. Moreover, some peptides will never be detected by MS due to high hydrophobicity and poor ionization efficiency. Therefore, as the immunopeptidomics community grows, and the relevance of immunopeptidomics for cancer immunology has become undisputable, there is a need to further develop and standardize these existing HLA extraction processes [200, 215]. Only when HLA extraction methodologies are both robust, and provide high coverage and reproducibility, can they form a good foundation for the exploration of novel antigens and thus help dictate the success of cancer immunotherapy.

1.5 Mass Spectrometry For Peptide Identification

To measure, identify and quantify HLA peptides from biological samples, relevant MS techniques are required. These techniques, along with their corresponding data interpretation processes and the varied acquisition options, are illustrated in this section. Thereafter, MS-based proteomics approaches are introduced, along with their use alongside immunopeptidomics.

Historically, peptide sequencing was performed by Edman degradation, which is time consuming and allows the analysis of only one protein at a time [219]. The method requires a homogenous protein sample as a starting point and the amino acid sequence is identified based on sequential cleavages from the amino terminus end of the protein. In contrast, the development of MS hyphenated with liquid chromatography (LC-MS) enables high-throughput information on the quality, and the quantity of a complex peptide mixture [220]. Thus, LC-MS is undoubtedly the state-of-the-art instrumentation for identifying and determining the abundance of peptides, and their source proteins.

There are a variety of mass spectrometers that detect and identify the mass-to-charge ratios (*m/z*) of ions. These contain core components, including a sample introduction device, a source to produce ions, one or more mass analyzers, a detector to measure ion abundance, and a computer for data processing. While there are a variety of different MS techniques tailored to specific disciplines, these are generally beyond the scope of this thesis and are extensively explained elsewhere [221]. The focus here will be on the techniques that are routinely employed to analyze biologically complex peptide mixtures by two stages of mass analysis (tandem mass spectrometry; MS/MS). The typical peptide separation technique prior to MS analyses is high performance liquid chromatography (HPLC), and is directly followed by electrospray ionization (ESI). Upon ionization, tandem mass spectrometry is performed to measure the *m/z*'s of biomolecular ions with a simple scan (MS1). Subsequently, ion activation using collision-induced dissociation (CID) results in ion fragmentation. From this, a product ion scan (MS/MS) is generated and the structural information of the ion can be determined [222, 223].

1.5.1 High Performance Liquid Chromatography

HPLC is a liquid sample separation technique, and an invaluable tool for the field of proteomics by greatly simplifying the complexity of a sample and facilitating unambiguous downstream peptide and protein identification [224]. While there are separation approaches that exploit different physico-chemical peptide properties, such as separation through ion exchange, hydrophilic interaction and affinities, the most common is reverse phase chromatography based on hydrophobicity. This latter approach utilizes a liquid chromatography microscale capillary column consisting of a stationary phase, for example, silica covalently bound to C18 alkyl chains. The mobile phase, that is more polar than the stationary phase, is composed of a mixture of water with various water-miscible organic solvents, such as methanol, acetonitrile and isopropanol. Before separation, the analyte is dissolved in an acidified solution and added to the mobile phase. Interactions between the stationary phase and the mobile phase with the analyte occur. The analyte distribution is dependent on the type of stationary phase, composition of mobile phase, and hydrophobicity of the analyte itself. In gradient elution, as the concentration of the organic solvent in the mobile phase increases over time, analytes are separated and eluted from the stationary phase based on their increasing hydrophobicity.

1.5.2 Ionization

In order for an analyte to be measured by MS, it first needs to be ionized into a stable gas phase ion. There are different techniques of ionization depending on the molecule of interest. For large biomolecules, such as peptides, the preferred ionization method for MS analyses is by “soft” ionization. Analytes are transferred from the solid or liquid phase to stable gas-phase ions without extensive fragmentation during ionization, either through matrix laser desorption/ionization (MALDI) or ESI, respectively. In MALDI, the laser heat is absorbed by the matrix and the energy is conveyed to the fixed analytes, resulting in their release as gas phase ions. In ESI, a potential is applied between the ESI source and the counter electrode, which produces an electrospray that is aided by the nebulization of an inert gas or by high temperature [225]. The charged droplets reduce their size by coulomb explosions, until ions are ejected from very small droplets as gas phase ions, following either the ion evaporation model or charge residue model. Typically, the ions are multiply charged, however, this largely depends on the peptide sequence and the properties of the amino acids. These are thereafter transferred into the near-vacuum system of the mass spectrometer.

1.5.3 Mass Analyzer

A mass analyzer is an ion detector that measures the abundance of a gas-phase ion by its *m/z*. A mass spectrum is generated where x and y axes correspond to *m/z* and ion abundance, respectively. The mass analyzers can be grouped into three different operation modes: continuous (magnetic sector, quadrupole), pulsed (time-of -flight (TOF)) and ion trapping devices (ion trap, Fourier transform ion cyclotron resonance and Orbitrap) [221, 223]. One of the most sophisticated instruments incorporates an Orbitrap, which achieves an unprecedented high resolution and accuracy in determining *m/z*. The Orbitrap is a spindle-like shaped electrode enclosed by a barrel-shaped outer electrode. An electrostatic field is generated in the Orbitrap, thereby allowing the ions, in a manner dependent on their *m/z*, to orbit the electrode and oscillate in an axial direction. The signals are processed by Fourier transformation to determine the *m/z* of ions. A mass spectrometer can come in a hybrid form consisting of at least two mass analyzers to enable tandem MS.

1.5.4 Ion Activation By Collision Induced Dissociation

As the knowledge of the precursor ion *m/z* is insufficient to distinguish between the thousands of peptides potentially found in the same sample, the fragmentation of peptides is required to deduce sequence information [226, 227]. A commonly used fragmentation technique is (lower-energy) CID, where precursor ions at a given *m/z* are accelerated in a trap containing inert gas. The ion collisions with the gas cause kinetic energy to be partially transformed into internal energy that break lower energy amide bonds within peptides. In higher-energy collisional dissociation (HCD), CID is separately performed in an HCD cell, before the product ions are introduced into the Orbitrap, allowing the determination of their *m/z*’s. The main difference between CID and HCD is the amount of energy being delivered to the ions. For CID, it ranges in the dozens of eV, whereas HCD can go up to keV levels and result in more information-rich spectra.

The resulting fragmentation pattern depends on multiple factors, such as the use of CID or HCD, the peptide charge and the position of basic and bulky residues, including proline. B- and y-ions are designated based on where their charge is retained, either on the amino- or on the carboxy-terminal part of the peptide, respectively. Apart from CID and HCD, other fragmentation techniques exist that increase the internal energy level of molecules, such as via electron capture dissociation (ECD), electron transfer dissociation (ETD), or pulsed Q CID [223]. These methods generate distinct fragmentation patterns that need to be considered during peptide sequencing.

1.5.5 Database-Dependent Search

Theoretically, a resulting MS/MS spectrum can be manually annotated and the underlying sequence determined. However, with the thousands of spectra generated from a fast and high-resolution MS, this attempt is impractical in reality. As such, MS search tools and sophisticated statistical algorithms exist to assign MS/MS spectra to specific sequences [228, 229]. A database-dependent search is a category of MS search tools that requires a protein sequence database as input **(Figure 4)**. The protein sequence database routinely used for MS-based searches is derived from the Uniprot Knowledgebase (UniprotKB) [230]. This database is enriched with manual annotation and highly detailed functional information on proteins. Sequence assignment occurs in a relatively straightforward manner: the provided protein database is processed *in silico* to generate a list of peptides determined by user-given parameters, such as peptide length specifications and enzyme specificity, and the theoretical *m/z* of peptide precursor ions is calculated. Peptides are retained in the list when the theoretical precursor *m/z* matches the experimental *m/z*, within a user-defined mass tolerance. Following this, the tool generates ideal fragment ions (depending on the fragmentation strategy) along with *m/z*, which are then compared to the experimentally acquired tandem MS.

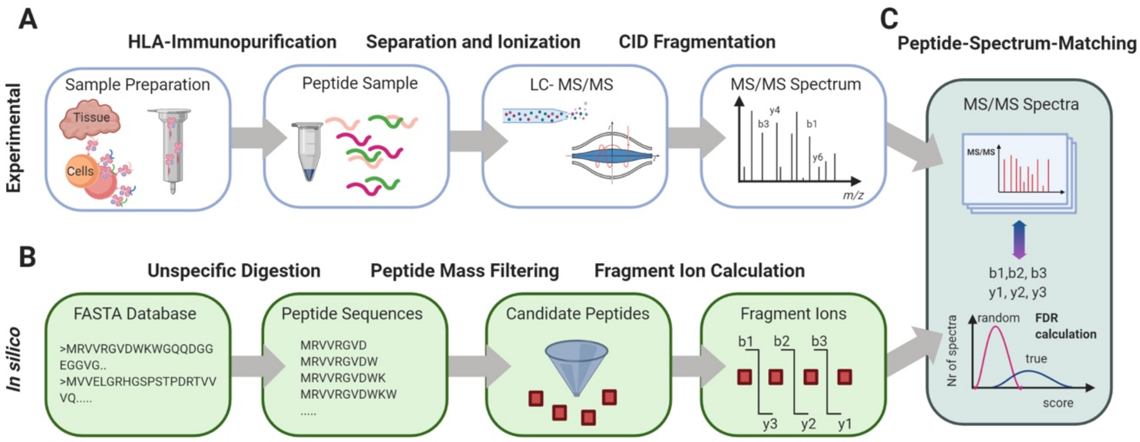


Figure 4 –Data acquisition and database-dependent search with immunopeptidomics data. ***(A)*** *An experiment in immunopeptidomics starts with the immunoaffinity purification of HLA complexes and HLA peptide extraction from tissue specimens or cell lines. The peptides are injected into the LC-MS, and can be fragmented, leading to the generation of MS/MS spectra.* ***(B)*** *A MS/MS search tool is used for peptide identification, where a protein sequence reference database is provided as input. For immunopeptidomics, unspecific in silico digestion is enabled, and peptides are filtered based on the m/z of the precursor ions. Theoretical fragment ion masses are calculated for every peptide.* ***(C)*** *A score that determines the similarity between the theoretical and experimental spectra is calculated, and statistical algorithms are applied to identify the best-matching peptide sequence and adjust the false discovery rate. Inspired by Eng et al., Molecular and Cellular Proteomics, 2011 [229].*

DATABASE SEARCH SCORING ALGORITHMS

There are several database search scoring algorithms that have been developed, which essentially score peptides based on the similarity of the *in silico* generated and experimentally measured MS/MS spectra. These algorithms can assess similarity in various ways [231, 232]. For example, cross-correlations are used in the Comet tool, where a similarity score is computed for each peptide through pattern recognition of theoretical versus experimental MS/MS spectra [233]. In contrast, the tools for Mascot, or the Andromeda engine within MaxQuant [234], calculate the probability of the observed number of experimental versus theoretical matches of fragment masses occurring by chance.

Ultimately, many sequences within a given database could theoretically match a spectrum by chance. Therefore, it is crucial to assign any match with a statistical significance, in order to have confidence that the matching did not occur at random. For this purpose, statistical methods exist that can estimate the level of false discoveries in a dataset. One approach is to utilize decoy databases, such as the original protein database in the reverse or scrambled form, together with the actual peptide sequences, in order to estimate the level of false discoveries [235]. This estimation is then used as a guideline to further filter the sequences based on certain criteria, for example, score threshold, to retain only those in the final dataset that are below the user-specified false discovery rate (FDR).

Additionally, delta scores are computed, where the highest ranked peptide that matched a specific spectrum is compared to a pool of peptides which also match the same spectrum, albeit with lower scores [229]. The differences in the score distributions provide confidence on whether the highest-ranked peptide was correctly, or incorrectly, determined. Therefore, it is crucial that a sufficient number of peptides is provided from the database and pass the search filter criteria, so that these comparisons can occur. Thus, if a database of insufficient size is used, the statistical significance of these calculations may be negatively impacted.

DE NOVO SEQUENCING

In addition to database-dependent searches and spectral library analyses (see Section 1.5.6), MS results can be analyzed by *de novo* sequencing algorithms [236, 237]. This method does not require prior input from protein sequence databases or libraries, and sequences are reconstructed directly from the MS/MS spectra. Therefore, *de novo* sequencing is thought to be particularly useful to increase the range of peptides that can be identified in immunopeptidomics samples [237, 238]. However, as this method relies upon low interferences, the quality of MS spectra provided must be very high, and thus there is currently limited adoption of *de novo* sequencing. Furthermore, in order to eventually define the source protein, the peptide sequence is then mapped back to a reference.

1.5.6 Mass Spectrometry Acquisition Techniques For Immunopeptidomics

DATA DEPENDENT ACQUISITION

There are different acquisition techniques with which a peptide sample can be measured by MS. The data-dependent-acquisition (DDA) method selects the top most abundant ions (often between top10-top20) for fragmentation, with dynamic exclusion for a few seconds to avoid the oversampling of abundant ions **(Figure 5)** [205]. This method accurately reveals a snapshot of complex protein/peptide samples, called shotgun-proteomics or –immunopeptidomics, and is typically analyzed via database-dependent searches. A drawback of DDA is that it suffers from the under-sampling of low abundance ions and low reproducibility between sample injections. The method can therefore lack "completeness", especially when analyses of specific pathways are of interest [205]. That being said, features such as sensitivity, mass accuracy and resolution (in TOF and Orbitrap mass analyzers) are being continuously improved and should lead to more robust results from shotgun methodologies. Overall, the DDA method is the most commonly used approach by the immunopeptidomics community, and represents an indispensable method for generating novel biological discoveries [209, 239, 240].

Data independent acquisition

An alternative approach, data independent acquisition (DIA), isolates and fragments all precursor ions in an unbiased manner within shifting and overlapping isolation windows [205, 241, 242]. Here, highly complex spectra are generated that incorporate signals from multiple peptides, and thus these results are not easily compatible with database-dependent searches. As such, tools that allow sequence identification from DIA rely on the prior construction of peptide spectral libraries generated initially through extensive DDA analyses. When using the DIA approach, the number of peptide identifications can be boosted by many factors compared to DDA. This is especially beneficial for low abundance peptides, and greatly enhances peptide reproducibility and quantification across multiple samples. Over a wide range of samples, this approach can advance the monitoring of both biomarkers for treatment stratification and predictors to specific responses. DIA has been recently optimized for immunopeptidomics, facilitating comparative analyses across biological samples, and importantly, reducing the need for large amounts of sample input due to higher sensitivity [205, 241].

While many spectral libraries are being made available from proteomics studies (www.swathatlas.org) [243], there still remains a significant need to generate large and high quality libraries for HLA peptides. A community-driven endeavor is currently being undertaken to achieve this, and should reflect the immense diversity of peptides bound by the different HLA allotypes [200, 215]. Importantly, a tissue-based atlas of the healthy murine MHC class I immunopeptidome has demonstrated the possibility of generating spectral libraries at a large scale [244]. Needless to say, this method presents various technical difficulties, and relies upon both comprehensive expertise and advanced data processing methodologies.

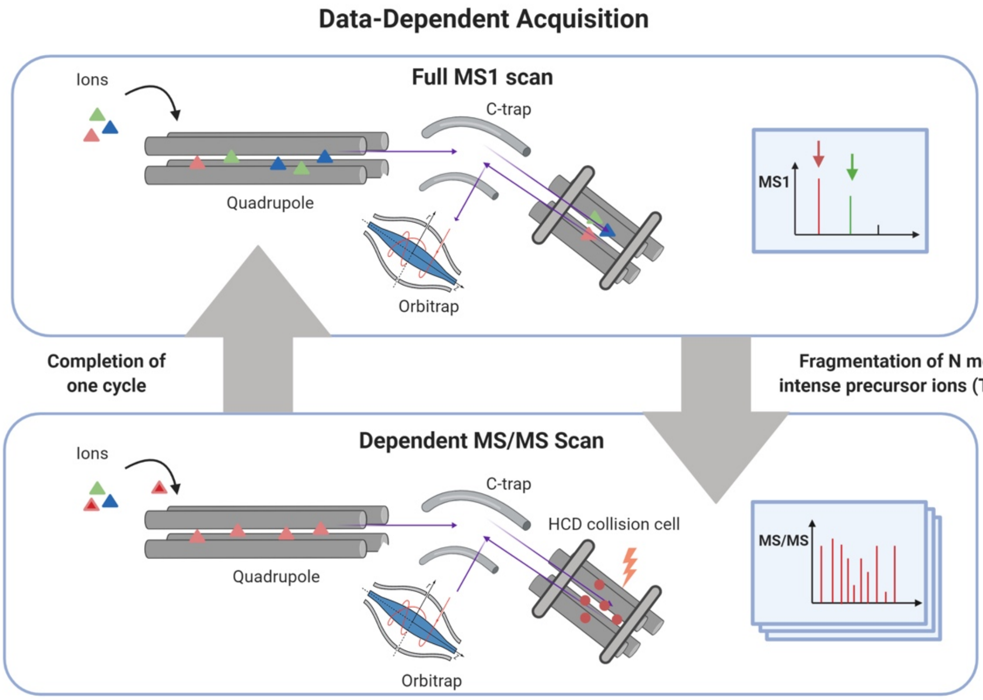


Figure 5 – In a ***data dependent acquisition*** *mode using the Q Exactive HF-X instrument, ions are filtered through the quadrupole, and a full MS1 scan is performed to deduce the m/z of precursor ions. Then, the most abundant N precursor ions (usually N=10-20) are selected for further fragmentation through dependent MS/MS scans, completing one cycle. The acquisition cycles are repeated throughout the LC gradient. MS/MS database search tools can be used to identify the sequences from the generated MS/MS spectra.*

Targeted Tandem mass spectrometry

Aside from DDA and DIA, selected or parallel reaction monitoring (SRM or PRM, respectively) are hypothesis-driven methodologies to selectively target and track a defined set of ions in complex peptide mixtures. These methods can reproducibly profile and quantify desired peptides over a period of MS acquisitions [245]. SRM is a more traditional approach in targeted proteomics, and is employed in triple quadrupole MS. Here, both the precursor, and the product ions are pre-defined for targeting. In PRM, only the precursor ion needs to be defined in advance, and all ensuing product ions are measured on hybrid quadrupole-Orbitrap instruments [246, 247]. PRM has now become the state-of-the-art approach for targeted proteomics, and offers several advantages over SRM, such as high sensitivity, the unambiguous confirmation of target sequences, and in ease in experimental execution as the prior determination of fragment *m/z*'s is not required **(Figure 6)**. Furthermore, the high accuracy of Orbitrap MS instrumentations enables PRM techniques to be less prone to interference from background noise, and thereby can better distinguish peaks, especially for low abundance ions.

Targeted MS techniques have been applied in immunopeptidomics in various ways and are considered the most robust method to validate a peptide sequence [78, 217, 238, 248-250]. For example, targeted MS is used in sequence-specific validation, in quantifying the abundance and copy number of specific antigens on cells' surfaces over time, and in monitoring peptide losses during immunoaffinity procedures. Importantly, while some antigens may be below the identifiable threshold in discovery approaches, targeted MS techniques can identify pre-defined predicted neoantigens that are of low abundance [251].

1.5.7 MS-Based Proteomics In Combination With Immunopeptidomics

In MS-based proteomics, different strategies, such as top-down, or bottom-up approaches are employed to identify and deduce the abundance of proteins. The most commonly used technique is the bottom-up approach, which first requires the enzymatic digestion of proteins into peptide fragments, usually using trypsin. The digested peptides are then analyzed by MS, and protein inference is performed post-measurement. Specifically, significant peptide-spectrum-matches can provide support for the presence of the same source protein. Generally, accurate protein inference takes into consideration situations where a peptide matches to different proteins, the occurrence of single-peptide-evidence, and incomplete tryptic signatures. In top-down approaches, intact proteins are measured and analyzed without prior digestion in specific mass spectrometers that allow the direct measurement of large molecules.

Aside from protein identification, MS and related technologies allow the accurate quantification of the proteome. This particularly powerful approach enables the analysis of proteins in different conditions, such as in disease or upon stimuli. Often, for relative abundance comparisons, proteins from different samples are distinctly labelled and simultaneously analyzed by MS, for example, through stable isotope labeling of amino acids. On the other hand, search tools for proteomics can include label-free analysis options, enabling larger-scale studies. Importantly, the high resolution, as well as high mass accuracy, of MS measurements are required for accurate label-free quantification [252]. A detailed explanation of the label-free quantification technique applied in MaxQuant is given by Cox et al., where the intensity-based precursor signals are used for quantification [253]. These steady advances in MS have enabled vast steps in proteomics research [254]. Including, but not limited to, generating the first human draft of the proteome [255], mapping the diverse roles of proteins in disease states [256], elucidating organelle-specific functions [257], studying protein folding [258], and unraveling the dynamic roles of protein PTMs [259].

By combining proteomics with immunopeptidomics, it becomes possible to explore the rules that govern peptide selection and presentation from cellular proteins, i.e. to study how the proteome can shape the immunopeptidome. However, studies incorporating proteomics to support immunopeptidomics findings have reported contradicting results to date [96, 260-265]. Factors such as protein abundance, turnover, degradation and translation are recognized to dictate the presented immunopeptidome, but lack clear consensus on their specific contributions. Interestingly, several targeted studies have shown that abundant proteins may not generate HLA peptides, while lower expressed proteins might [81, 266]. Therefore, the rules that dictate proteome sampling for HLA presentation remains to be properly dissected. Evidently, such insights will advance our understanding of the sampling of mutated and pathogen-derived proteins, and novel translation products.

1.5.8 Additional Considerations For MS-based Immunopeptidomics

As MS-based technologies were originally optimized on proteomics samples, several considerations arise when applying the technology for immunopeptidomics, as samples inherently differ in terms of their nature, required preparation and methods for identification [238]. For example, HLA peptides are typically shorter, often singly charged, and have a lower prevalence for basic residues. These features usually cause HLA peptide fragmentation to be more challenging than enzymatically digested (often tryptic) peptides. This issue results in a lower percentage of spectra being identified when compared to proteomics experiments, creating a bottleneck that needs to be considered in any antigen discovery endeavor. Notably, while the use of HCD is compatible with most MS/MS database search algorithms, other (complementary) fragmentation techniques performed on HLA peptides have been reported to significantly improve identification rates in immunopeptidomics. Specifically, the use of a combination of electron-transfer/higher-energy CID generated dual-fragment ion series and resulted in highly information-rich spectra [267, 268]. This was shown to improve the identification of HLA peptides by approximately three-fold, and led to the superior localization of HLA-associated PTMs. However, due to the complex nature of these alternative techniques, the analyses of their different fragmentation patterns require more advanced computational infrastructure.

Furthermore, compared to proteomics, immunopeptidomics generally utilizes larger protein reference sources for database-dependent searches. Notably, the increased database size has implications on the confidence of the statistical calculations used in MS/MS search tools for peptide identification [200, 205, 238]. In proteomics searches, an enzyme specificity is set, resulting in a more restricted list of *in silico* digested peptides. In contrast, for immunopeptidomics, enzyme specificity cannot be set and thus a significantly larger list of theoretical peptides is generated for comparison to experimental data **(Figure 4)**. Additionally, the identification of HLA peptides is more challenging due to the high similarity of peptide sequences, when compared to enzymatically digested samples. For statistical calculations, decoy-target searches are routinely applied [269, 270]. However, due to the larger database size, the probability of a randomly assigned false match increases, and can ultimately lead to an underestimation of the FDR [271-273].

These important considerations in MS-based immunopeptidomics must be taken into account as the search for mutated neoantigens continues, and remain especially critical when exploring novel sources of antigens.

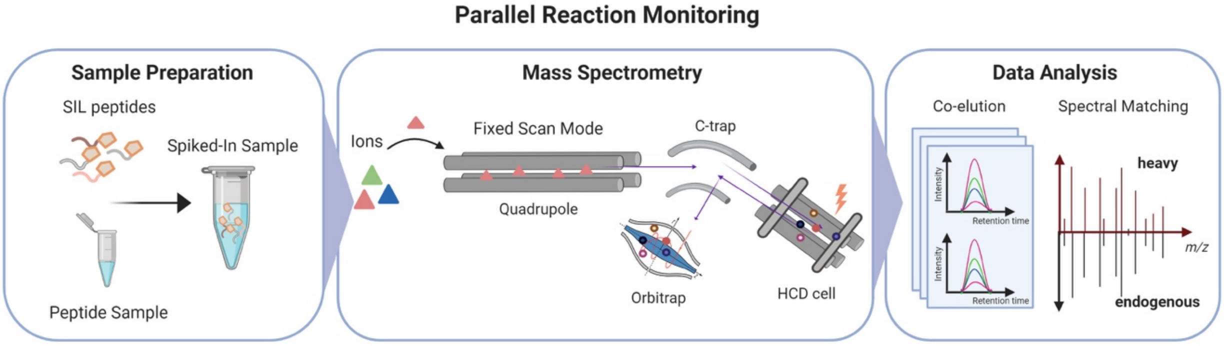


Figure 6 – Parallel reaction monitoring. *Selected peptides are synthesized in their heavy-labelled forms (SIL peptides) and spiked back into the original sample. The mass spectrometer selects precursor ions of interest in a fixed scan mode, fragment these, and monitors all transitions. To confirm the existence of the endogenous peptide, co-elution of the «heavy» and «endogenous» peptides must be found. Further, the fragmentation patterns should be highly similar, except for the characteristic mass shift derived from the stable isotope labelling. Abbreviations are as follows: SIL: stable isotope labelled*

1.6 Proteogenomics

Identifying novel, yet un-annotated, antigens from biological samples by MS is not a trivial task. For antigen discovery, one key fact to be considered in any MS database-dependent search is that only peptide sequences already provided in the database can be identified. Importantly, a routinely used protein sequence database, such as UniprotKB, potentially lacks many peptide sequences, for example due to mis- or non-annotation. Therefore, novel peptides would never be discovered in this way. That being said, in recent years, thanks to the acceleration of advances in high-throughput DNA-, RNA- and Ribo-Seq techniques, a new field, termed proteogenomics, has emerged [274]. Proteogenomics essentially involves the integration of either MS-based proteomics or immunopeptidomics data with information from genomics and/or transcriptomics. The aim of proteogenomics is to expand the interpretation of MS-based data and provide protein evidence from gene-level studies. In this last section, the background of proteogenomics is discussed, followed by a summary of MS-based discoveries in the area of non-canonical peptides. Lastly, the current challenges surrounding the discovery of non-canonical peptides via proteogenomics are summarized.

1.6.1 Proteogenomics Background

Since its initial introduction in 2004 [274, 275], proteogenomics approaches have led to various scientific discoveries. For example, using database-dependent searches, mutated neoantigens were found by including information on nonsynonymous somatic mutations derived from WES [209], and proteasome-generated spliced peptides from the inclusion of predicted splice variants [276].

However, the challenges associated with the often immense size of the non-canonical space has to be tackled in any proteogenomics approach. In an attempt to specify the search space for MS-based analyses, RNA expression data is often used to predict the potentially translated peptide products *in silico*. In a regular RNA-Seq experiment, the original mRNA strand information is lost. Strand-specific RNA-Seq circumvents this issue and allows more accurate transcript expression analysis and the determination of read direction. In this manner, using strand-specific RNA-Seq reduces 6-frame translation on both strands to 3-frame translation on one strand [277, 278]. Despite this, incorporating novel sequences using 3- or 6-frame translations to existing protein sequences results in a significant enlargement of the search space [274, 279]. Thus, data processing strategies and statistical analyses for proteogenomics have become challenging, with the associated risks discussed in Section 1.6.3. Depending on the biological question in hand, researchers are forced to evaluate a trade-off between database completeness, and increased search time and higher FDRs. Therefore, novel peptides should always be treated carefully, and compared to other reference databases to test whether they fit another “known” sequence, or contaminant. These peptides should be further validated by complementary experimental, analytical and targeted MS-based approaches.

1.6.2 MS-Based Non-Canonical Peptide Discovery

In this section, a number of publications are discussed that have applied in-house bioinformatic workflows to generate customized databases and search MS data for the identification and validation of (tumor-specific) non-canonical peptides **(Figure 7)**. This is in contrast to Section 1.3.3, where non-canonical antigens were first introduced and the majority found in single case targeted, and reductionist studies.

Conceptual Non-Canonical Peptide Discovery

With the advent of "omics" and bioinformatic developments, several groups have identified alternative translation products at a larger scale, potentially increasing the pool of non-canonical antigens to explore. For example, indirect evidence showing the potential of generating HLA peptides include investigations into short ORFs, which have previously been largely disregarded due to difficulties in their annotation. Thus far, hundreds of potentially coding short ORFs, including those with alternative start codons derived from coding genes and non-coding RNAs, have been identified with computational, MS-based proteomics and RNA-Seq approaches [280-283].

Conversely, direct evidence of non-canonical peptides at a larger-scale was reported, such as proteasome-generated spliced peptides in *cis*, where distant peptide fragments from the same protein are linked [276]. These peptides were identified by MS using a customized database of predicted spliced protein sequences concatenated to the normal proteome. Using lymphoblastoid and lymphoid cell lines, Liepe et al. reported approximately 30% of the immunopeptidome to be of spliced origin. *Trans* spliced peptides, i.e. linked peptide fragments from different proteins, were interrogated by the group of Purcell, using mono-allelic cell lines and an extensive bioinformatics workflow employing a combination of *de novo* sequencing and library (re)-searches [284]. The researchers found the number of *trans* spliced peptides to be similar to the levels reported previously for *cis* spliced peptides. This was challenged shortly after by Mylonas et al., by discerning that the identified *cis* spliced peptides had low HLA binding affinities and poor binding motifs [271]. Therefore, an alternative workflow was recommended employing *de novo* sequencing and multiple search tools, at 1% FDR. Consequently, the proposed number of *cis*-proteasome-generated spliced peptides were estimated to make up at most 2-6% of the entire immunopeptidome.

Moreover, B cell lines were used to identify cryptic peptides derived from non-canonical reading frames [285]. The protein sequence database used was constructed from 6-frame translation of non-coding regions of sample-matched RNA-Seq data. Notably, a 9% FDR was reported, and 10% of the entire B cell immunopeptidome was estimated to be of non-canonical origin. Although an intriguing finding, the significance of the peptides' contribution to the immunopeptidome remains unclear, primarily due to the high FDR reported.

Furthermore, ribosome profiling (Ribo-Seq), which pinpoint transcript regions of active translation through sequencing of ribosome-protected fragments, has shown that pervasive translation occurs outside of protein-coding genes [286-288]. Going beyond protein sequence databases inferred from RNA-Seq data, databases assembled via Ribo-Seq information for MS searches have been used for benchmarking the accuracy of Ribo-Seq protocols. Specifically, it was shown that an advanced Ribo-Seq pipeline enabled the identification of cryptic translation events, which were validated by MS in fibroblast samples [287].

Non-Canonical Peptide Discovery in Disease States

The variety of conceptual studies indicating the existence of non-canonical peptides at a systems-level has led to a number of researchers evaluating the relevance of non-canonical peptides in disease states, especially in cancer [145]. This concept is of particular interest, as non-canonical antigens could arise from the tumor-related aberrant translation of non-coding regions and be shared across patients. This is in contrast to mutated neoantigens, which are mostly private and thus require individual interrogation per patient. Thus, the potential pool of shared tumor-specific non-canonical peptides could outcompete mutated neoantigens for off-the-shelf cancer treatments [289]. Below, a selection of studies identifying disease-specific non-canonical peptides are outlined.

Through a combination of an *in silico* approach supported with MS validation, Smart et al. reported the presence of intron-retained neo-epitopes in cancer cell lines and across patient datasets [290]. The customized database included the prediction of intron-retained epitopes from RNA-Seq data, which were further filtered for patient HLA restrictions prior to MS-based search.

Shortly after, Laumont et al. utilized two murine cancer cell lines CT26 and EL4 in mouse models to validate the existence and relevance of tumor-specific non-canonical peptides [250]. The researchers built the customized database by only taking into account sequences that were predicted MHC binders. Furthermore, sample specific thresholds were applied, rather than a global FDR. Tumor specificity was set by removing normal RNA-Seq reads of matched murine thymic epithelial cells (TECs) from the tumor transcriptome. Ultimately, tumor-specific non-canonical peptides were shown to elicit anti-tumor responses in mice after vaccination.

Furthermore, researchers have recently focused on the discovery of RNA edited neoepitopes [291]. For this purpose, a computational pipeline was developed to annotate editing sites and concatenate the RNA editome to the canonical protein sequence database. MS/MS spectra from human tumors were used to screen for the existence of these peptides with a non-personalized database, and five were found to be derived from RNA editing processes. RNA edited neoepitope specific CD8$^+$ T cells were present in human tumors, and initiated killing in tumor cells presenting these epitopes.

Finally, TE-derived antigen discovery was tackled by the group of Chen-Harris through the robust annotation of TEs using RNA-Seq data and subsequent interrogation of MS data [292]. Initially, TE-derived transcription was found across TCGA samples, and validated in glioblastoma cells. When DNA demethylating agents were applied to a glioblastoma cell line, TE-derived peptides were found to be upregulated. Notably, the protein sequence database for MS search was tailored to only include overexpressed TE elements upon treatment, along with 6-frame translation.

1.6.3 Current Challenges In Proteogenomics For Non-Canonical Peptides

Collectively, the findings above provide direct and indirect evidence that alternative peptides represent untapped sources of immunogenic antigens for cancer immunotherapy. However, as 75% of the genome is transcribed and could therefore theoretically be translated, the resulting search space is incredibly vast [293]. This presents a significant and overarching challenge when identifying non-canonical peptides. While the strategies described in Section 1.3.4 for peptide identification, prioritization and evaluation have been applied specifically to tumor-associated and mutated antigens, these alone are currently not sufficient to determine the relevance of the large pool of potential non-canonical peptides.

Although research into more robust experimental and computational developments is needed, gene expression and MS-based analyses offer a solution to narrow down the peptides being considered. However, there remain central issues when undertaking efforts to map the non-canonical immunopeptidome by MS. First and foremost, there are FDR issues inherent to MS-based searches that need to be thoroughly evaluated [270, 274, 279, 294]. This is especially true when 6-frame translations of RNA species are used for the MS search, which significantly increases the size of the database. Furthermore, there are large differences in the type of non-canonical peptides being explored, with very little consensus between studies. Moreover, pre-filtering options prior to MS-based searches, such as restricting the peptides based on HLA binding prediction or tumor specificity, could skew any downstream results obtained, and the basic biological significance of non-canonical presentation would remain understudied. Additionally, while personalized approaches can potentially narrow down the novel targets specific to a patient, the existence of these peptides are rarely experimentally validated. Validation strategies, preferably directly on patient tumor tissue, are key in supporting the existence of any novel peptide. Ultimately, the immunogenicity of any non-canonical peptide should ideally be tested on autologous immune cells to determine its clinical relevance.

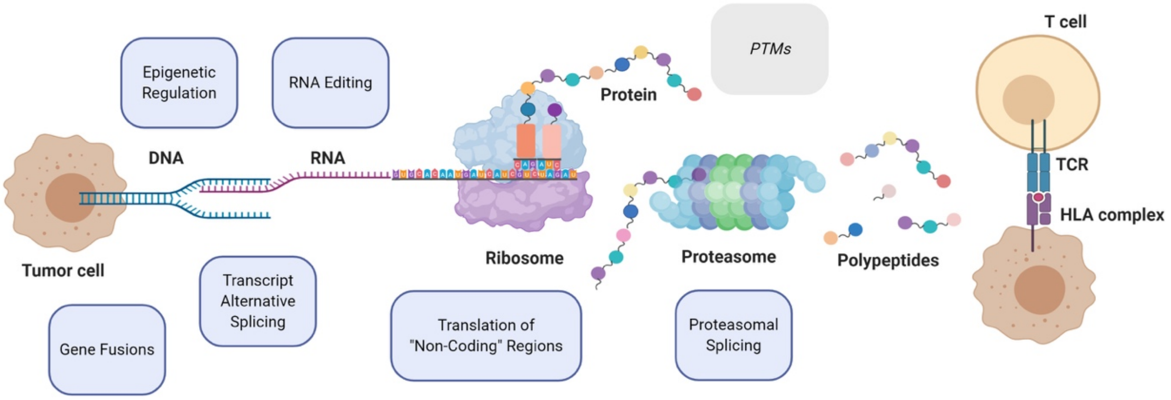


***Figure 7 – The processes that potentially generate non-canonical peptides** are illustrated in blue, from the genomic level through to proteasomal splicing. For completeness, PTMs are included in the diagram and shown in grey, and are considered as canonical in this thesis.*

Chapter 2

Aims and Objectives

Chapter 2 AIMS AND OBJECTIVES

Despite the steady advance in the field of cancer immunotherapy, the search for the "ideal" tumor antigen that can be exploited in combination with other immunotherapy modalities remains a key challenge. Many researchers are integrating a variety of sequencing technologies and rapidly-evolving neoantigen prediction tools in order to pinpoint targetable antigens. In comparison, MS-based immunopeptidomics allows the direct characterization of the tumor antigenic repertoire, and is thus highly attractive for the screening of presented tumor antigens. However, a significant number of issues need to be addressed before the research can be effectively translated to the clinic. Therefore, the focus of this thesis is to offer solutions to the gaps that are hindering the advancements of immunopeptidomics. Specifically, this endeavor is performed with two overarching aims, as illustrated in detail below.

AN IMMUNOPEPTIDOMICS PLATFORM FOR BASIC AND TRANSLATIONAL APPLICATIONS

Existing immunopeptidomics analysis pipelines continue to lack robustness and standardization. This translational gap was discussed among scientific representatives in the first international Human Immunopeptidome Project workshop in 2017, co-organized by the Bassani-Sternberg laboratory [215]. Based on the community's opinion, one of the major caveats in current immunopeptidomics workflows is the HLA immunoaffinity purification. This is often impeded by low reproducibility and sample-throughput, uncertain peptide yields, and the dependency on large sample amounts for downstream analysis.

Therefore, the initial aim of this thesis is focused on designing and optimizing a novel HLA immunoaffinity purification system that enables the streamlined extraction of HLA-I and –II peptides. When compared to the attributes of existing systems, the goal is to improve on multiple features ranging from increased speed, sensitivity, reproducibility and scalability, while systematically validating these improvements using patient-derived cell lines and tissue samples. The applicability of this method for drug screening will be assessed at the peptide level upon treatment with inflammatory agents. A pipeline that addresses the current challenges should help accelerate MS-based immunopeptidomics implementations into clinical settings, and provide a reliable framework to explore further biological topics, such as the identification of non-canonical cancer epitopes.

DISCOVERING NON-CANONICAL PEPTIDES IN TUMOR IMMUNOPEPTIDOMES

Over the last 30 years, targeted molecular approaches have led to the identification of several immunogenic epitopes derived from alternative ORFs, intronic regions, and retroviral elements, with seminal studies published by the research groups of both Boon and Rosenberg [146, 148-150, 295-298]. These findings have motivated researchers to exploit alternative antigens for cancer therapies. With the advent of NGS, several studies have started large-scale investigations into the existence and relevance of HLA peptides derived from presumed non-coding genomic regions, especially in the context of cancer [250, 290, 292]. However, the presented workflows to perform this endeavor vary greatly between studies. Principally, both the computational and validation approaches typically employed require optimization to confidently evaluate the clinical significance of non-canonical peptides.

As such, the core aim of this thesis is to develop a state-of-the-art integrated immunopeptidomics and proteogenomics framework, robustly identifying and characterizing presented non-canonical HLA peptides (noncHLAp) in patient-derived melanoma cell lines and lung cancer tissue samples. Patient-specific non-canonical peptides will be identified by combining immunopeptidomics with genomics, transcriptomics and translatomics analyses. Further, a MS-based computational module will be developed to control for the identification error of non-canonical peptides and limit false positives due to large search spaces. Following this, identifications can be validated through complementary analytical and targeted MS-based experimental methods. The clinical relevance of the non-canonical peptides will be investigated by examining their tumor specificity via comparison against publicly available healthy tissue RNA-Seq data. Moreover, the potential of re-identifying shared actionable antigens, beneficial for faster "off-the-shelf" therapies, will be investigated both across the patient samples by targeted MS, and within a large in-house generated immunopeptidomics database. Finally, *in vitro* cellular assays will be performed to validate the immunogenicity of the identified non-canonical peptides.

Overall, the combined work presented in this thesis aims to contribute to the research field of personalized antigen discovery for cancer immunotherapy in two ways. First, by facilitating the implementation of immunopeptidomics in translational research by providing an improved step-by-step guide for HLA immunoaffinity purification. Second, by systematically assessing non-canonical peptide identification with a MS-based immunopeptidomics, proteogenomics and analytical approach, furthering the surge of interest in determining immunogenic tumor non-canonical peptides.

Chapter 3

Summary of Results

Manuscript 1

Manuscript 2

Chapter 3 SUMMARY OF RESULTS

Two central manuscripts present and discuss the results of my thesis work, and are summarized in this chapter. The first, Manuscript 1, highlights the immunopeptidomics platform, and the second, Manuscript 2, discusses the proteogenomics workflow for tumor non-canonical peptide identification. Below, the results are outlined for each manuscript, followed by copies of the published Manuscript 1 (Molecular and Cellular Proteomics, 2017 [299]) and accepted Manuscript 2 (Nature Communications, 2020). Figure references in this chapter refer to the original articles, and the supplementary tables and datasets for Manuscript 1 can be found online. Supplementary Information for each of the manuscripts are included in the Appendix.

3.1.1 Manuscript 1

MS-based immunopeptidomics is the only unbiased method allowing the interrogation of the repertoire of naturally presented HLA peptides and the most critical step in this approach is the sample preparation, as it determines the coverage and reproducibility. Commonly, immunoaffinity purification of HLA complexes has been performed with anti-HLA antibody-crosslinked beads in relatively large individual chromatography columns. Samples were lysed and incubated with these beads from several hours to overnight at 4°C. Thereafter, the beads were washed and the HLA complexes eluted. The HLA peptides were then separated from the HLA molecules by a molecular weight cut-off size filter and concentrated by applying once or twice C18-based reversed-phase extraction ([299], Table S5). In our experience, this method is both time-consuming and involves extensive sample handling. Moreover, it suffers from low-throughput issues and is composed of many steps that result in the significant loss of both quality and quantity of HLA peptides. These challenges represent severe bottlenecks for the implementation of immunopeptidomics in robust clinical antigen discovery applications.

A HIGH-THROUGHPUT PLATFORM FOR HLA IMMUNOAFFINITY PURIFICATION

With these existing issues in mind, a high-throughput platform for HLA immunoaffinity purification was designed, using 96-well plates operated with a customized positive pressure instrument (Fig. 1). With the implementation of this system, several technical refinements to the traditional protocol were achieved. First, due to the streamlined application of positive pressure and plate stacking, the sequential purification HLAIp and HLAIIp was enabled, and substantially reduced pipetting steps and the time of traditional HLA immunoaffinity purifications to just a few hours. Second, the column volume was reduced, due to the purification in 2 mL plate wells as opposed to >10 mL columns. This significantly reduced the amount of expensive material used, such as the antibody-crosslinked beads. Third, the plate format allows up to 96 samples to be processed simultaneously, thereby increasing the speed, as well as the reproducibility of HLA peptide extraction.

Multiple validations were performed to support the superiority of the presented purification framework and the quality of the acquired immunopeptidome. Through LC-MS-based analyses and from 21 simultaneously processed samples, approximately 50,000 unique HLAIp and HLAIIp derived from human B and T cell lines and meningioma tissue samples were obtained at a 1% FDR (Fig. 2A-B). *Bona fide* HLA peptide characteristics, such as their length properties (Fig. 2C-D) and HLA binding motifs (Supplemental Fig. S2) were confirmed, and a high intra- and inter-plate reproducibility was observed (Pearson correlation coefficients "r" ranging from 0.89 to 0.98) across biological and technical replicates (Fig. 3 and Supplemental Fig. S3). In addition, to evaluate both peptide recovery and the potential risk of carry-overs between the plate wells, 15 heavy-labelled synthetic peptides were spiked into the B cell line CD165 prior to the desalting step. No synthetic peptides were detected in neighboring samples, and all 15 heavy-labelled peptides, along with their endogenous counterparts, were re-identified in CD165 replicates (Supplemental Table S3 and S4).

Furthermore, considerable obstacles in immunopeptidomics stem from sensitivity issues, and therefore, large sample amounts are required which is often not feasible for precious clinical material. Cell number dilution experiments were performed to compare the amounts of HLA peptides obtained through our high-throughput extraction method. Significant improvement on the sensitivity of HLA peptide extraction (1,846 HLAIp and 2,633 HLAIIp from 10 million cells) was seen when compared to other methods (Supplemental Table S5), mainly due to the reduced column volume and sample handling (Fig. 2E-F).

High Reproducibility Facilitates The Exploration of the Drug-Modulated Peptidome

To test the robustness of our pipeline for label-free comparative and quantitative immunopeptidomics analyses, the IFNγ modulated peptidome on UWB.1 289 ovarian cancer cells was mapped, and the overall properties of the presented peptide repertoire upon stimulation was explored. The inflammatory cytokine IFNγ enhances surface presentation of HLA complexes that could lead to increased peptide presentation and a higher probability to discover immunogenic epitopes. High biological reproducibility within the control (r=0.97) and IFNγ replicates (r=0.95) was observed, and the increased repertoire and abundance in HLA presentation upon IFNγ treatment was noted (Fig. 4A-C, Supplemental Fig. S4). The results showed the differential presentation of peptides from source proteins that were upregulated, such as STAT1 and STAT2, WARs, and importantly, peptides derived from the immunoproteasome subunits (Fig. 4D). Furthermore, these observations were supported at the proteomics level (Fig. 4E).

As a result of the high quality immunopeptidomics repertoire obtained upon IFNγ treatment, further interesting and novel aspects were found, potentially related to the IFNγ-induced proteasome to immunoproteasome switch. Notably, the proteasome determines the C-terminal cleavage specificity of HLAIp. While the constitutive proteasome exhibits both tryptic and chymotryptic-like activities, the IFNγ-induced immunoproteasome demonstrates quantitatively higher chymotryptic-like activity [73]. In line with these findings, IFNγ led to enhanced presentation of peptides that bind HLA-B*07:02, displaying C-terminal chymotryptic-like amino acid specificity (Fig. 5A-C). The presentation of longer peptides harboring C-terminal chymotryptic-like amino acids were also induced, and showed a significant preference of these peptides over their shorter tryptic counterparts (p-value <0.01) (Fig. 5D-E). To this end, these comparative immunopeptidomics analyses allowed insights into both the well-known quantitative changes associated with the immunopeptidome, as well as the more sophisticated fine-tuning of peptide processing upon IFNγ treatment.

Finally, this optimized method formed the basis to the second part of this thesis work, which focuses on the exploration of the non-canonical space for novel alternative antigens.

3.1.2 Manuscript 2

Recent clinical data provide clear evidence that due to molecular genomic alterations, tumors express unique mutated antigens, the so-called neoantigens, that could play a key role in tumor immune recognition, and have been implicated in the therapeutic efficacy of immune checkpoint inhibitor antibodies [101]. The exon-coded proteome has provided limited opportunities to identify tumor neoantigens that are shared across patients, especially in tumors with low to moderate mutational burden. On the other hand, alterations within non-coding regions, if expressed, could represent a rich source of tumor-specific neoantigens. Such alternative non-canonical antigens are widely regarded to have relevance and potential to increase the breadth of targetable epitopes for cancer immunotherapy. Although this area of research has rapidly expanded over the last years, the systematic identification and evaluation of non-canonical peptides remains a challenge due to limitations in sensitivity and specificity, and thus their clinical relevance is still questionable.

An Integrated Proteogenomics Workflow For Non-Canonical Peptide Identification

In order to determine and characterize non-canonical peptide presentation in patient-derived melanoma cell lines and matched tumor/healthy lung tissues, a systems-level approach was adopted by integrating immunopeptidomics, genomics, transcriptomics and translatomics (Fig. 1a and Supplementary Data 1). The approach was specifically focused on non-canonical sources derived from (long) non-coding RNAs, pseudogenes, 5' and 3' untranslated regions, novel ORFs, and TEs. The personalized information of expressed non-canonical elements for every sample was derived from RNA-Seq data and subsequently *in silico* translated into three forward ORFs. Finally, Ribo-Seq was performed for the representative melanoma sample 0D5P.

A systematic challenge in proteogenomics studies, which utilize RNA-Seq information to generate translation products for MS-based approaches, stems from the use of large database search spaces [274]. This is especially true when all potential 6- or 3-frame translation products are constructed for transcripts that are presumed non-coding. Probabilistic-based algorithms for peptide-spectrum-matches in MS-based searches applied to large databases inherently cause a higher proportion of incorrect identifications. To overcome this FDR issue for large search spaces, a computational module, NewAnce (A new analytical approach for non-canonical element identification), was developed, which combines two MS-based search tool results (MaxQuant and Comet) (Fig. 1c and Supplementary Fig. 1a). Specifically, FDRs were calculated separately for proteome-derived HLA peptides and noncHLAp and only the consensus (intersection) peptide-spectrum-matches from both Comet and MaxQuant were retained. This strategy was applied to reliably identify noncHLAp from large sample-specific databases that included all potential 3-frame translations of expressed non-coding genes or TEs.

The accuracy of peptide identification was assessed for all samples using the following two methods. First, immunoaffinity purified peptides should be enriched with ligands that are predicted to bind the expressed HLA allotypes, and can be assessed using HLA binding prediction tools [175]. Typically, more than 90% of the identified proteome-derived HLA-I peptides (protHLAIp) are predicted as HLA binders, thus, similar levels for non-canonical HLA-I peptides (noncHLAIp) are expected. Second, as peptides elute from the analytical HPLC system with acetonitrile according to their hydrophobic properties, true peptide sequences should display a strong positive correlation between their observed retention time (RT) and the calculated hydrophobicity index (HI) [300].

When applying NewAnce, up to 148 novel noncHLAIp per individual sample were identified, with a combined total of 452 unique noncHLAIp (Supplementary Data 2 & 3). These peptides were evaluated based on both their HLA binding specificities and sequence-specific hydrophobicity, and demonstrated that the results derived from NewAnce were significantly superior over using either MS search tool alone (Fig. 2, Supplementary Fig. 2).

Non-Canonical Peptide Validation

MS-based targeted analyses were further performed to experimentally validate the proportion of true non-canonical peptide identifications in the representative melanoma sample 0D5P. For this purpose, peptide candidates of interest were synthesized in their heavy isotope-labelled forms through the incorporation of carbon-13 and nitrogen-15. Synthetic peptides were mixed, spiked into the sample of eluted HLA peptides from 0D5P cells and measured by targeted MS. Co-elution of heavy-labelled and endogenous peptides, along with nearly identical MS/MS fragmentation patterns, was used to confirm the existence of a novel sequence. With this technique, the targeted analyses were executed for the noncHLAIp identified with NewAnce, and directly compared to selected tumor-associated (protHLAIp) antigens. PRM for the different peptide classes showed that the rate of protHLAIp confirmation was superior to that of noncHLAIp (78.5% for TAAs versus 55.2% for lncRNAs and 27.7% for TEs) (Fig. 3a and Supplementary Data 6 & 7).

Furthermore, the Ribo-Seq method pinpoints actively translated regions, hence, it could circumvent the need to utilize RNA-Seq data and large databases for MS-based searches. An investigation into whether Ribo-Seq could detect active translation of noncHLAIp-derived ORFs (from RNA-Seq inferred immunopeptidomics data) showed that 22.2% of TE- and 21.3% of lncRNA-derived peptides were translated in the correct frame encoding the novel peptide sequences, compared to 100% of TAAs (Fig. 3b).

Insights Into Translation and Expression Levels Of Non-Canonical Elements

An independent MS-based discovery method was adopted using Ribo-Seq data with the representative sample 0D5P. Here, all actively translated ORFs were extracted, *in silico* translated, and included into the MS-based search. Using the smaller Ribo-Seq inferred database, in comparison to RNA-Seq inferred database, led to the conclusion that the overall immunopeptidome is better captured by the translatome than the transcriptome. With NewAnce, the Ribo-Seq inferred database led to a deeper coverage of the immunopeptidome, as well as the additional identification of novel HLAIp derived from currently un-annotated ORFs in coding genes (Fig. 4e-j).

As a limited proportion of noncHLAIp were re-confirmed by targeted MS, the expression patterns of source non-coding genes were investigated. Many of these source non-coding genes were observed to be lowly expressed, and treatment with either IFNγ or DAC did not significantly upregulate non-canonical peptide presentation (Fig. 4a-b, Supplementary Fig. 4m-r). However, many of the lowly expressed transcripts generated HLA peptides that were still confirmed by PRM (Fig. 4c-d). It was theorized that a subset of cells expressing the same gene at sufficient levels could enable their HLA presentation and detection by MS. Thus, single cell RNA sequencing (scRNA-Seq) was performed to gain deeper insights into the underlying profiles of non-coding gene expression. In this manner, a subset of cells was found in the melanoma cell line that co-expressed the non-coding source gene LINC00520, with marker genes ATP-binding cassette sub-family B member 5 (ABCB5), catenin beta 1 (CTNNB1) and microphthalmia-associated transcription factor (MITF) (Fig. 5e-h). The three latter genes are known for their cancer stem cell properties and are important drivers of melanoma progression [301-303]. In 0D5P, a novel downstream ORF of ABCB5 was detected by Ribo-Seq, resulting in the identification of a non-canonical ABCB5 peptide, KYKDRTNILF. Importantly, this ABCB5 noncHLAIp was found to be immunogenic, as assessed by IFNγ secretion in both autologous TILs and CD8$^+$ T cells from peripheral blood lymphocytes (Fig. 9a-c) following *in vitro* peptide stimulation. This finding has exciting implications, potentially allowing immune targeting of melanoma stem cell subpopulations to inhibit tumor progression.

Tumor Specificity of Non-Canonical Peptides And Re-Identification Across Patient Samples

At the clinical level, the interrogation of tumor specificity is key to limit toxicity and on-target off-tumor effects. Publicly available databases such as GTEx [304](The Genotype-Tissue Expression project; consisting of healthy human tissue RNA-Seq data) were employed to retrospectively evaluate tumor specificity of source non-coding genes. Of these, 23% were found to be specific to the tumor samples (Fig. 6). Additionally, the ideal situation was evaluated, where both tumor and healthy tissue from the same patient is available. In the case of the lung cancer/healthy tissue samples, nearly all non-canonical peptides found were patient-specific. However, these were not necessarily tumor-specific, as they were additionally identified by MS in the healthy tissue counterpart (Fig. 7a-b). This suggests that thorough investigation of non-canonical peptides across healthy tissues, ideally from the same patient, is necessary to ensure tumor specificity.

Finally, in order to develop rapid and “off-the-shelf” cancer treatment options there is profound interest in identifying immunogenic antigens that are shared among patients. Therefore, the prevalence of common noncHLAIp was investigated in the nine tumor samples. Twenty-seven shared peptides were detected by MS, and 15 of these events were re-confirmed by PRM, thereby validating for the first time that noncHLAIp can be shared across patient samples (Fig. 8a). Lastly, using an in-house curated immunopeptidomics MS database, ipMSDB [305], the prevalence of common noncHLAIp across a larger set of patients was assessed. A large-scale non-canonical presentation signature was obtained over the 91 biological cancer and 35 healthy tissues/cell line sources (Fig. 8b and Supplementary Data 8). Sixty of the re-identified tumor-specific noncHLAIp were limited to cancer samples in ipMSDB, and an enrichment trend of noncHLAIp was observed across cancer samples (Fig. 8c). Moreover, fourteen peptides were detected in at least one additional cancer sample, with the immunogenic non-canonical ABCB5 peptide shared across three melanoma samples in ipMSDB. Overall, these findings highlight the potential of non-canonical peptides to be shared across patients, and in a variety of different cancer types.

Chapter 4

Manuscript 1

High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferon γ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome

Chloe Chong, Fabio Marino, HuiSong Pak, Julien Racle, Roy T. Daniel, Markus Müller, David Gfeller, George Coukos, Michal Bassani-Sternberg

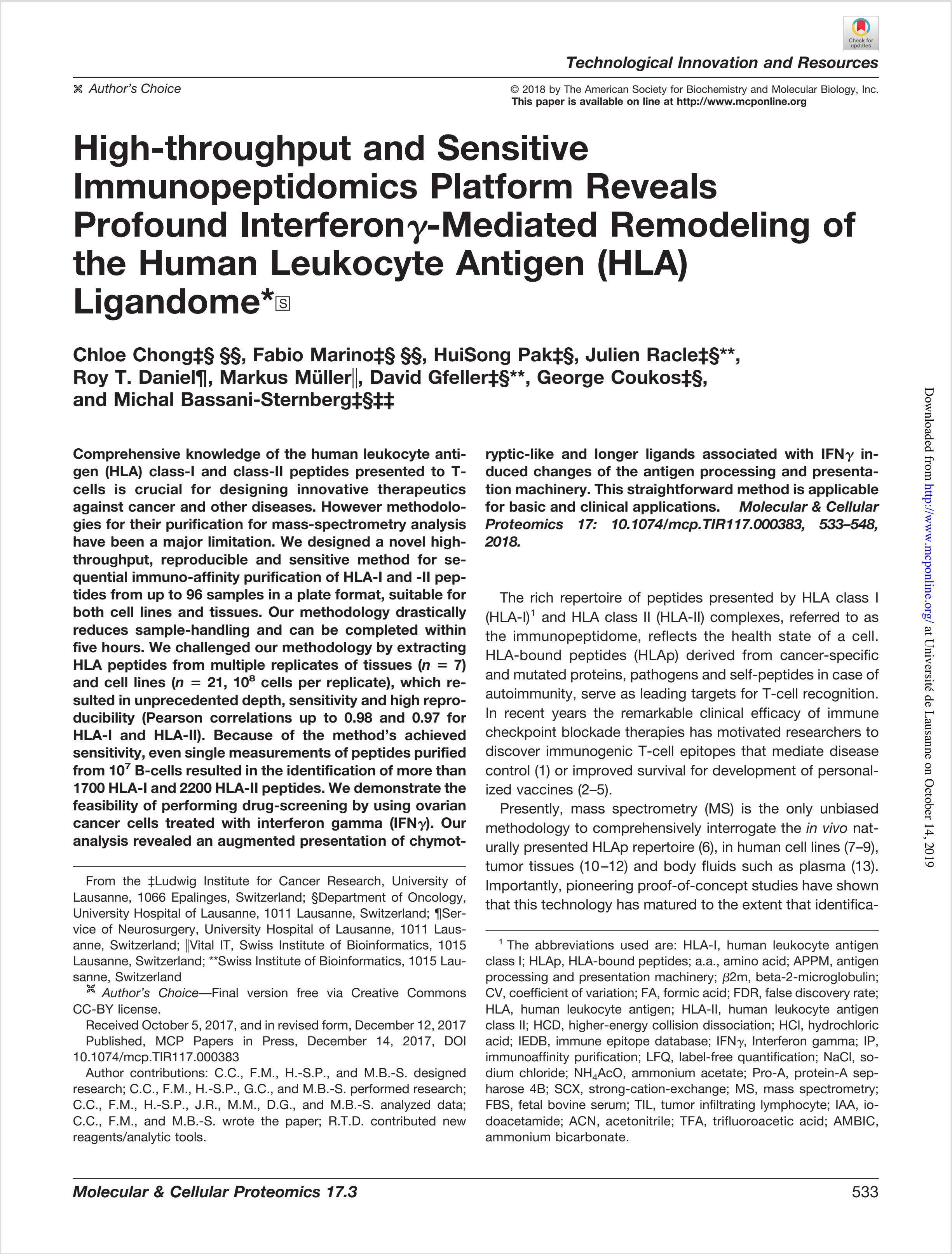
Molecular and Cellular Proteomics, 2017

Chapter 4 MANUSCRIPT 1

The method was published in Molecular and Cellular Proteomics in December 2017, where I am first co-author together with Dr. Fabio Marino, a former postdoc in the Bassani-Sternberg lab. The study was designed and results were interpreted together with Dr. Michal Bassani-Sternberg, while I and Dr. Marino, performed and analyzed all wet lab, cell culture and MS-based experiments, and wrote the manuscript together with Dr. Michal Bassani-Sternberg and Prof. George Coukos.

Furthermore, and as also outlined in Chapter 9, I co-developed a book chapter that describes our above step-by-step HLA purification protocol for basic and translational applications. Following this, in collaboration with Prof. David Gfeller's group at the Ludwig Institute for Cancer Research, Lausanne, the high quality immunopeptidomics datasets we generated with the above method were used to refine HLA-I and -II binding motifs and to improve the performance of HLA binding prediction algorithms. Lastly, I contributed to two additional studies by applying the described method for the discovery of immunogenic epitopes in ovarian cancer patient samples in collaboration with the group of Prof. Inge Marie Svane from Copenhagen and in a pre-clinical humanized mouse model (manuscript in preparation).

4.1 Published Manuscript



Technological Innovation and Resources

⌘ Author's Choice

© 2018 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome*Ⓢ

Chloe Chong‡§ §§, Fabio Marino‡§ §§, HuiSong Pak‡§, Julien Racle‡§**, Roy T. Daniel¶, Markus Müller‖, David Gfeller‡§**, George Coukos‡§, and Michal Bassani-Sternberg‡§‡‡

Comprehensive knowledge of the human leukocyte antigen (HLA) class-I and class-II peptides presented to T-cells is crucial for designing innovative therapeutics against cancer and other diseases. However methodologies for their purification for mass-spectrometry analysis have been a major limitation. We designed a novel high-throughput, reproducible and sensitive method for sequential immuno-affinity purification of HLA-I and -II peptides from up to 96 samples in a plate format, suitable for both cell lines and tissues. Our methodology drastically reduces sample-handling and can be completed within five hours. We challenged our methodology by extracting HLA peptides from multiple replicates of tissues (*n* = 7) and cell lines (*n* = 21, 10^8 cells per replicate), which resulted in unprecedented depth, sensitivity and high reproducibility (Pearson correlations up to 0.98 and 0.97 for HLA-I and HLA-II). Because of the method's achieved sensitivity, even single measurements of peptides purified from 10^7 B-cells resulted in the identification of more than 1700 HLA-I and 2200 HLA-II peptides. We demonstrate the feasibility of performing drug-screening by using ovarian cancer cells treated with interferon gamma (IFNγ). Our analysis revealed an augmented presentation of chymotryptic-like and longer ligands associated with IFNγ induced changes of the antigen processing and presentation machinery. This straightforward method is applicable for basic and clinical applications. ***Molecular & Cellular Proteomics 17: 10.1074/mcp.TIR117.000383, 533–548, 2018.***

The rich repertoire of peptides presented by HLA class I (HLA-I)[1] and HLA class II (HLA-II) complexes, referred to as the immunopeptidome, reflects the health state of a cell. HLA-bound peptides (HLAp) derived from cancer-specific and mutated proteins, pathogens and self-peptides in case of autoimmunity, serve as leading targets for T-cell recognition. In recent years the remarkable clinical efficacy of immune checkpoint blockade therapies has motivated researchers to discover immunogenic T-cell epitopes that mediate disease control (1) or improved survival for development of personalized vaccines (2–5).

Presently, mass spectrometry (MS) is the only unbiased methodology to comprehensively interrogate the *in vivo* naturally presented HLAp repertoire (6), in human cell lines (7–9), tumor tissues (10–12) and body fluids such as plasma (13). Importantly, pioneering proof-of-concept studies have shown that this technology has matured to the extent that identifica-

From the ‡Ludwig Institute for Cancer Research, University of Lausanne, 1066 Epalinges, Switzerland; §Department of Oncology, University Hospital of Lausanne, 1011 Lausanne, Switzerland; ¶Service of Neurosurgery, University Hospital of Lausanne, 1011 Lausanne, Switzerland; ‖Vital IT, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; **Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

⌘ *Author's Choice*—Final version free via Creative Commons CC-BY license.

Received October 5, 2017, and in revised form, December 12, 2017

Published, MCP Papers in Press, December 14, 2017, DOI 10.1074/mcp.TIR117.000383

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

[1] The abbreviations used are: HLA-I, human leukocyte antigen class I; HLAp, HLA-bound peptides; a.a., amino acid; APPM, antigen processing and presentation machinery; β2m, beta-2-microglobulin; CV, coefficient of variation; FA, formic acid; FDR, false discovery rate; HLA, human leukocyte antigen; HLA-II, human leukocyte antigen class II; HCD, higher-energy collision dissociation; HCl, hydrochloric acid; IEDB, immune epitope database; IFNγ, Interferon gamma; IP, immunoaffinity purification; LFQ, label-free quantification; NaCl, sodium chloride; NH_4AcO, ammonium acetate; Pro-A, protein-A sepharose 4B; SCX, strong-cation-exchange; MS, mass spectrometry; FBS, fetal bovine serum; TIL, tumor infiltrating lymphocyte; IAA, iodoacetamide; ACN, acetonitrile; TFA, trifluoroacetic acid; AMBIC, ammonium bicarbonate.

tion of clinically relevant mutated antigens in humans has become reality (14–17). In the field of immunology, this methodology is perceived as highly promising although not ready yet for its implementation in clinical settings because of its low sensitivity and robustness (2).

HLA-I and HLA-II complexes have key roles in modulation of immune responses and are distinguishable by the type of cells that express and recognize them and by the distinct biogenesis of the presented peptides (18). The repertoire of the presented immunopeptidome is constantly modulated by source protein expression levels, post translational modifications, and by several enzymes, chaperones and transporters that comprise the cellular antigen processing and presentation machinery (APPM). Cellular perturbation could affect this machinery at multiple levels, leading to the presentation of an altered peptidome. So far, the assessment of differential immunopeptidomics has been mainly unexplored because of technical limitations related to low throughput and reproducibility of existing methodologies (19, 20).

Immunopeptidomics is based on immunoaffinity purification (IP) of HLA complexes from mild detergent solubilized lysates, followed by extraction of the HLAp. The extracted peptides are then separated by chromatography and directly injected into a mass spectrometer. With the new generation of mass spectrometer instrumentations, thousands of HLAp can be readily identified per sample (7, 21).

The most critical step in the immunopeptidomics pipeline is the sample preparation as it determines the overall peptide yield and reproducibility. The entire workflow is laborious, typically spanning over 3 to 5 days, and is often limited to a few samples at a time (22). The above-mentioned bottlenecks pose severe restrictions on implementing this methodology for robust clinical applications and for LFQ comparative studies such as antigen presentation on infection (23), drug treatments or association of particular HLA alleles with autoimmunity (24).

In this work we set out to develop the first high-throughput method for IP of HLAp for MS-based immunopeptidomics, suitable for both basic and translational studies, where thousands of unique HLA-Ip and -IIp can be readily identified in a single IP procedure from cell lines and tissue samples. As IP of clinically relevant samples are often hindered by scarcely available amounts (12, 25–27), we decided to challenge the sensitivity of our platform by immunopurifying HLAp from as low as 10^7 B-cells. Furthermore, we also demonstrated the feasibility of performing comparative screening using an ovarian cancer cell line treated with the pro-inflammatory cytokine IFNγ. IFNγ is a well-known master regulator of immune modulation that up-regulates antigen presentation on target cells (28). Here, for the first time, we captured IFNγ-mediated modulation of specific components of the APPM which resulted in qualitative and quantitative alterations of the presented HLAp repertoire. Specifically, we discovered an enhanced presentation of chymotryptic-like ligands, as well as longer ligands deriving from nested sets on IFNγ treatment.

EXPERIMENTAL PROCEDURES

Cell Lines—EBV-transformed human B-cell lines JY (ATCC® 77442™, Manassas, Virginia), CD165, PD42, CM467, RA957 (a gift from Pedro Romero, Ludwig Cancer Research Lausanne) were maintained in RPMI 1640 + GlutaMAX medium (Life Technologies, Carlsbad, CA) supplemented with 10% heat-inactivated fetal bovine serum (FBS) (Dominique Dutscher, Brumath, France) and 1% Penicillin/Streptomycin Solution (BioConcept, San Diego, CA). UWB.1 289 ovarian carcinoma cells (ATCC® CRL-2945™) were maintained in a 1:1 mix of HuMEC Ready medium (Thermo Fisher Scientific, Waltham, MA) supplemented with HuMEC Supplement Kit (Thermo Fisher Scientific) and RPMI 1640 + GlutaMAX medium, with addition of 1% Penicillin/Streptomycin Solution and 3% heat-inactivated FBS.

Cells were grown to the required cell amount, collected by centrifugation at 1200 rpm for 5 min, washed twice with ice cold PBS and stored as dry cell pellets at −20 °C until use. For the *in vitro* treatment of UWB. 1 289 cells with human IFNγ (Miltenyl Biotec, Bergisch Gladbach, Germany), cells were grown to 1.5×10^8 in quadruplicates both for control and treatment. For treatment, cells were exposed to 100 IU/ml IFNγ for 24 h, detached with Accutase (Thermo Fisher Scientific), counted and washed twice with cold PBS before storage at −20 °C.

All cells were tested negative for mycoplasma contamination. High resolution 4-digit HLA-I and HLA-II typing was performed for all cell lines at the Laboratory of Diagnostics, Service of Immunology and Allergy, CHUV, Lausanne and provided in supplemental Table S1.

Patient Material—T-cells were expanded from two melanoma tumors as previously described (29) following established protocols (30, 31). Briefly, fresh tumor samples were cut in small fragments and placed in 24-well plate containing RPMI CTS grade (Life Technologies), 10% Human serum (Valley Biomedical, Winchester, VA), 0.025 M HEPES (Life Technologies), 55 μmol/l 2-Mercaptoethanol (Life Technologies) and supplemented with a high concentration of IL-2 (Proleukin, 6,000 IU/ml, Novartis, Basel, Switzerland) for 3 to 5 weeks. Following this initial pre-rapid expansion, tumor infiltrating lymphocytes (TILs) were then expanded in using a rapid expansion protocol approach. To do so, 25×10^6 TILs were stimulated with irradiated feeder cells, anti-CD3 (OKT3, 30 ng/ml, Miltenyl biotec) and high dose IL-2 (3,000 IU/ml) for 14 days. The final cell product was washed and prepared using a cell harvester (LoVo, Fresenius Kabi, Lake County, IL). On receival of TIL samples, the cells were washed with PBS on ice, aliquoted to a cell count of 1×10^8 and stored as dry pellets at −80 °C until use.

Snap frozen meningioma tissues from patients (3830-NJF, 3849-BR, 3912-BAM, 3865-DM) were obtained from the University Hospital of Lausanne (CHUV, Lausanne, Switzerland).

Informed consent of the participants was obtained following requirements of the institutional review board (Ethics Commission, CHUV). Protocol F-25/99 has been approved by the local Ethics committee and the biobank of the Lab of Brain Tumor Biology and Genetics.

Generation of Antibody-crosslinked Beads—W6/32 and HB145 monoclonal antibodies were purified from the supernatant of HB95 (ATCC® HB-95™) and HB145 cells (ATCC® HB-145™) grown in CELLLine CL-1000 flasks (Sigma-Aldrich, St. Louis, MI) using protein-A Sepharose 4B (Pro-A) beads (Invitrogen, Carlsbad, CA). Antibodies were cross-linked to Pro-A beads at a concentration of 5 mg of antibodies per 1 ml volume of beads. For this purpose, the antibodies were incubated with the Pro-A beads for 1 h at room temperature. Chemical cross-linking was performed by addition of Dimethyl pimelimidate dihydrochloride (Sigma-Aldrich) in 0.2 M Sodium Borate buffer pH 9 (Sigma-Aldrich) at a final concentration of 20 mM for 30 min. The reaction was quenched by incubation with 0.2 M ethanolamine pH 8 (Sigma-Aldrich) for 2 h. Cross-linked antibodies were kept at 4°C until use.

High-throughput Purification of HLA Class-I and -II Complexes—For high-throughput HLA-I and -II purification, we employed the Waters Positive Pressure-96 Processor (Waters, Milford, MA). For IPs, we used the 96-well single-use micro-plate with 3 μm glass fiber and 10 μm polypropylene membranes which are compatible with the processor and are commercially available (ref number: 360063, Seahorse Bioscience, North Billerica, MA). The positive pressure processor was used in each step of the procedure to generate homogenous flow of liquid through the plates. The suggested applied pressure is in the range of 3–5 psi. The following procedure is also exemplified in Fig. 1.

Preparation of lysates: In the Plate 1 experiment (see supplemental Table S2) we purified the HLA-I and II peptidome from JY, CD165, PD42, CM467, RA957, TIL1, and TIL3. Cell lysis was performed with PBS containing 0.25% sodium deoxycholate (Sigma-Aldrich), 0.2 mM iodoacetamide (IAA) (Sigma-Aldrich), 1 mM EDTA, 1:200 Protease Inhibitors Mixture (Sigma-Aldrich), 1 mM Phenylmethylsulfonylfluoride (Roche, Basel, Switzerland), 1% octyl-beta-D glucopyranoside (Sigma-Alrich) at 4 °C for 1 h. In general, lysis buffer was added to the cells at a concentration of 1 × 10^8 cells/ml. Lysates were cleared by centrifugation with a table-top centrifuge (Eppendorf Centrifuge, Hamburg, Germany) at 4 °C at 14,200 rpm for 50 min. For each cell line, lysate from a total of 3 × 10^8 cells were pooled and evenly distributed as 1 × 10^8 triplicates into designated wells. Mock wells were incorporated into the experimental set-up, whereby wells contained anti-HLA-I and HLA-II cross-linked beads without addition of lysate. In the Plate 2 experiment (supplemental Table S2), snap-frozen meningioma tissue samples were placed in tubes containing ice cold lysis buffer (mentioned above) and homogenized on ice in 3–5 short intervals of 5 s each using an Ultra Turrax homogenizer (IKA, T10 standard, Staufen, Germany) at maximum speed. For one gram of tissue, 10 ml of lysis buffer was required. Lysates were cleared by centrifugation at 25,000 rpm in a high-speed centrifuge (Beckman Coulter, JSS15314, Nyon, Switzerland) at 4 °C for 50 min. To test the sensitivity of our method (Plate 3, see supplemental Table S2), we extracted HLA-I and -II peptides from 10, 30, 50, and 70 million cells as described above and we split the lysate of 1.6 × 10^8 CD165 B-cells proportionally to the desired cell amount; this was performed in triplicates. Lastly, four biological replicates of UWB.1 289 cells untreated and treated with IFNγ (1.5 × 10^8 cells each replicate) were processed in parallel for HLAp purification (Plate 4, see supplemental Table S2).

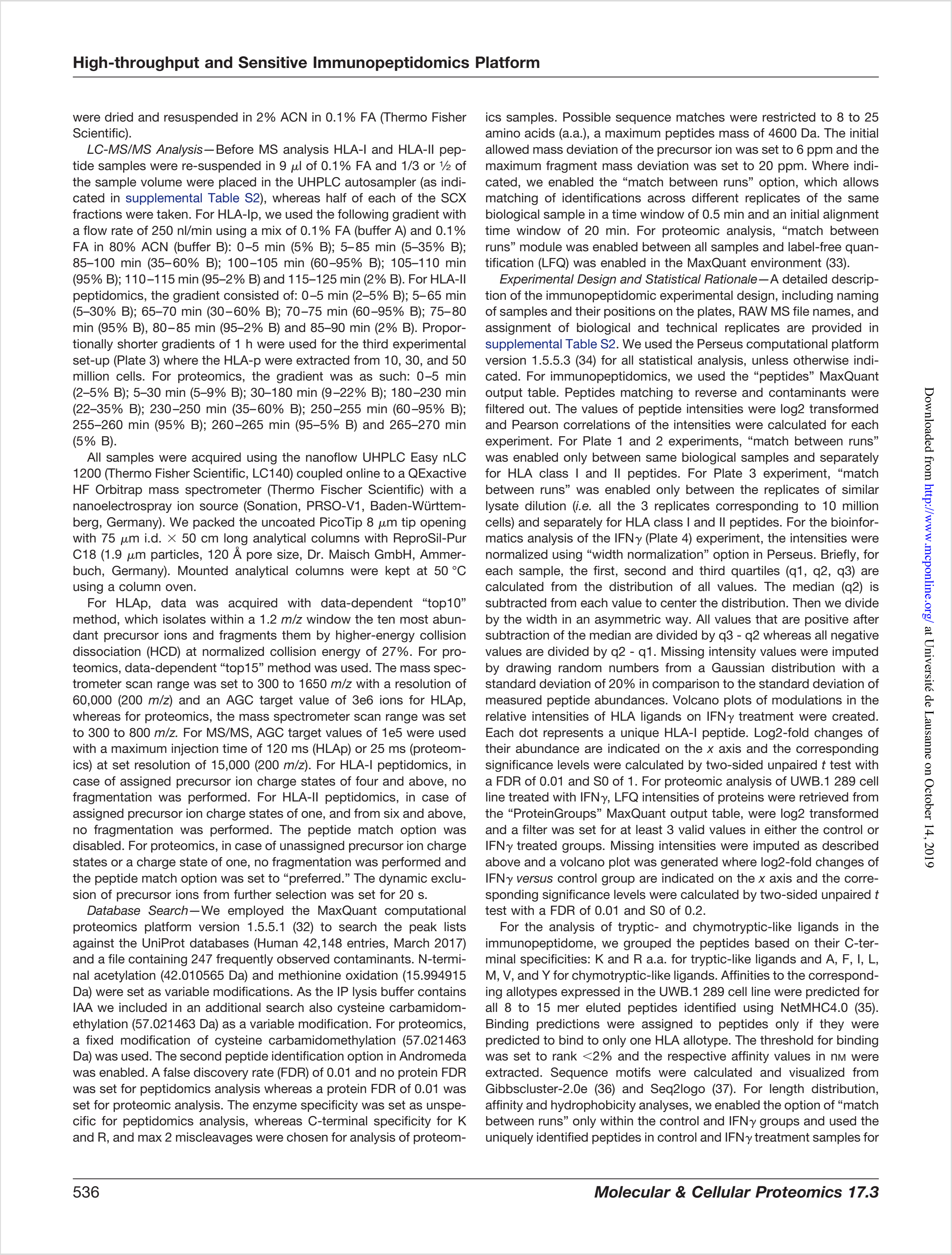
Preparation of plates: First, empty plates' wells were washed and equilibrated with 1 ml of 100% ACN (Sigma-Aldrich), followed by 1 ml of 0.1% TFA (Merck Millipore, Billerica, Massachusetts) and lastly with 2 ml of 0.1 M Tris-hydrochloric acid (HCl) pH 8 (Thermo Fisher Scientific). Anti-pan HLA-I and HLA-II antibodies cross-linked to beads were loaded on their respective plates (named "HLA class I" and "HLA class II," see Fig. 1) at a final bead volume of 75 μl in 0.1 M Tris-HCl. For tissue samples, a depletion step of endogenous antibodies was required. Therefore, an additional plate (named "Pre-clear" plate) with wells containing 100 μl Pro-A beads was prepared. The beads alone or antibodies cross-linked to beads were conditioned with lysis buffer before lysate loading.

Affinity purification of HLA complexes using the processor: As represented in Fig. 1, for tissue purification, three plates were sequentially stacked together; the Pre-clear on top, followed by the HLA class I, HLA class II and lastly, collection or waste plates. In this manner, we sequentially depleted the endogenous antibodies and immuno-affinity purified HLA class I and II complexes without intermediate steps. For cell line preparation, the pre-clear plate is not necessary. The lysates were loaded on the first plate and flowed by gravity through the preclear (for tissues only), HLA class I and II plates at 4 °C. HLA class I and II plates were then washed separately (Fig. 1) using the processor as follows: 4 times 2 ml of 150 mM sodium chloride (NaCl) (Carlo-Erba, Val de Reuil, France) in 20 mM Tris-HCl pH 8, 4 times 2 ml of 400 mM NaCl in 20 mM Tris-HCl pH 8 and again with 4 times 2 ml of 150 mM NaCl in 20 mM Tris-HCl pH 8. Finally, we washed the beads twice with 2 ml of 20 mM Tris-HCl pH 8.

Purification of HLA-I and HLA-II peptides: Two Sep-Pak tC18 100 mg Sorbent 96-well plates (named "C18 solid phase extraction" plate) (ref number: 186002321, Waters) were required for the purification and concentration of HLA-I and HLA-II peptides. Each C18 plate was handled separately. Firstly, we conditioned the plates with 1 ml of 80% ACN in 0.1% trifluoroacetic acid (TFA) and then with 2 ml of 0.1% TFA. The affinity plate was stacked on top of the C18 plate to achieve direct elution of the HLA complexes and the bound peptides with 500 μl 1% TFA. The use of TFA leads to complete denaturation of antibodies and results in a high recovery of HLAp. This is followed by washing the C18 wells with 2 ml of 0.1% TFA. Thereafter, we eluted the HLA-I peptides with 500 μl of 28% ACN in 0.1% TFA. HLA-II peptides were eluted from the class II C18 plate with 500 μl of 32% ACN in 0.1% TFA. Both HLA-I and -II peptides elutions were transferred into eppendorf tubes. Recovered HLA-I and -II peptides were dried using vacuum centrifugation (Concentrator plus Eppendorf) and stored at −20 °C. The overall time required for sample drying may vary according to the specification of the vacuum centrifuge, the user settings and amount of samples.

HLA class I and II heavy chains and the β2m molecules were recovered from the C18 plates using 300 μl of 80% ACN in 0.1% TFA. The samples were dried down and re-suspended in 30 μl 0.1% TFA. One-third of each fraction was loaded onto an SDS-gel for visual inspection of HLA complexes by SDS-electrophoresis.

Sample Preparation for Proteomics Analysis—The four biological replicates of IFNγ treated and untreated UWB.1 289 cells were re-suspended in lysis buffer composed of 8 M Urea (Biochemica, Billingham, UK) and 50 mM ammonium bicarbonate (AMBIC, Sigma-Aldrich) pH 8. The cell lysates were sonicated in the Bioruptor instrument (Diagenode, B01020001, Seraing, Belgium) for 15 cycles, maximum mA for 30 s each cycle. Subsequently, centrifugation at 20,000 × *g* at 4 °C for 30 min separated the soluble from the insoluble protein fractions. The soluble fraction was collected and the protein concentration of the lysates was determined by a Bradford protein assay. Proteins were reduced with a final concentration of 5 mM DTT (Sigma-Aldrich) at 37 °C for 60 min, followed by alkylation with a final concentration of 15 mM iodoacetamide (IAA, Sigma-Aldrich) at room temperature for 60 min in the dark. After the alkylation step the digestion was carried out with a mixture of endoproteinase Lys-C and Trypsin (Trypsin/Lys-c Mix, Promega, Madison, WI). The first step consists of endoproteinase Lys-C digestion for 4 h at 37 °C with a protein to enzyme ratio of 50:1 (w/w). Subsequently, the samples were diluted 8 times with 50 mM AMBIC to a Urea concentration of 1 M. The second step of digestion was performed with Trypsin overnight at 37 °C with a substrate to enzyme ratio of 50:1 (w/w). After digestion, the samples were acidified with formic acid (FA) and desalted on C18 spin columns (Harvard Apparatus, Holliston, MA). Samples were further fractionated using 2 layers of strong-cation-exchange (SCX) discs (Empore, Sigma-Aldrich) inserted into 20 μl StageTips generated in-house. Centrifugation was performed at up to 500 rcf on a tabletop centrifuge. Three fractions were collected by eluting with 75 mM ammonium acetate (NH_4AcO) pH 4 (Sigma-Aldrich), 200 mM NH_4AcO pH 5 and 5% Ammonia (Merck, Corsier-sur-Vevey, Switzerland) in 80% ACN pH 12. The fractions were dried and resuspended in 0.1% TFA for desalting on C18 spin columns. Finally, the samples



were dried and resuspended in 2% ACN in 0.1% FA (Thermo Fisher Scientific).

LC-MS/MS Analysis—Before MS analysis HLA-I and HLA-II peptide samples were re-suspended in 9 μl of 0.1% FA and 1/3 or ½ of the sample volume were placed in the UHPLC autosampler (as indicated in supplemental Table S2), whereas half of each of the SCX fractions were taken. For HLA-Ip, we used the following gradient with a flow rate of 250 nl/min using a mix of 0.1% FA (buffer A) and 0.1% FA in 80% ACN (buffer B): 0–5 min (5% B); 5–85 min (5–35% B); 85–100 min (35–60% B); 100–105 min (60–95% B); 105–110 min (95% B); 110–115 min (95–2% B) and 115–125 min (2% B). For HLA-II peptidomics, the gradient consisted of: 0–5 min (2–5% B); 5–65 min (5–30% B); 65–70 min (30–60% B); 70–75 min (60–95% B); 75–80 min (95% B), 80–85 min (95–2% B) and 85–90 min (2% B). Proportionally shorter gradients of 1 h were used for the third experimental set-up (Plate 3) where the HLA-p were extracted from 10, 30, and 50 million cells. For proteomics, the gradient was as such: 0–5 min (2–5% B); 5–30 min (5–9% B); 30–180 min (9–22% B); 180–230 min (22–35% B); 230–250 min (35–60% B); 250–255 min (60–95% B); 255–260 min (95% B); 260–265 min (95–5% B) and 265–270 min (5% B).

All samples were acquired using the nanoflow UHPLC Easy nLC 1200 (Thermo Fisher Scientific, LC140) coupled online to a QExactive HF Orbitrap mass spectrometer (Thermo Fischer Scientific) with a nanoelectrospray ion source (Sonation, PRSO-V1, Baden-Württemberg, Germany). We packed the uncoated PicoTip 8 μm tip opening with 75 μm i.d. × 50 cm long analytical columns with ReproSil-Pur C18 (1.9 μm particles, 120 Å pore size, Dr. Maisch GmbH, Ammerbuch, Germany). Mounted analytical columns were kept at 50 °C using a column oven.

For HLAp, data was acquired with data-dependent "top10" method, which isolates within a 1.2 *m/z* window the ten most abundant precursor ions and fragments them by higher-energy collision dissociation (HCD) at normalized collision energy of 27%. For proteomics, data-dependent "top15" method was used. The mass spectrometer scan range was set to 300 to 1650 *m/z* with a resolution of 60,000 (200 *m/z*) and an AGC target value of 3e6 ions for HLAp, whereas for proteomics, the mass spectrometer scan range was set to 300 to 800 *m/z.* For MS/MS, AGC target values of 1e5 were used with a maximum injection time of 120 ms (HLAp) or 25 ms (proteomics) at set resolution of 15,000 (200 *m/z*). For HLA-I peptidomics, in case of assigned precursor ion charge states of four and above, no fragmentation was performed. For HLA-II peptidomics, in case of assigned precursor ion charge states of one, and from six and above, no fragmentation was performed. The peptide match option was disabled. For proteomics, in case of unassigned precursor ion charge states or a charge state of one, no fragmentation was performed and the peptide match option was set to "preferred." The dynamic exclusion of precursor ions from further selection was set for 20 s.

Database Search—We employed the MaxQuant computational proteomics platform version 1.5.5.1 (32) to search the peak lists against the UniProt databases (Human 42,148 entries, March 2017) and a file containing 247 frequently observed contaminants. N-terminal acetylation (42.010565 Da) and methionine oxidation (15.994915 Da) were set as variable modifications. As the IP lysis buffer contains IAA we included in an additional search also cysteine carbamidomethylation (57.021463 Da) as a variable modification. For proteomics, a fixed modification of cysteine carbamidomethylation (57.021463 Da) was used. The second peptide identification option in Andromeda was enabled. A false discovery rate (FDR) of 0.01 and no protein FDR was set for peptidomics analysis whereas a protein FDR of 0.01 was set for proteomic analysis. The enzyme specificity was set as unspecific for peptidomics analysis, whereas C-terminal specificity for K and R, and max 2 miscleavages were chosen for analysis of proteomics samples. Possible sequence matches were restricted to 8 to 25 amino acids (a.a.), a maximum peptides mass of 4600 Da. 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. Where indicated, we enabled the "match between runs" option, which allows matching of identifications across different replicates of the same biological sample in a time window of 0.5 min and an initial alignment time window of 20 min. For proteomic analysis, "match between runs" module was enabled between all samples and label-free quantification (LFQ) was enabled in the MaxQuant environment (33).

Experimental Design and Statistical Rationale—A detailed description of the immunopeptidomic experimental design, including naming of samples and their positions on the plates, RAW MS file names, and assignment of biological and technical replicates are provided in supplemental Table S2. We used the Perseus computational platform version 1.5.5.3 (34) for all statistical analysis, unless otherwise indicated. For immunopeptidomics, we used the "peptides" MaxQuant output table. Peptides matching to reverse and contaminants were filtered out. The values of peptide intensities were log2 transformed and Pearson correlations of the intensities were calculated for each experiment. For Plate 1 and 2 experiments, "match between runs" was enabled only between same biological samples and separately for HLA class I and II peptides. For Plate 3 experiment, "match between runs" was enabled only between the replicates of similar lysate dilution (*i.e.* all the 3 replicates corresponding to 10 million cells) and separately for HLA class I and II peptides. For the bioinformatics analysis of the IFNγ (Plate 4) experiment, the intensities were normalized using "width normalization" option in Perseus. Briefly, for each sample, the first, second and third quartiles (q1, q2, q3) are calculated from the distribution of all values. The median (q2) is subtracted from each value to center the distribution. Then we divide by the width in an asymmetric way. All values that are positive after subtraction of the median are divided by q3 - q2 whereas all negative values are divided by q2 - q1. Missing intensity values were imputed by drawing random numbers from a Gaussian distribution with a standard deviation of 20% in comparison to the standard deviation of measured peptide abundances. Volcano plots of modulations in the relative intensities of HLA ligands on IFNγ treatment were created. Each dot represents a unique HLA-I peptide. Log2-fold changes of their abundance are indicated on the *x* axis and the corresponding significance levels were calculated by two-sided unpaired *t* test with a FDR of 0.01 and S0 of 1. For proteomic analysis of UWB.1 289 cell line treated with IFNγ, LFQ intensities of proteins were retrieved from the "ProteinGroups" MaxQuant output table, were log2 transformed and a filter was set for at least 3 valid values in either the control or IFNγ treated groups. Missing intensities were imputed as described above and a volcano plot was generated where log2-fold changes of IFNγ *versus* control group are indicated on the *x* axis and the corresponding significance levels were calculated by two-sided unpaired *t* test with a FDR of 0.01 and S0 of 0.2.

For the analysis of tryptic- and chymotryptic-like ligands in the immunopeptidome, we grouped the peptides based on their C-terminal specificities: K and R a.a. for tryptic-like ligands and A, F, I, L, M, V, and Y for chymotryptic-like ligands. Affinities to the corresponding allotypes expressed in the UWB.1 289 cell line were predicted for all 8 to 15 mer eluted peptides identified using NetMHC4.0 (35). Binding predictions were assigned to peptides only if they were predicted to bind to only one HLA allotype. The threshold for binding was set to rank <2% and the respective affinity values in nM were extracted. Sequence motifs were calculated and visualized from Gibbscluster-2.0e (36) and Seq2logo (37). For length distribution, affinity and hydrophobicity analyses, we enabled the option of "match between runs" only within the control and IFNγ groups and used the uniquely identified peptides in control and IFNγ treatment samples for comparison. The same list of peptides was used for comparing predicted binding affinities between control and IFNγ treated samples. Hydrophobicity scores were calculated online with https://www.protpi.ch/Calculator. Their significance levels of control and IFNγ treated samples were calculated using a two-sided unpaired *t* test. IceLogo was used to calculate the statistics to find over-represented a.a. in each position of HLA-B*07:02, -A*68:01 and -A*03:01 predicted binders of the IFNγ dataset compared with the control, with a *p* value cut-off of 0.01(38). The normalized output tables were parsed using a Java program to retain proteins with matched overlapping HLA-Ip sequences. We determined nested pairs of peptides containing the same core region (referred as "short") differing by up to five a.a. to the left (N-terminal) or to the right (C-terminal) (referred as "long"). Intensity changes on IFNγ treatment were calculated as Normalized log2-intensity difference = log2((IFNγ_{long}-ctrl$_{\text{long}}$)/(IFNγ_{short}-ctrl$_{\text{short}}$)). For in-depth analysis of C-terminal extensions, we paired short and long peptide versions based on whether they remained tryptic-like, chymo tryptic-like, or if their cleavage specificities were switched. The *p* values were calculated using a one-sided *t* test, where the null hypothesis represented zero change. Statistical calculations and plots were performed in R (www.r-project.org).

Synthetic Peptides—15 peptides (PEPotech Heavy grade 3, Thermo Fisher Scientific) with Alanine and Leucine C-terminal were selected based on their high intensities and retention time distribution from previously measured CD165 HLA-I samples. We mixed all the heavy-labeled peptides (listed in supplemental Table S3) together and desalted them on a C18 spin column. Peptides were dried to obtain 10,000 pmol of each peptide in the mixture. To test the level of cross-contamination between wells as well as reproducibility, we spiked-in the 15 peptides at 50 pmol immediately after the peptides were placed onto the C18 plate for the three replicates of CD165 HLA-I samples.

To measure the total abundance of synthetic peptides, the area under the curve (AUC) of extracted ion chromatograms for charge states z = 1+, z = 2+, and z = 3+ were calculated and summed to obtain the total signal of a given peptide. The log2 ratio between heavy and light peptides was then calculated and the mean, standard deviation and coefficient of variation (CV) were assigned for 3 exemplary synthetic peptides (see supplemental Table S4) as an example of reproducibility between the three replicates.

Gibbs Clustering Analysis for HLA-II Peptides—Gibbscluster-2.0e (31) was run independently for each sample using all HLA-IIp identified in a given sample, with the default options except that the number of clusters was tested between 1 and 6, the number of seeds for initial conditions was set to 5, the initial Monte Carlo temperature was 1.5, and we enabled the preference for hydrophobic a.a. at P1. The number of motifs plotted for each sample in supplemental Fig. S2 corresponds to the best number of motifs as determined by GibbsCluster.

We determined the reference binding motifs for each HLA-II allele based on peptides annotated in the immune epitope database (IEDB) as positive, positive-high, positive-intermediate and positive-low (39). Here Gibbscluster-2.0e was run separately per allele, with the same parameters mentioned above yet by considering a single cluster. Sequence logos were drawn with Seq2logo, based on Shannon entropy, without any sequence weighting nor Blosum correction (37).

FACS Analysis of HLA-I and -II Expression—To analyze cell surface expression of HLA-I and -II of UWB.1 289 cells on IFNγ treatment for 24 h, cells were stained with anti-HLA-A,B,C PerCP/Cy5.5 and anti-HLA-DR DP DQ FITC, or isotype-matched controls (Biolegend, San Diego, CA). Dead cells were measured using DAPI staining (PanReac Applichem, Darmstadt, Germany). Data was acquired using a LSR II SORP instrument (Beckton Dickinsons) and analyzed with the FlowJo Software version 10.3.

RESULTS

Development of a High-throughput and In-depth Immunopeptidomics Method—In an attempt to improve the sample preparation for MS-based immunopeptidomics, we revisited several recently published studies (7, 12, 17, 25–27, 40–50). Although most reported methods are similar and based on common IP procedures, our literature study systematically revealed insufficient description of experimental methodologies (supplemental Table S5) such as IP conditions, amount of cells or tissue used and the throughput of the experiment. Furthermore, the yields of quantified and identified peptides by MS may vary drastically between research labs; consequently, no fair comparisons could be conducted. Importantly, all the screened methodologies were found to have limited throughput because of lengthy (2–5 days) and laborious procedures.

We envisioned that reducing sample handling throughout all the purification steps would minimize peptide losses and significantly improve reproducibility. Thus, we designed a high-throughput 96-well plate format workflow for the simultaneous processing of tens of samples with commercially available reagents and consumables (Fig. 1). The platform employs a positive pressure processor which ensures a controlled and reproducible flow through the wells.

Briefly, tissue lysates are loaded on the first plate (Pre-clear plate in Fig. 1) containing Pro-A beads for clearance of endogenous antibodies, whereas cell lysates are loaded directly onto the plate (HLA class I plate in Fig. 1) containing anti-HLA-I antibodies covalently cross-linked to Pro-A beads for IP of HLA-I complexes. Lysates then drop directly from the first affinity plate onto the second plate (HLA class II plate in Fig. 1) that contains anti-HLA-II cross-linked to Pro-A beads. HLA class I and II plates are washed separately and each plate is then positioned on top of distinct C18 96-well plates (C18 solid phase extraction plate in Fig. 1). The HLA complexes are eluted from each of the affinity plates with TFA directly onto the corresponding C18 plate. After adequate washing of the C18 plates, the HLAp are eluted with ACN into collection plates and are ready to be dried by vacuum centrifugation and stored. The immunopurification procedure takes on average five hours including the desalting step and thus eliminates the in-process temporary storage of samples. To complement the immunopeptidomic analyses, total protein extracts and DNA can be collected from the investigated samples for shot-gun proteomics and genomics.

High-throughput Purification of HLA-Ip and HLA-IIp from Tissues and Cell Lines—To assess the throughput and overall performance of our method, we first purified HLA-Ip and HLA-IIp in a single IP procedure (Plate 1) from a total of twenty one samples (with 3 additional mock samples), which included three replicates each from five B- and two T-cell lines (10^8 cells per replicate). In a second experiment (Plate 2) we processed four primary meningioma tissues, using 0.7 to 1.47

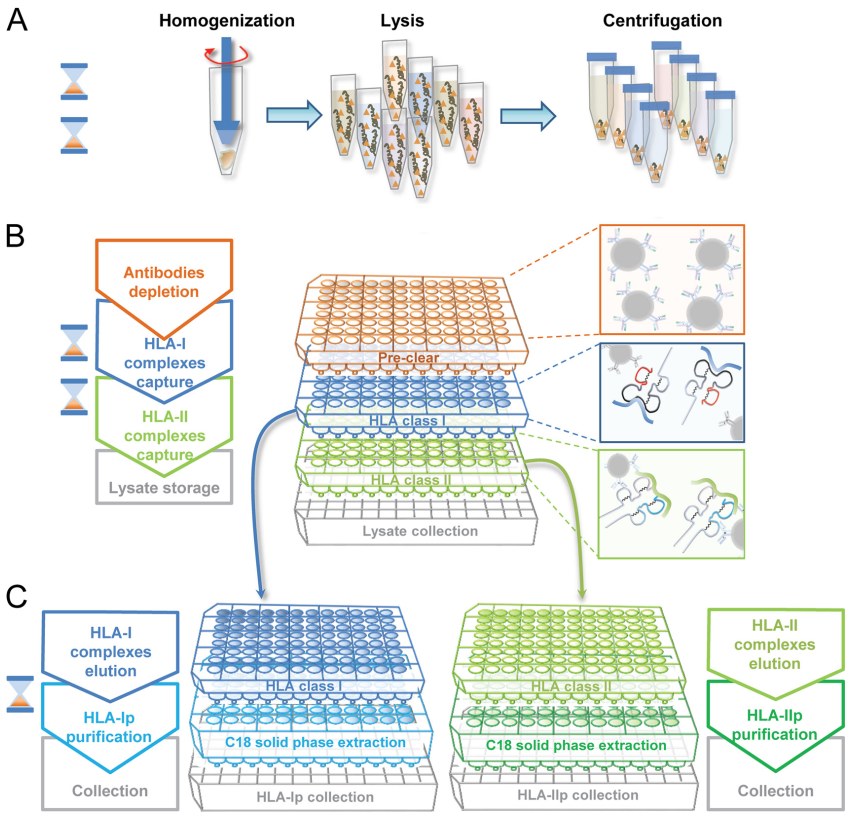


FIG. 1. **Outline of the high-throughput immunopurification workflow using a plate format.** *A*, Tissues are first homogenized, lysed with mild detergents and cleared with a centrifugation step. *B*, To enable sequential loading of the lysates on multiple affinity resins, cleared lysates are loaded on stacked plates containing firstly, Pro-A beads for depletion of tissue endogenous antibodies, then anti-HLA class I and II antibodies cross-linked to Pro-A beads for direct enrichment of HLA class I and II complexes. *C*, Affinity plates containing the captured HLA complexes are separated, washed individually and stacked on C18 plates. HLA class I and II complexes are then eluted on the C18 plates. Peptide and protein fractions are then recovered separately. Each step is timed with the hourglass symbol that is equivalent to about one hour.

grams per biological replicate. Detailed information about the experimental design is provided in supplemental Table S2 and clinical information and HLA typing are provided in supplemental Table S1. From plate number 1, a total of 42,556 unique HLA-Ip from 8975 source proteins and 43,702 unique HLA-IIp from 4501 source proteins were identified using a 1% peptide spectrum match FDR. The number of unique HLA-Ip in B- and T- cell lines varied from 3293 to 13,696 and from 7210 to 10,060 for HLA-IIp (Fig. 2*A*–2*B* and supplemental Table S6). Unlike the high concentration of about 15 mM used for carbamidomethylation of cysteines in shotgun proteomics workflows, the low concentration of 0.2 mM IAA in the IP lysis buffer facilitates irreversible inhibition of cysteine proteases, like caspases (51). Therefore, we identified a small percentage of on average 1.2% HLA-Ip and 1.9% HLA-IIp containing carbamidomethylated cysteines (supplemental Table S7). To exclude carry-over between wells during the affinity purification steps, we incorporated in plate 1 cross-linked beads not loaded with lysate (mock samples), and indeed no HLA-I or HLA-II complexes were detected here (supplemental Fig. S1*A*). In plate 2 we identified from 3497 to 14,213 HLA-Ip and from 5047 to 7972 HLA-IIp (at 1% FDR) from four patient-derived primary meningioma tissues (Fig. 2*A*–2*B* and supplemental Table S8).

In-depth and Accurate Immunopeptidomics Enables Determination of Consensus Binding Motifs—HLA-Ip datasets were highly enriched for ligands of typical length distribution for HLA-I (Fig. 2*C*). The consensus binding motifs of respective HLA-I alleles can be accurately de-convoluted from the identified peptides and the motifs match remarkably well to the

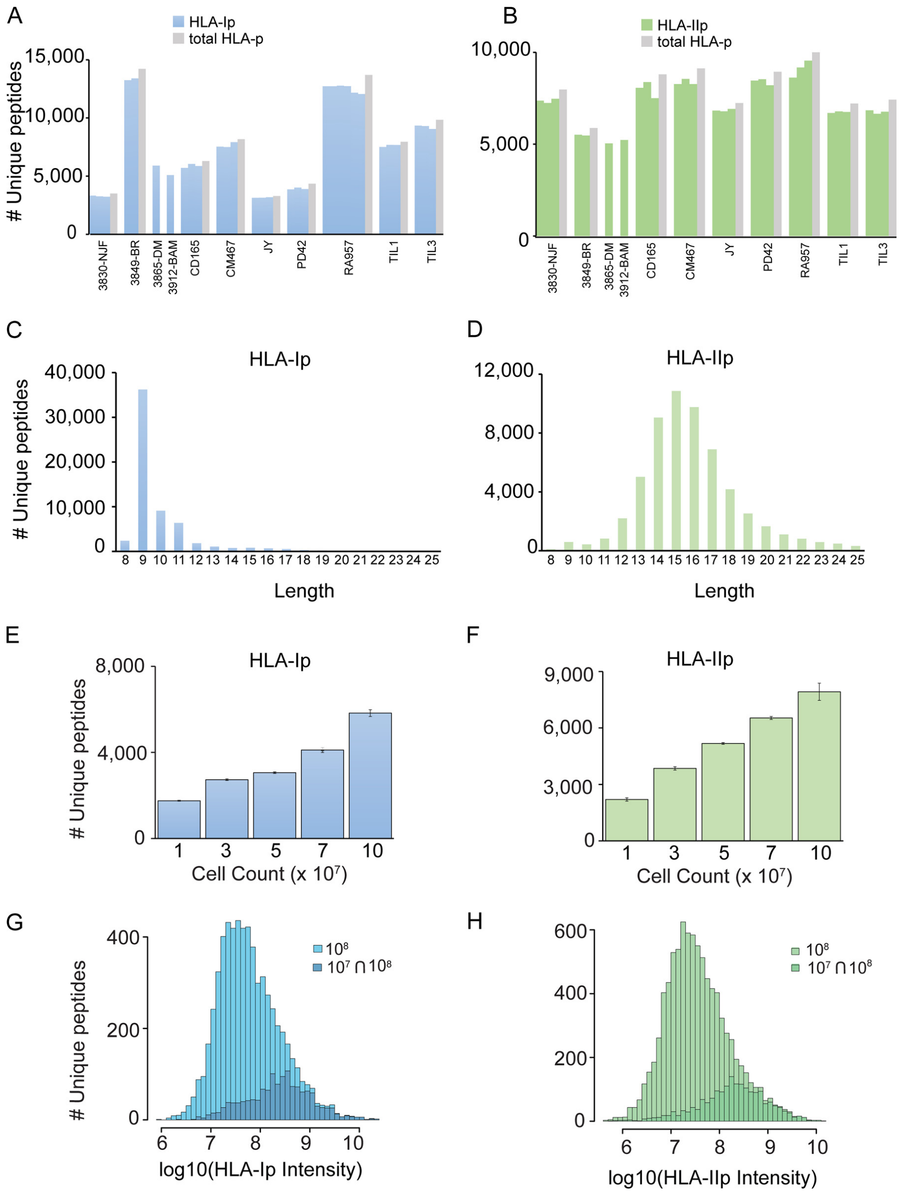


FIG. 2. **In-depth and sensitive analysis of HLA-Ip and HLA-IIp at 1% FDR for peptide identifications.** *A*, Number of unique HLA-Ip (blue bars) and (*B*) HLA-IIp (green bars) identified for B- and T-cell lines and individual tissue samples, and in total (gray bars). *C*, Length distribution of HLA-Ip and (d) of HLA-IIp. *D*, Average number of HLA-Ip (blue bars) and (*E*) HLA-IIp (green bars) identified in triplicates in lysate volumes equivalent to 10, 30, 50, 70 and 100 -million CD165 cells. Data is represented as mean ± S.D. *F*, Distribution of intensities of HLA-Ip and (*G*) HLA-IIp detected in the samples of 100 million cells and those detected in samples of both 10 million and 100 million cells.

known ones (7, 29, 52). However, in contrast to the HLA-I motifs, the core binding preferences of HLA-IIp are still poorly defined (36, 53). HLA-II molecules present longer peptides (mainly 12–19 mer and average a.a length of 15) (Fig. 2*D*) often sharing a binding core of typically 8–9 a.a. We anticipate that the great depth of our data will facilitate HLA-II motif determination. Similarly to HLA-I motif analysis (29), we de-convoluted the peptidomics data per sample and searched for the concordant motif between samples sharing the same HLA-DRB1 alleles (36). We further compared them to the motifs derived from assembled IEDB data (39). We were able to determine at least one HLA-DR motif in each of the samples with defined anchor residues typically located at positions 1, 4, 6, and 9. Furthermore, motifs of shared alleles showed a high degree of similarity between samples (supplemental Fig. S2).

Challenging the Sensitivity of the Immunopeptidomics Platform for Samples of Limited Amount—Sample amount availability poses a major limitation for the recovery of HLA class I and II peptides, especially in clinically relevant samples (12, 25–27). We reasoned that because of the fast recovery of HLA complexes and minimal sample handling, our method would also achieve substantial peptide yields even from samples of limited amount. Thus, we decided to challenge the sensitivity of our immunopurification platform by assessing HLA-I and -II peptide yields for decreasing cell amounts, down to 10^7 B-cells. We selected a B-cell line (CD165) characterized with an average yield of peptides (from the B-cell lines analyzed). Lysate volumes equivalent to 10, 30, 50, and 70 million cells in triplicates were loaded on the plate. The linear recovery of the HLA heavy chains and β2m was visualized on a SDS-gel (supplemental Fig. S1*B*). Similarly, the unique HLA-Ip and -IIp identified linearly correlated with the amount of cells (Fig. 2*E*–2*F*). From as little as 10 million cells we identified a total of 1846 HLA-Ip and 2633 HLA-IIp peptides (Fig. 2*E*–2*F* and supplemental Tables S9 and S10) and as expected, the peptides identified from 10 million cells were among the most abundant ones detected in the samples containing a 100 million cells (Fig. 2*G*–2*H*).

Assessment of Intra- and Interplate Reproducibility—For a thorough evaluation of the reproducibility, we distributed the same amount of lysates from each of the B- and T-cell lines into triplicate wells within the same plate (Plate 1). First, we assessed the overlap of detection of HLA-Ip and HLA-IIp in one, two or all three replicates of the RA957 cell line. 84% of HLA-Ip overlapped in all 3 replicates, 12% in 2 out of 3 and only 4% in one replicate. In the case of HLA-IIp, 79% of peptides were found in all 3 replicates, 15% in 2 out of 3 and only 6% in one replicate (Fig. 3*A*–3*B*). The overall reproducibility of the MS signal at the peptide level displayed Pearson correlation coefficients (r) ranging from 0.89 to 0.98 for HLA-Ip, and from 0.89 to 0.97 for HLA-IIp (Fig. 3*C*–3*D*). Notably, the reproducibility between wells was as good as the reproducibility of MS-technical duplicates of the RA957 samples (Fig. 3*E*–3*F*). Additionally, CD165, CM647 and JY samples were distributed in Plate 1 (supplemental Table S2) to non-adjacent wells to assess how plate-positional effects would affect reproducibility; no evident plate-positional effects were observed (Fig. 3*C*–3*D*).

High correlations (r) of 0.93 were also observed between the peptides extracted from different sections of 3849-BR and 3830-NJF meningioma tissues (Fig. 3*C*–3*D*, see Plate 2, supplemental Table S2), emphasizing the applicability of our platform for more challenging clinical tissue samples. In addition, JY cells of similar amounts were purified on different days, with new reagents and using orthogonal wells across the plates to evaluate interplate performance (supplemental Tables S2 and S11). Average correlations (r) of 0.93 for HLA-Ip and 0.9 for HLA-IIp were observed (supplemental Fig. S3*A*–3*B*). The peptide recovery was further evaluated by spiking 15 heavy-labeled peptides into CD165 HLA-Ip samples. All 15 heavy-labeled and their endogenous counterparts were identified in each of the replicates. Their retention times are reported in supplemental Table S3. The CV of the ratio between the heavy-labeled and endogenous peptides was calculated for three exemplary cases resulting in a CV of 1% between the replicates (supplemental Table S4). The synthetic peptides were additionally used to evaluate carry-overs between wells during desalting steps and no synthetic peptides were detected in neighboring samples after manual inspection of the RAW MS data (supplemental Table S3).

Highly Reproducible Analysis Facilitates Label-free Comparative Study of the Drug-modulated Immunopeptidome—We reasoned that our streamlined method would enable a qualitative and quantitative assessment of HLAp alterations on external stimuli, potentially revealing the mechanistic mode of action. Thus, as a proof of concept we interrogated alterations induced by IFNγ on the UWB.1 289 ovarian cancer cell line. IFNγ is a key cytokine that activates multiple immune related signaling pathways and hence modulates the expression of hundreds of genes. Specifically, it is known to up-regulate the expression of HLA-I complexes as well as other key proteins involved in the antigen processing and presentation pathway in tumor cells (54). Indeed, on 24 h of IFNγ treatment of the UWB.1 289 cells, we detected enhanced cell surface expression of HLA-I by FACS and a global increase of total HLA-I and β2m by SDS-gel analysis (supplemental Fig. S4*A*–S4*B*). Average Pearson correlations of 0.95 between IFNγ-treated and of 0.97 between control replicates were obtained (Fig. 4*A*). We identified on average 4090 unique HLA-Ip in controls and 5195 peptides in IFNγ treated samples (Fig. 4*B*). However, with the "match between runs" option which enables the assignment of identifications to MS features that were not selected for fragmentation in all replicates (55), the number of identified peptides in controls evened up to the number detected in IFNγ treated samples (Fig. 4*B* and supplemental Table S12). The overlap of the two datasets was then as high as 91%. This observation, together with the sum of peptide

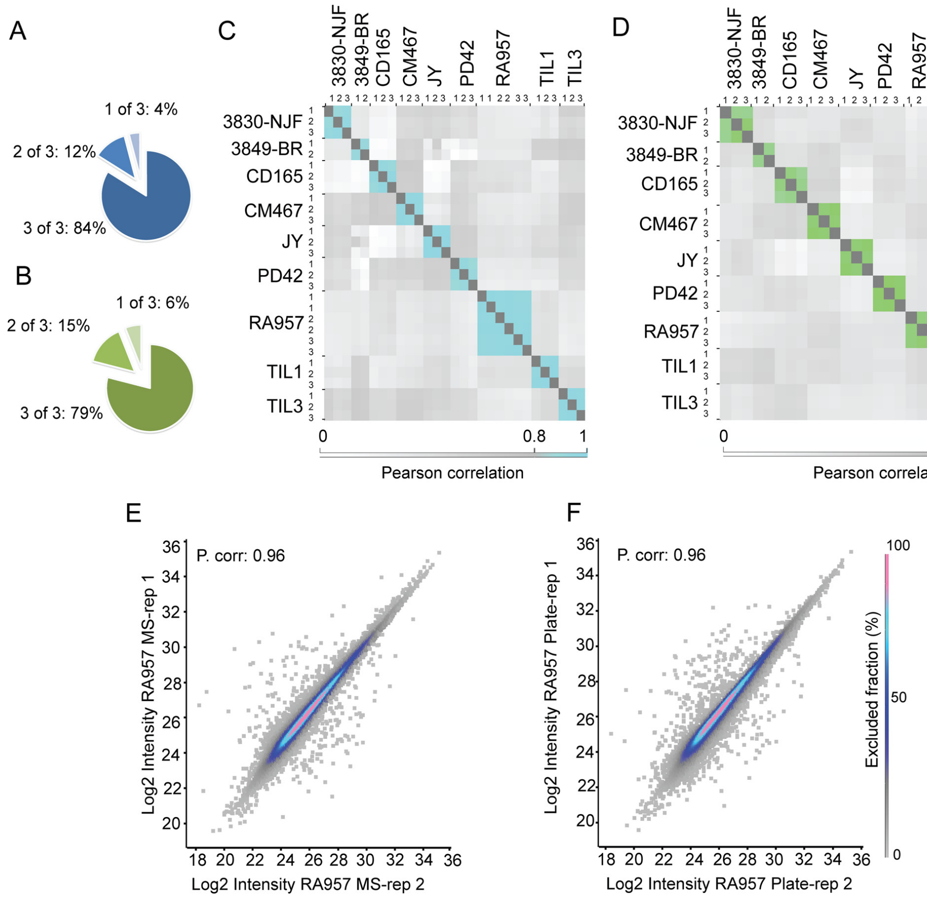


FIG. 3. **Assessment of intra-plate reproducibility.** *A*, Overlap in the frequency of HLA-Ip and (*B*) HLA-IIp identified in three plate replicates of RA957 samples. *C*, Intra-plate reproducibility calculated by Pearson correlations of log2 transformed intensities of HLA-Ip and (*D*) HLA-IIp identified across the different MS measurements. *E*, Examples of comparative semi-quantitative analysis of HLA-Ip detected in two MS measurements (referred here as technical MS replicates) of one RA957 sample and (*F*) of two representative plate replicates of RA957 samples. Values of the Pearson correlation are indicated.

signal intensities (Fig. 4*C*) suggested that IFNγ led to quantitative reshaping of the repertoire.

Immunopeptidomics and Proteomics Capture Similar Global Changes on IFNγ Treatment—We further explored qualitative global changes in the HLAp repertoire modulated on IFNγ treatment and detected 1157 HLA-Ip that were significantly up-regulated and 551 down-regulated HLA-Ip (*t* test FDR = 0.01, S0 = 1). Among the up-regulated HLA-Ip we detected peptides derived from well-known intracellular mediators of IFNγ (Fig. 4*D*) (56) and this observation was confirmed with our proteomics analysis (supplemental Table S13). Proteins involved in antigen processing and presentation and consequently in the IFN-mediated immune response, were significantly up-regulated, among them STAT1 and 2, TAP1 and 2, β2m, OAS3, WARS, IFI16, and IRF (Fig. 4*D*). The constitutive proteasomal subunits (*i.e.* β5 (PSMB5), β1 (PSMB6) and β2 (PSMB7)) were not found to be differentially regulated on IFNγ treatment. On the other hand, the immunoproteasomal subunits (*i.e.* β5i (PMSB8) β1i (PSMB9), and β2i (PSMB10)) were up-regulated on exposure to IFNγ in both peptidomics and proteomics datasets (Fig. 4*D*–4*E*), supporting the switch from the constitutive- to the immunoproteasome (57). Notably, the constitutive proteasome has both tryptic and chymotryptic-like activities, whereas the IFNγ-

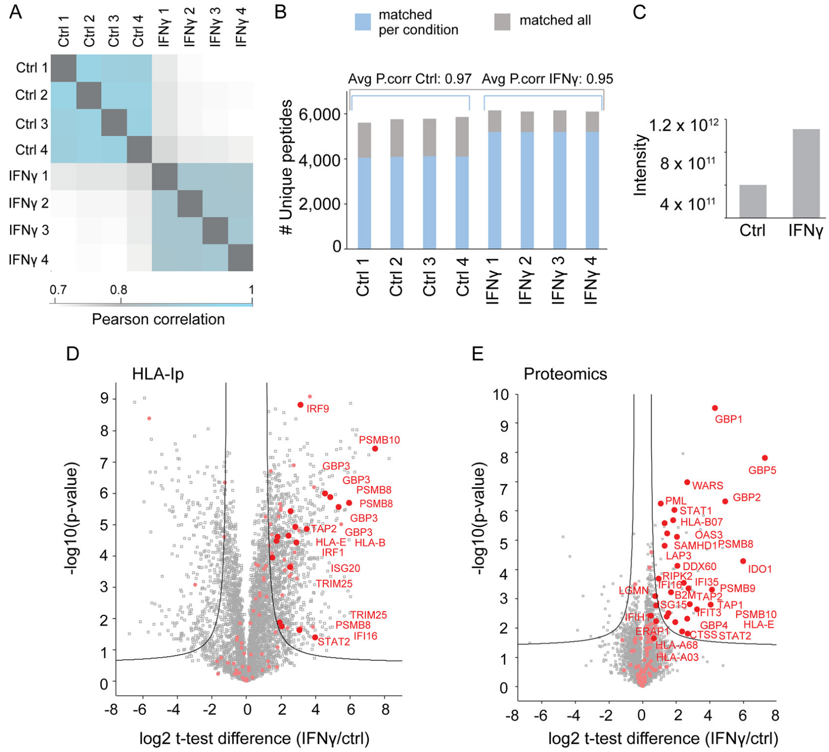
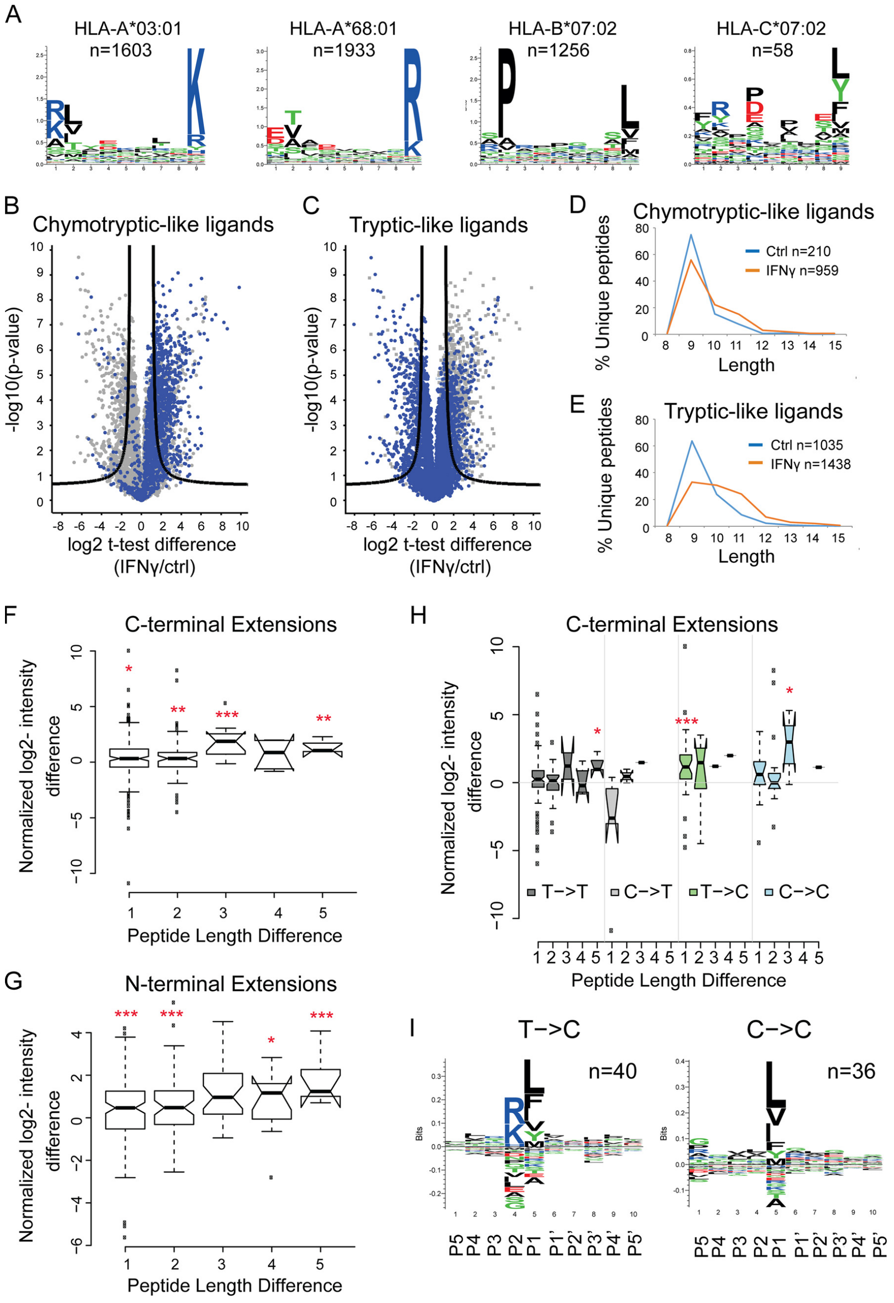


FIG. 4. **Label-free semi-quantitative comparative analysis of IFNγ modulated immunopeptidome.** *A*, Reproducibility calculated by Pearson correlations of log2 transformed intensities from HLA-Ip of control and IFNγ -treated samples across the different MS measurements. *B*, Number of HLA-Ip identified from UWB.1 289 ovarian cancer cells untreated (control) and treated with IFNγ. The number of peptides identified with (gray) and without (blue) matching identifications across the treated and untreated samples and the average values of the Pearson correlations are indicated. *C*, Summed peptide intensities identified in each of the IFNγ treated and control samples. *D*, Volcano plot summarizing unpaired *t* test analysis of the immunopeptidome of IFNγ treated *versus* untreated cells. Peptides located above the lines are statistically significantly modulated in their level of presentation (FDR = 0.01, S0 = 1). All peptides derived from proteins related to immunity are highlighted in pink. Selected up-regulated peptides were highlighted in red, corresponding to well known intracellular mediators of IFNγ signaling. *E*, Volcano plot of unpaired *t* test analysis of the proteome of IFNγ treated *versus* untreated cells. Proteins located above the lines are statistically significantly modulated in their expression level (FDR = 0.01, S0 = 0.2). Selected proteins involved were similarly highlighted.

induced immunoproteasome has been shown to exhibit quantitatively higher chymotryptic-like activity (58).

IFNγ Induced Presentation of Longer Peptides and of Peptides Harboring C-terminal Chymotryptic-like Residues—As the chosen UWB.1 289 cell line expresses HLA-I alleles of C-terminal tryptic motifs (A03:01, A68:01) and C-terminal chymotryptic-like motifs (B07:02) (Fig. 5*A*) we hypothesized that analyzing the repertoire changes on IFNγ could uncover the impact of the immunoproteasome on the presented ligandome. First, we grouped the peptides based on their chymotryptic- or tryptic-like C-terminal (regardless of their HLA allele preferences). This revealed an enhanced presentation of chymotryptic-like ligands (Fig. 5*B*), whereas the presentation of tryptic-like ligands did not differ substantially on IFNγ treatment (Fig. 5*C*). Another global effect of the treatment was a general distribution of longer peptides that were uniquely detected on IFNγ treatment compared with those in control samples (Fig. 5*D*–5*E*).

Previous reports have shown peptides containing the binding motif within a common core region but extending beyond the motif in either N- or C-terminal directions (21, 42, 59, 60). About 7% of the peptides in this dataset were found overlapping entirely with longer peptide sequences. When the short and long peptide pairs were found to start at the same position we named them C-terminal elongated pairs. Similarly, short and long peptide pairs ending at the same position were named N-terminal elongated pairs (supplemental Table S14). We observed that after IFNγ stimulation, the longer peptides,


A
HLA-A*03:01
n=1603
HLA-A*68:01
n=1933
HLA-B*07:02
n=1256
HLA-C*07:02
n=58
B
Chymotryptic-like ligands
-log10(p-value)
log2 t-test difference
(IFNγ/ctrl)
C
Tryptic-like ligands
-log10(p-value)
log2 t-test difference
(IFNγ/ctrl)
D
Chymotryptic-like ligands
% Unique peptides
Ctrl n=210
IFNγ n=959
Length
E
Tryptic-like ligands
% Unique peptides
Ctrl n=1035
IFNγ n=1438
Length
F
C-terminal Extensions
Normalized log2- intensity difference
Peptide Length Difference
H
C-terminal Extensions
Normalized log2- intensity difference
T−>T
C−>T
T−>C
C−>C
Peptide Length Difference
G
N-terminal Extensions
Normalized log2- intensity difference
Peptide Length Difference
I
T−>C
n=40
C−>C
n=36
Bits
P5
P4
P3
P2
P1
P1'
P2'
P3'
P4'
P5'

in both N- and C-terminal directions, were often significantly more abundant compared with their shorter counterparts (Fig. 5*F*–5*G*), thereby contributing to the global length shift (Fig. 5*D*–5*E*). The extensions range between one to five a.a. along both termini.

Because the proteasome is known to determine the C-terminal cleavage specificity of HLA-Ip, we further explored if the enhanced presentation of the C-terminal elongated peptides is in agreement with the global increase in presentation of chymotryptic-like peptides on IFNγ treatment. Thus, we grouped the C-terminal elongated pairs based on whether their C-terminal remained tryptic-like, chymotryptic-like, or if their specificity was switched. We observed that on treatment, elongated chymotryptic-like peptides were significantly more abundant than their shorter tryptic-like peptides, mainly by one a.a. (supplemental Table S14 and Fig. 5*H*). The switch in cleavage specificity can be further visualized by comparing long peptides changing from tryptic- to chymotryptic-like (n = 40) and those maintaining their chymotryptic-like (n = 36) C-terminal cleavage specificity (Fig. 5*I*).

Allele-specific Analysis of the Immunopeptidome—Allele-specific presentation on IFNγ can be defined by other factors apart from the immunoproteasomal cleavage preferences. Therefore, we predicted the binding affinities with NetMHC 4.0 for HLA-A*03:01, -A*68:01, -B*07:02 and -C*07:02 (Fig. 5*A* and supplemental Table S15). We could not predict binding to the HLA-C*03:32 allele as the motif for this allele is currently still unknown and we excluded the HLA-C*07:02 binders because of their small population size (n = 58). Finally, we assigned the allele specificities to peptides that were predicted to bind to only one allele.

In line with the analysis for chymotryptic-like ligands, we detected an enhanced presentation of the HLA-B*07:02 peptides which contains C-terminal chymotryptic-like a.a. (Fig. 5*B* and supplemental Fig. S4*C*). Increased expression of HLA-B*07:02 heavy chain molecules was evident also at the proteomic level (Fig. 4*E*). Both HLA-A*03:01 and -A*68:01 have tryptic-like binding motifs, but with marked differences at the C-terminal (Fig. 5*A*). When examining these allele-specific populations, we observed that peptides predicted to bind the HLA-A*03:01 molecules were slightly down-regulated in contrast to HLA-A*68:01 ones (supplemental Fig. S4*C*). Overall, the predicted binding affinities of HLA-B*07:02 ligands were similar regardless of IFNγ treatment, whereas HLA-A*03:01 and -A*68:01 ligands were predicted to bind with higher affinity on IFNγ stimulation, although not statistically significant in all peptide lengths (supplemental Fig. S5*A*). We detected a statistically significant increase in the composition of hydrophobic a.a. (as indicated by the hydrophobicity score Φ, supplemental Fig. S5*B*) in peptides across all three alleles.

DISCUSSION

The extraction procedure of HLA ligands for deep MS analysis has been a major limitation (19). We present here a greatly improved IP-based HLA-Ip and -IIp purification pipeline which has been rigorously optimized and encompasses several new features and advantages: (1) the fast IP step minimizes artifacts possibly introduced during long incubations, (2) minimal in-process sample handling and freezing steps allow competitive recovery and sensitivity, (3) drastic reduction in the amount of expensive antibody-crosslinked beads, (4) parallel processing of dozens of samples and (5) elimination of error prone steps, making the pipeline suitable for processing valued patient-derived tissue samples. We demonstrated the high-throughput nature of the workflow by purifying in a single IP procedure HLA-Ip and HLA-IIp from twenty one samples (only 10^8 cells per replicate). The depth and reproducibility of the enriched HLA-I and -II peptidomes were outstanding with an average of more than 7500 unique peptides identified in single IP for both class I and II (Fig. 2*A*–2*B*). Furthermore, the overall reproducibility (Pearson correlation coefficient) ranged from 0.89 to 0.98 for HLA-Ip, and from 0.89 to 0.97 for HLA-IIp (Fig. 3*C*–3*D*). We affirmed the pipeline's robust performance and clinical applicability by parallel processing of four primary meningioma tissues of different quantities, which matched well with their peptide yields (Figs. 2*A*–2*B* and 3*C*–3*D*).

Importantly, a major bottle-neck of immunopeptidomics is still the requirement of a relatively large amount of cells or tissue material, which is not always feasible to obtain from clinical samples. We showed that our methodology reached a degree of sensitivity that enables us to identify 1846 HLA-Ip and 2633 HLA-IIp from as little as 10 million cells (Fig. 2*E*–2*F*). This achievement therefore highlights the possibility to drastically scale down the sample amount, when required.

The purity and depth of the extracted peptidomes allowed us to determine high-resolution HLA class II motifs compa-

FIG. 5. **Impact of IFNγ on global features of HLA class-I repertoire.** *A*, Peptides were assigned to the different HLA allotypes based on binding affinity predictions and their binding motifs depicted with sequence logos. *B*–*C*, Volcano plots summarizing unpaired *t* test analysis of the immunopeptidome of IFNγ treated *versus* untreated cells. Peptides located above the lines are statistically significantly modulated in their level of presentation (FDR = 0.01, S0 = 1). All chymotryptic-like (*B*) and tryptic-like ligands (*C*) were highlighted, respectively. *D*–*E*, Length distribution of peptides uniquely identified in IFNγ treated (orange) or control (blue) samples according to their chymotryptic- (*D*) or tryptic-like (*E*) properties. *F*–*G*, Intensity changes on IFNγ treatment were calculated for longer peptides against their shorter versions for both C- or N-terminal extensions: Normalized log2-intensity difference = log2((IFNγ_{long}-ctrl$_{long}$)/(IFNγ_{short}-ctrl$_{short}$)). (*H*) C-terminal nested versions were grouped based on whether their extended peptides remained tryptic-like (T→T), chymo-tryptic like (C→C), or if their specificities were switched (C→T or T→C). Log2-intensity changes on IFNγ treatment were calculated for longer peptides against their shorter versions (*I*) For T→C and C→C peptide pairs, the sequence logos around the cleavage site of the long peptides (C-terminal P1–5, downstream P'1–5) are depicted (one-sided *t* test, *p* value * < 0.1; ** < 0.05; and *** < 0.01).

rable to the IEDB data (supplemental Fig. S2). A current limitation to obtain the HLA class II binding motifs from this approach arises when the alleles from an individual sample share very similar binding motifs (*i.e.* DRB1*01:01 and DRB1*07:01). In such eventuality, GibbsCluster cannot cluster the peptides into distinct motifs. Motif annotation can also be hindered by the lack of existing data and reference motifs in the IEDB database.

The current data allowed the identification of motifs for most of the HLA-DR alleles present in each sample. However, these samples also contained HLA-DP and HLA-DQ alleles. As such we cannot exclude that some of the motifs predicted by Gibbscluster for which we could not find any HLA-DR may correspond to -DP or -DQ motifs. Unfortunately, both the alpha and beta chains show high variability in these alleles, such that each sample could contain up to four different HLA-DP and four different HLA-DQ combinations, which makes motif identification and annotation much more challenging. Our work suggests that samples expressing monoallelic selected HLA-DP and HLA-DQ alleles should be explored to reveal their specificities. We anticipate that with the growing number of samples analyzed with the method described in this work, the number of newly identified HLA-II binding motifs will quickly grow and, re-interrogation of this data can be of use for improving HLA-II ligand predictions in machine learning studies.

The development of a reproducible and high-throughput methodology allows us to generate high quality data to gain more insight into biological questions. Thus, as a proof of concept we demonstrated the feasibility of performing comparative screening on IFNγ treatment. Exposure of tumor cells to IFNγ is known to induce pro-inflammatory gene signatures that consequently lead to enhanced recognition by cytotoxic T-lymphocytes (28, 61) mediated by the up-regulation of the APPM and HLA surface expression. Augmented surface presentation of HLA complexes could lead to a higher probability of presentation and, hence, recognition of immunogenic epitopes. However, no high-quality mapping of the IFNγ modulated peptidome has been reported so far and therefore, the overall properties of the presented repertoire on stimulation remain unknown.

Here, we uncovered a global modulation in the immunopeptidome on exposure to IFNγ. We estimated that the HLA-Ip repertoire increased by 170%, as depicted from the differential MS intensities (Fig. 4*C*). The boost in HLA-I expression was also validated by FACS, as well as by semi-quantification of the HLA heavy chains and β2m (supplemental Fig. S4*A*–S4*B*). In addition, the proteomics analysis confirmed the up-regulation of intracellular sensors and mediators of the IFNγ signaling pathway as well as several of its effectors (Fig. 4*E*), as previously reported (56). The IFNγ signature was clearly conveyed at the peptide repertoire (Fig. 4*D*–4*E*), supporting the reported positive correlation between proteome expression and antigen presentation (7). In fact, both at the peptidome and proteome level, various key components of the APPM were up-regulated, such as TAP 1 and 2, β2m, HLA heavy chains and the immunoproteasomal subunits (*i.e.* β5i (PMSB8) β1i (PSMB9), and β2i (PSMB10)). Our results clearly showed a marked shift in presentation of chymotryptic-like ligands (*i.e.* HLA-B*07:02 binders) on IFNγ stimulation (Fig. 5*A* and supplemental Fig. S4*C*). This can be explained by the combined effect of the increased expression of HLA-B*07:02 molecules (Fig. 4*E*) and by the switch from the constitutive proteasome to immunoproteasome. In fact, proteasomal switching may have led to a more efficient generation of peptides harboring C-terminal chymotryptic-like residues (*i.e.* same as HLA-B*07:02 binders) (58), whereas the overall tryptic-like ligand presentation remain largely unchanged (Fig. 5*B*–5*C* and supplemental Fig. S4*C*). However, when allele specificities were taken into account, the presentation of HLA*03:01 and A*68:01 binders were differentially regulated on treatment possibly because of subtle proteasomal (or other peptidases) cleavage preferences or to the slightly different expression levels of these HLA alleles (supplemental Fig. S4*C*).

A general tuning of the APPM toward presentation of longer peptides was globally detected (Fig. 5*D*–5*E*). These longer peptides may bind via canonical residues facilitated by the bulging of the middle part of the peptide or they could bind with inner anchors leaving the extension to protrude from one end of the binding groove (42, 60, 62, 63). Uniquely to our study, we were able to quantitatively assess the enhanced presentation of peptides varying in length on IFNγ stimulation. N-terminal extended peptides did not show cleavage patterns; this is expected because of the downstream trimming events that takes place in the ER. N-terminal extended versions of canonical peptides still may contain appropriate HLA binding motifs (59). Intriguingly, we also detected C-terminal tryptic-like peptides that have statistically significant enriched chymotryptic-like longer versions (one a.a.), possibly hinting toward an enhanced chymotryptic-like activity of the immunoproteasome (Fig. 5*H*). We speculate that the production of longer peptides may have been favored, not only by proteasome switching but also by TAP shuttling (64), which was also significantly up-regulated in our proteomic analysis. Interestingly, it was observed that the first three N-terminal residues and the C-terminal residue were the most critical for TAP-binding (65). Therefore, the longer peptides we detected may have been favored because of their specific physicochemical properties for binding to the TAP (64).

Altogether, in this proof of concept study, we shed light on the quantitative re-shaping of the presented peptidome imposed by IFNγ treatment. Further investigation across additional cell lines covering more HLA allotypes is required to precisely determine the exact molecular mechanisms underlying these changes and if the observed modulation of the peptidome can be generalized. Further research is also needed to reveal the cellular conditions that facilitate the

generation and surface presentation of peptides of higher quality in terms of fitness to the HLA alleles, such as binding affinity, sequence hydrophobicity and peptide length. This would enable the incorporation of additional features in prediction algorithms of antigen presentation and hence improve their accuracy (43, 66). We present a robust methodology that enables LFQ comparative immunopeptidomics applicable for the investigation of perturbations in the antigen presentation caused for example by pathogenic infections, autoimmunity and cancer. Several pioneering vaccine companies and research labs have recently incorporated MS-based immunopeptidomics as a discovery tool to gain knowledge of the presented peptidome for improved binding predictions or to directly identify mutated antigens and clinically relevant targets for vaccinations (67, 68). We hope that our method will accelerate the development of vaccines and will assist in implementing this technology in clinical practice.

Acknowledgments—We thank the Center of Experimental Therapeutics team and specifically to Lana Kandalaft and Philippe Gannon for providing us with the patient-derived tissue samples. We are thankful to Camilla Jandus and Pedro Romero for sharing the B-cell lines with us. We thank Monika Hegi for insightful discussion about the meningioma collection. We also thank Justine Michaux for her valuable support throughout the experiments. We thank the Ludwig Cancer Research Lausanne Branch for financial support for the immunopeptidomics lab in Lausanne. We would like to thank the Vital-IT team for its support for the computational infrastructure and Ioannis Xenarios for the support of the project.

DATA AVAILABILITY

All the RAW data, the FASTA reference file, MaxQuant parameters and output tables, including selected MaxQuant output tables used for analyses (filtered for contaminants and reverse hits) have been deposited to the ProteomeXchange Consortium (69) via the PRIDE partner repository with the dataset identifier PXD006939.

* This project is supported by the ISREC Foundation for a period of three years, thanks to a donation from the Biltema Foundation.

S This article contains supplemental material.

‡‡ To whom correspondence should be addressed: Ludwig Institute for Cancer Research, University of Lausanne, 1066 Epalinges, Switzerland. E-Mail: Michal.bassani@chuv.ch.

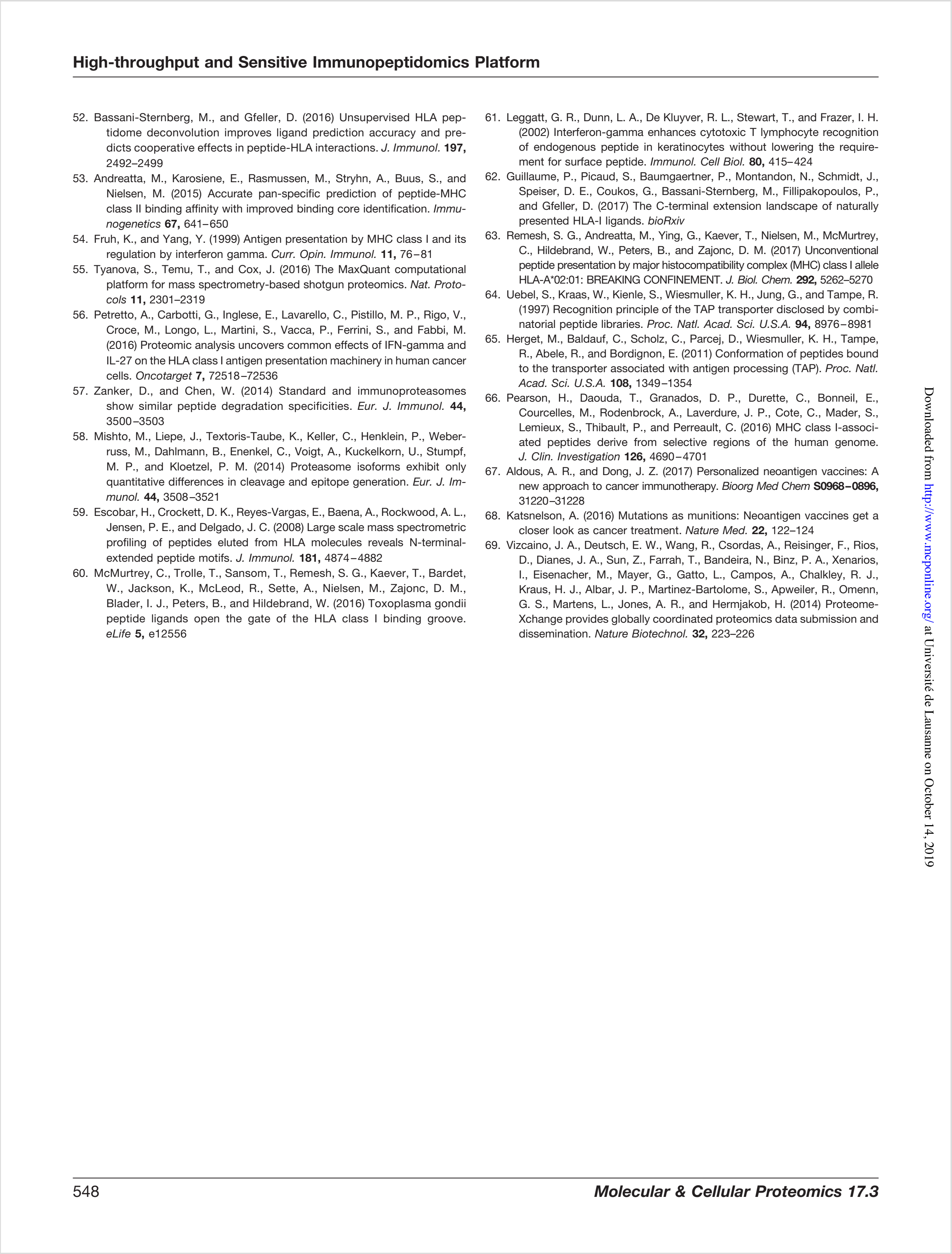
§§ These authors contributed equally to this work.

REFERENCES

1. Schadendorf, D., Hodi, F. S., Robert, C., Weber, J. S., Margolin, K., Hamid, O., Patt, D., Chen, T. T., Berman, D. M., and Wolchok, J. D. (2015) Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. *J. Clin. Oncol.* **33,** 1889–1894
2. Schumacher, T. N., and Schreiber, R. D. (2015) Neoantigens in cancer immunotherapy. *Science* **348,** 69–74
3. Yarchoan, M., Johnson B. A., 3rd, Lutz, E. R., Laheru, D. A., and Jaffee, E. M. (2017) Targeting neoantigens to augment antitumour immunity. *Nature Rev.* **17,** 209–222
4. Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., Javeri, I., Nellaiappan, K., Salazar, A. M., Daley, H., Seaman, M., Buchbinder, E. I., Yoon, C. H., Harden, M., Lennon, N., Gabriel, S., Rodig, S. J., Barouch, D. H., Aster, J. C., Getz, G., Wucherpfennig, K., Neuberg, D., Ritz, J., Lander, E. S., Fritsch, E. F., Hacohen, N., and Wu, C. J. (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547,** 217–221
5. Sahin, U., Derhovanessian, E., Miller, M., Kloke, B. P., Simon, P., Lower, M., Bukur, V., Tadmor, A. D., Luxemburger, U., Schrors, B., Omokoko, T., Vormehr, M., Albrecht, C., Paruzynski, A., Kuhn, A. N., Buck, J., Heesch, S., Schreeb, K. H., Muller, 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., Bruck, A. K., Witt, M., Zillgen, M., Rothermel, A., Kasemann, B., Langer, D., Bolte, S., Diken, M., Kreiter, S., Nemecek, R., Gebhardt, C., Grabbe, S., Holler, C., Utikal, J., Huber, C., Loquai, C., and Tureci, O. (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547,** 222–226
6. Bassani-Sternberg, M., and Coukos, G. (2016) Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr. Opin. Immunol.* **41,** 9–17
7. 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
8. Singh-Jasuja, H., Emmerich, N. P., and Rammensee, H. G. (2004) The Tubingen approach: identification, selection, and validation of tumor-associated HLA peptides for cancer therapy. *Cancer Immunol., Immunother.* **53,** 187–195
9. Weinschenk, T., Gouttefangeas, C., Schirle, M., Obermayr, F., Walter, S., Schoor, O., Kurek, R., Loeser, W., Bichler, K. H., Wernet, D., Stevanovic, S., and Rammensee, H. G. (2002) Integrated functional genomics approach for the design of patient-individual antitumor vaccines. *Cancer Res.* **62,** 5818–5827
10. Dutoit, V., Herold-Mende, C., Hilf, N., Schoor, O., Beckhove, P., Bucher, J., Dorsch, K., Flohr, S., Fritsche, J., Lewandrowski, P., Lohr, J., Rammensee, H. G., Stevanovic, S., Trautwein, C., Vass, V., Walter, S., Walker, P. R., Weinschenk, T., Singh-Jasuja, H., and Dietrich, P. Y. (2012) Exploiting the glioblastoma peptidome to discover novel tumour-associated antigens for immunotherapy. *Brain* **135,** 1042–1054
11. Berlin, C., Kowalewski, D. J., Schuster, H., Mirza, N., Walz, S., Handel, M., Schmid-Horch, B., Salih, H. R., Kanz, L., Rammensee, H. G., Stevanovic, S., and Stickel, J. S. (2014) Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. *Leukemia* **29,** 647–659
12. Walz, S., Stickel, J. S., Kowalewski, D. J., Schuster, H., Weisel, K., Backert, L., Kahn, S., Nelde, A., Stroh, T., Handel, M., Kohlbacher, O., Kanz, L., Salih, H. R., Rammensee, H. G., and Stevanovic, S. (2015) The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for T-cell-based immunotherapy. *Blood* **126,** 1203–1213
13. Bassani-Sternberg, M., Barnea, E., Beer, I., Avivi, I., Katz, T., and Admon, A. (2010) Soluble plasma HLA peptidome as a potential source for cancer biomarkers. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 18769–18776
14. Bassani-Sternberg, M., Braunlein, 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. D, H. B., 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,** 13404
15. Carreno, B. M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A. A., Ly, A., Lie, W. R., Hildebrand, W. H., Mardis, E. R., and Linette, G. P. (2015) Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348,** 803–808
16. Kalaora, S., Barnea, E., Merhavi-Shoham, E., Qutob, N., Teer, J. K., Shimony, N., Schachter, J., Rosenberg, S. A., Besser, M. J., Admon, A., and Samuels, Y. (2016) Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* **7,** 5110–5117
17. Khodadoust, M. S., Olsson, N., Wagar, L. E., Haabeth, O. A., Chen, B., Swaminathan, K., Rawson, K., Liu, C. L., Steiner, D., Lund, P., Rao, S., Zhang, L., Marceau, C., Stehr, H., Newman, A. M., Czerwinski, D. K.,

Carlton, V. E., Moorhead, M., Faham, M., Kohrt, H. E., Carette, J., Green, M. R., Davis, M. M., Levy, R., Elias, J. E., and Alizadeh, A. A. (2017) Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature* **543,** 723–727

18. Neefjes, J., Jongsma, M. L., Paul, P., and Bakke, O. (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Rev. Immunol.* **11,** 823–836
19. 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
20. Caron, E., Vincent, K., Fortier, M. H., Laverdure, J. P., Bramoulle, A., Hardy, M. P., Voisin, G., Roux, P. P., Lemieux, S., Thibault, P., and Perreault, C. (2011) The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol. Systems Biol.* **7,** 533
21. Mommen, G. P., Frese, C. K., Meiring, H. D., van Gaans-van den Brink, J., de Jong, A. P., van Els, C. A., and Heck, A. J. (2014) Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci. U.S.A.* **111,** 4507–4512
22. Kowalewski, D. J., and Stevanovic, S. (2013) Biochemical large-scale identification of MHC class I ligands. *Methods Mol. Biol.* **960,** 145–157
23. Croft, N. P., Smith, S. A., Wong, Y. C., Tan, C. T., Dudek, N. L., Flesch, I. E., Lin, L. C., Tscharke, D. C., and Purcell, A. W. (2013) Kinetics of antigen expression and epitope presentation during virus infection. *PLoS Pathog.* **9,** e1003129
24. Martin-Esteban, A., Guasp, P., Barnea, E., Admon, A., and Lopez de Castro, J. A. (2016) Functional interaction of the ankylosing spondylitis-associated endoplasmic reticulum aminopeptidase 2 with the HLA-B*27 peptidome in human cells. *Arthritis Rheumatol.* **68,** 2466–2475
25. Heyder, T., Kohler, M., Tarasova, N. K., Haag, S., Rutishauser, D., Rivera, N. V., Sandin, C., Mia, S., Malmstrom, V., Wheelock, A. M., Wahlstrom, J., Holmdahl, R., Eklund, A., Zubarev, R. A., Grunewald, J., and Ytterberg, A. J. (2016) Approach for Identifying Human Leukocyte Antigen (HLA)-DR Bound Peptides from Scarce Clinical Samples. *Mol. Cell. Proteomics* **15,** 3017–3029
26. Ciudad, M. T., Sorvillo, N., van Alphen, F. P., Catalan, 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
27. Klatt, M. G., Kowalewski, D. J., Schuster, H., Di Marco, M., Hennenlotter, J., Stenzl, A., Rammensee, H. G., and Stevanovic, S. (2016) Carcinogenesis of renal cell carcinoma reflected in HLA ligands: A novel approach for synergistic peptide vaccination design. *Oncoimmunology* **5,** e1204504
28. Zhou, F. (2009) Molecular mechanisms of IFN-gamma to up-regulate MHC class I antigen processing and presentation. *Int. Rev. Immunol.* **28,** 239–260
29. 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 Computational Biol.* **13,** e1005725
30. Dudley, M. E., Gross, C. A., Langhan, M. M., Garcia, M. R., Sherry, R. M., Yang, J. C., Phan, G. Q., Kammula, U. S., Hughes, M. S., Citrin, D. E., Restifo, N. P., Wunderlich, J. R., Prieto, P. A., Hong, J. J., Langan, R. C., Zlott, D. A., Morton, K. E., White, D. E., Laurencot, C. M., and Rosenberg, S. A. (2010) CD8+ enriched "young" tumor infiltrating lymphocytes can mediate regression of metastatic melanoma. *Clin. Cancer Res.* **16,** 6122–6131
31. Donia, M., Larsen, S. M., Met, O., and Svane, I. M. (2014) Simplified protocol for clinical-grade tumor-infiltrating lymphocyte manufacturing with use of the Wave bioreactor. *Cytotherapy* **16,** 1117–1120
32. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372
33. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526
34. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–740
35. Jurtz, V. I., 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. *bioRxiv* **199,** 3360–3368
36. Andreatta, M., Alvarez, B., and Nielsen, M. (2017) GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45,** 458–463
37. Thomsen, M. C., and Nielsen, M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40,** W281–W287
38. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009). Improved visualization of protein consensus sequences by iceLogo. *Nature methods* **6,** 786–787
39. 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
40. Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., and Fugmann, T. (2016) High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics* **16,** 1570–1580
41. Ternette, N., Yang, H., Partridge, T., Llano, A., Cedeno, S., Fischer, R., Charles, P. D., Dudek, N. L., Mothe, B., Crespo, M., Fischer, W. M., Korber, B. T., Nielsen, M., Borrow, P., Purcell, A. W., Brander, C., Dorrell, L., Kessler, B. M., and Hanke, T. (2016) Defining the HLA class I-associated viral antigen repertoire from HIV-1-infected human cells. *Eur. J. Immunol.* **46,** 60–69
42. Pymm, P., Illing, P. T., Ramarathinam, S. H., O'Connor, G. M., Hughes, V. A., Hitchen, C., Price, D. A., Ho, B. K., McVicar, D. W., Brooks, A. G., Purcell, A. W., Rossjohn, J., and Vivian, J. P. (2017) MHC-I peptides get out of the groove and enable a novel mechanism of HIV-1 escape. *Nature Structural Mol. Biol.* **24,** 387–394
43. 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
44. 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. U.S.A.* **103,** 14889–14894
45. Hassan, C., Kester, M. G. D., de Ru, A. H., Hombrink, P., Drijfhout, J. W., Nijveen, H., Leunissen, J. A. M., Heemskerk, M. H. M., Falkenburg, J. H. F., and van Veelen, P. A. (2013) The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell. Proteomics* **12,** 1829–1843
46. Alpizar, A., Marino, F., Ramos-Fernandez, A., Lombardia, 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
47. Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H. G., and Stevanovic, 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
48. van Haren, S. D., Herczenik, E., ten Brinke, A., Mertens, K., Voorberg, J., and Meijer, A. B. (2011) HLA-DR-presented peptide repertoires derived from human monocyte-derived dendritic cells pulsed with blood coagulation factor VIII. *Mol. Cell. Proteomics* **10,** M110-002246
49. Adamopoulou, E., Tenzer, S., Hillen, N., Klug, P., Rota, I. A., Tietz, S., Gebhardt, M., Stevanovic, S., Schild, H., Tolosa, E., Melms, A., and Stoeckle, C. (2013) Exploring the MHC-peptide matrix of central tolerance in the human thymus. *Nat. Commun.* **4,** 2039
50. Costantino, C. M., Spooner, E., Ploegh, H. L., and Hafler, D. A. (2012) Class II MHC self-antigen presentation in human B and T lymphocytes. *PloS one* **7,** e29805
51. Vaux, D. L., Wilhelm, S., and Hacker, G. (1997) Requirements for proteolysis during apoptosis. *Mol. Cell. Biol.* **17,** 6502–6507



52. 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

53. Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S., and Nielsen, M. (2015) Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* **67,** 641–650

54. Fruh, K., and Yang, Y. (1999) Antigen presentation by MHC class I and its regulation by interferon gamma. *Curr. Opin. Immunol.* **11,** 76–81

55. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protocols* **11,** 2301–2319

56. Petretto, A., Carbotti, G., Inglese, E., Lavarello, C., Pistillo, M. P., Rigo, V., Croce, M., Longo, L., Martini, S., Vacca, P., Ferrini, S., and Fabbi, M. (2016) Proteomic analysis uncovers common effects of IFN-gamma and IL-27 on the HLA class I antigen presentation machinery in human cancer cells. *Oncotarget* **7,** 72518–72536

57. Zanker, D., and Chen, W. (2014) Standard and immunoproteasomes show similar peptide degradation specificities. *Eur. J. Immunol.* **44,** 3500–3503

58. Mishto, M., Liepe, J., Textoris-Taube, K., Keller, C., Henklein, P., Weberruss, M., Dahlmann, B., Enenkel, C., Voigt, A., Kuckelkorn, U., Stumpf, M. P., and Kloetzel, P. M. (2014) Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation. *Eur. J. Immunol.* **44,** 3508–3521

59. Escobar, H., Crockett, D. K., Reyes-Vargas, E., Baena, A., Rockwood, A. L., Jensen, P. E., and Delgado, J. C. (2008) Large scale mass spectrometric profiling of peptides eluted from HLA molecules reveals N-terminal-extended peptide motifs. *J. Immunol.* **181,** 4874–4882

60. McMurtrey, C., Trolle, T., Sansom, T., Remesh, S. G., Kaever, T., Bardet, W., Jackson, K., McLeod, R., Sette, A., Nielsen, M., Zajonc, D. M., Blader, I. J., Peters, B., and Hildebrand, W. (2016) Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. *eLife* **5,** e12556

61. Leggatt, G. R., Dunn, L. A., De Kluyver, R. L., Stewart, T., and Frazer, I. H. (2002) Interferon-gamma enhances cytotoxic T lymphocyte recognition of endogenous peptide in keratinocytes without lowering the requirement for surface peptide. *Immunol. Cell Biol.* **80,** 415–424

62. Guillaume, P., Picaud, S., Baumgaertner, P., Montandon, N., Schmidt, J., Speiser, D. E., Coukos, G., Bassani-Sternberg, M., Fillipakopoulos, P., and Gfeller, D. (2017) The C-terminal extension landscape of naturally presented HLA-I ligands. *bioRxiv*

63. 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.* **292,** 5262–5270

64. Uebel, S., Kraas, W., Kienle, S., Wiesmuller, K. H., Jung, G., and Tampe, R. (1997) Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 8976–8981

65. Herget, M., Baldauf, C., Scholz, C., Parcej, D., Wiesmuller, K. H., Tampe, R., Abele, R., and Bordignon, E. (2011) Conformation of peptides bound to the transporter associated with antigen processing (TAP). *Proc. Natl. Acad. Sci. U.S.A.* **108,** 1349–1354

66. Pearson, H., Daouda, T., Granados, D. P., Durette, C., Bonneil, E., Courcelles, M., Rodenbrock, A., Laverdure, J. P., Cote, C., Mader, S., Lemieux, S., Thibault, P., and Perreault, C. (2016) MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Investigation* **126,** 4690–4701

67. Aldous, A. R., and Dong, J. Z. (2017) Personalized neoantigen vaccines: A new approach to cancer immunotherapy. *Bioorg Med Chem* **S0968–0896,** 31220–31228

68. Katsnelson, A. (2016) Mutations as munitions: Neoantigen vaccines get a closer look as cancer treatment. *Nature Med.* **22,** 122–124

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

Chapter 5

Manuscript 2

Integrated Proteogenomic Deep Sequencing and Analytics Accurately Identify Non-Canonical Peptides in Tumor Immunopeptidomes

Chloe Chong, Markus Müller, HuiSong Pak, Dermot Harnett, Florian Huber, Delphine Grun, Marion Leleu, Aymeric Auger, Marion Arnaud, Brian J. Stevenson, Justine Michaux, Ilija Bilic, Antje Hirsekorn, Lorenzo Calviello, Laia Simó-Riudalbas, Evarist Planet, Jan Lubiński, Marta Bryśkiewicz, Maciej Wiznerowicz, Ioannis Xenarios, Lin Zhang, Didier Trono, Alexandre Harari, Uwe Ohler, George Coukos, Michal Bassani-Sternberg

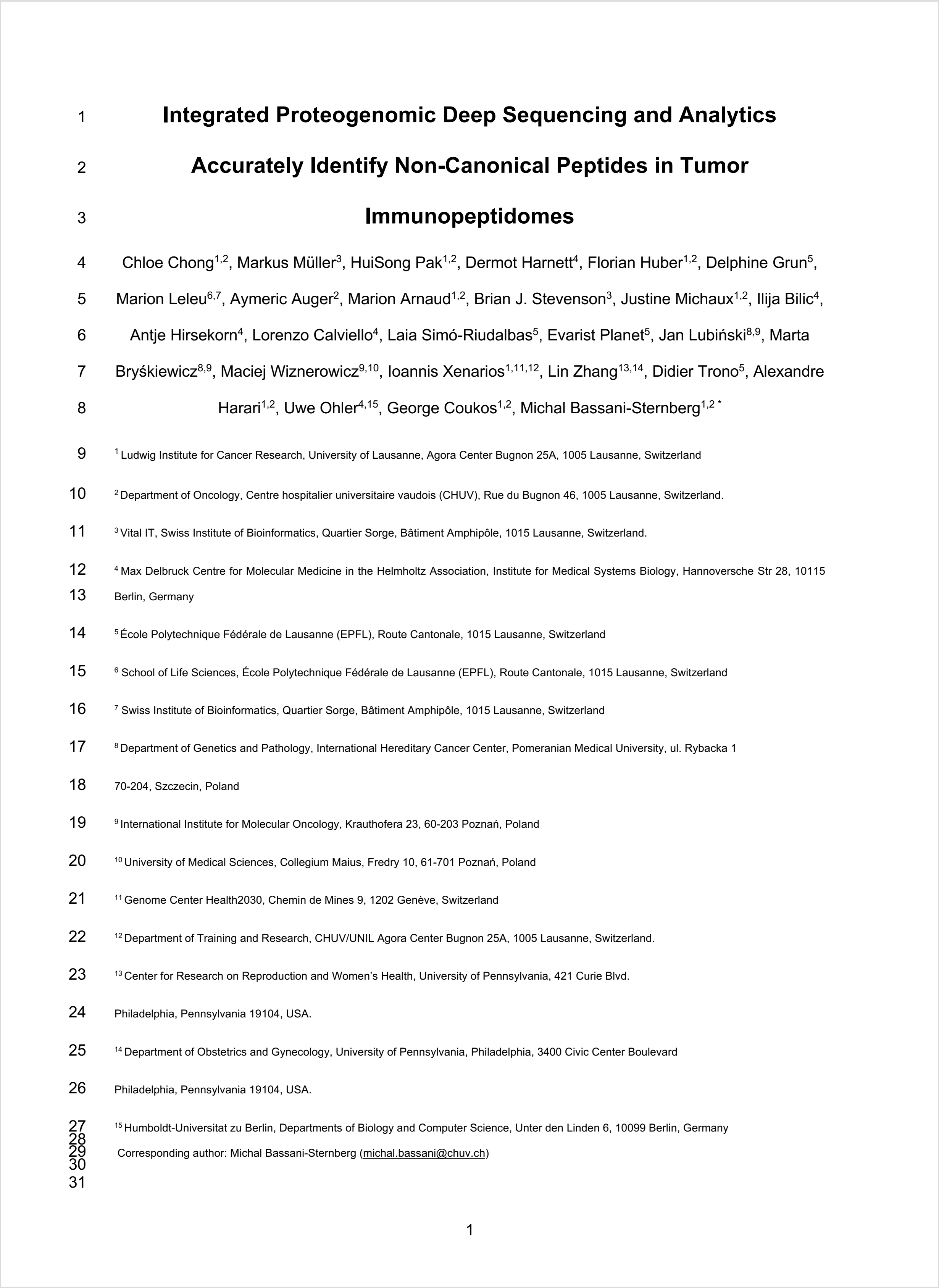
Nature Communications, accepted, February 2020

Chapter 5 MANUSCRIPT 2

This manuscript was accepted by Nature Communications on the 12th February 2020, where I am first author. This project was designed and supported by Dr. Michal Bassani-Sternberg and Prof. George Coukos. I managed the project, performed most of the wet lab, all cell culture, and all MS-based experiments (with guidance from MS expert Dr. Hui Song Pak). I analyzed, coordinated, integrated and interpreted the data, with significant support from all co-authors. Dr. Markus Müller (second author) significantly contributed to the project by developing the computational section of the pipeline and performed the analysis related to ipMSDB.

Overall, the project was a product of a large collaboration endeavor involving the support and expertise from all co-authors (please refer to "Author Contributions" in the attached original article), and the manuscript was written by myself, Dr. Markus Müller and Dr. Michal Bassani-Sternberg.

5.1 Accepted Manuscript



Integrated Proteogenomic Deep Sequencing and Analytics Accurately Identify Non-Canonical Peptides in Tumor Immunopeptidomes

Chloe Chong[1,2], Markus Müller[3], HuiSong Pak[1,2], Dermot Harnett[4], Florian Huber[1,2], Delphine Grun[5], Marion Leleu[6,7], Aymeric Auger[2], Marion Arnaud[1,2], Brian J. Stevenson[3], Justine Michaux[1,2], Ilija Bilic[4], Antje Hirsekorn[4], Lorenzo Calviello[4], Laia Simó-Riudalbas[5], Evarist Planet[5], Jan Lubiński[8,9], Marta Bryśkiewicz[8,9], Maciej Wiznerowicz[9,10], Ioannis Xenarios[1,11,12], Lin Zhang[13,14], Didier Trono[5], Alexandre Harari[1,2], Uwe Ohler[4,15], George Coukos[1,2], Michal Bassani-Sternberg[1,2 *]

[1] Ludwig Institute for Cancer Research, University of Lausanne, Agora Center Bugnon 25A, 1005 Lausanne, Switzerland

[2] Department of Oncology, Centre hospitalier universitaire vaudois (CHUV), Rue du Bugnon 46, 1005 Lausanne, Switzerland.

[3] Vital IT, Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Amphipôle, 1015 Lausanne, Switzerland.

[4] Max Delbruck Centre for Molecular Medicine in the Helmholtz Association, Institute for Medical Systems Biology, Hannoversche Str 28, 10115 Berlin, Germany

[5] École Polytechnique Fédérale de Lausanne (EPFL), Route Cantonale, 1015 Lausanne, Switzerland

[6] School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Route Cantonale, 1015 Lausanne, Switzerland

[7] Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Amphipôle, 1015 Lausanne, Switzerland

[8] Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University, ul. Rybacka 1

70-204, Szczecin, Poland

[9] International Institute for Molecular Oncology, Krauthofera 23, 60-203 Poznań, Poland

[10] University of Medical Sciences, Collegium Maius, Fredry 10, 61-701 Poznań, Poland

[11] Genome Center Health2030, Chemin de Mines 9, 1202 Genève, Switzerland

[12] Department of Training and Research, CHUV/UNIL Agora Center Bugnon 25A, 1005 Lausanne, Switzerland.

[13] Center for Research on Reproduction and Women's Health, University of Pennsylvania, 421 Curie Blvd.

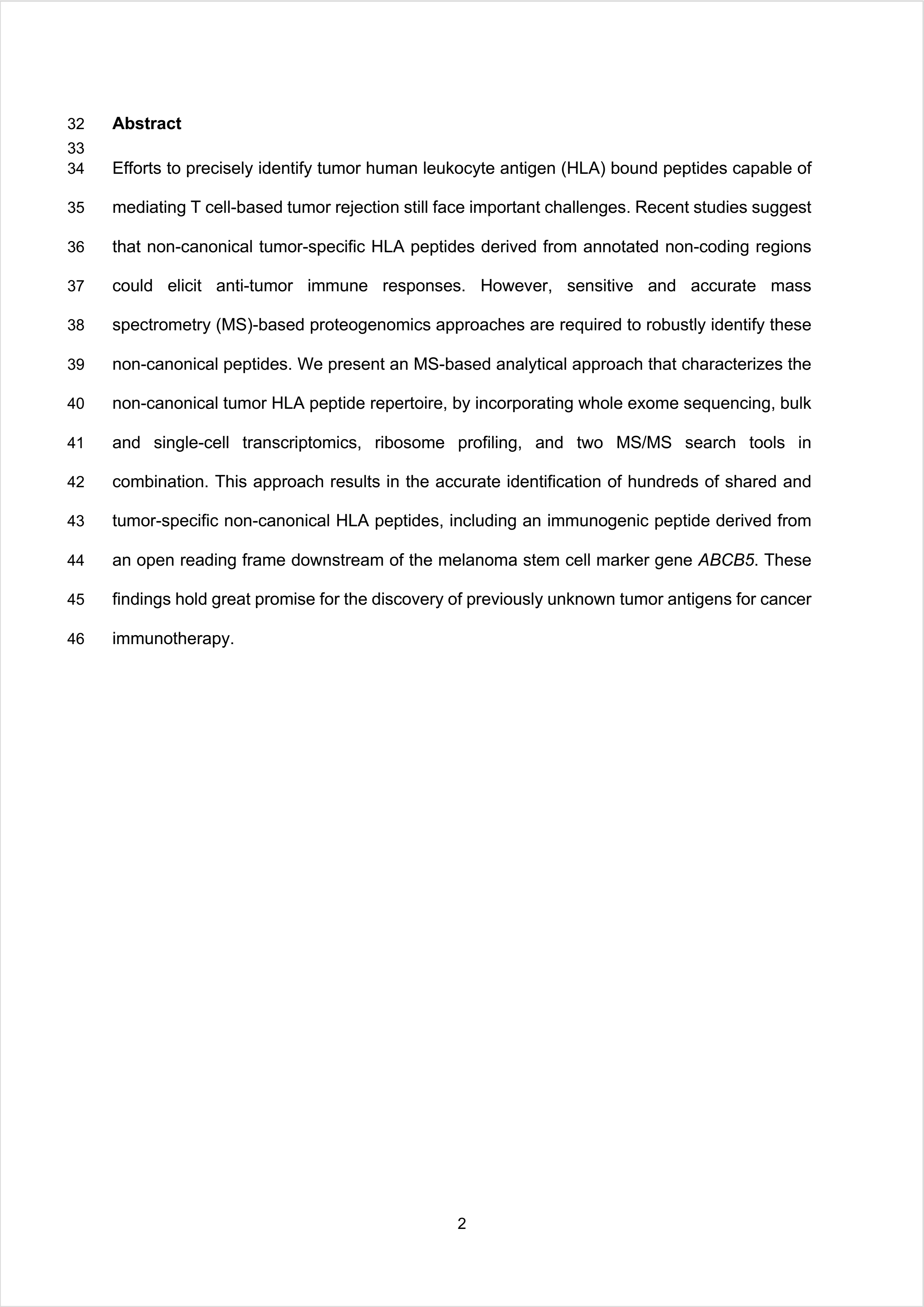
Philadelphia, Pennsylvania 19104, USA.

[14] Department of Obstetrics and Gynecology, University of Pennsylvania, Philadelphia, 3400 Civic Center Boulevard

Philadelphia, Pennsylvania 19104, USA.

[15] Humboldt-Universitat zu Berlin, Departments of Biology and Computer Science, Unter den Linden 6, 10099 Berlin, Germany

Corresponding author: Michal Bassani-Sternberg (michal.bassani@chuv.ch)



Abstract

Efforts to precisely identify tumor human leukocyte antigen (HLA) bound peptides capable of mediating T cell-based tumor rejection still face important challenges. Recent studies suggest that non-canonical tumor-specific HLA peptides derived from annotated non-coding regions could elicit anti-tumor immune responses. However, sensitive and accurate mass spectrometry (MS)-based proteogenomics approaches are required to robustly identify these non-canonical peptides. We present an MS-based analytical approach that characterizes the non-canonical tumor HLA peptide repertoire, by incorporating whole exome sequencing, bulk and single-cell transcriptomics, ribosome profiling, and two MS/MS search tools in combination. This approach results in the accurate identification of hundreds of shared and tumor-specific non-canonical HLA peptides, including an immunogenic peptide derived from an open reading frame downstream of the melanoma stem cell marker gene *ABCB5*. These findings hold great promise for the discovery of previously unknown tumor antigens for cancer immunotherapy.

Introduction

The efficacy of T cell-based cancer immunotherapy relies on the recognition of HLA-bound peptides (HLAp) presented on the surface of cancer cells. Characterizing and classifying immunogenic epitopes is an ongoing endeavour for developing cancer vaccines and adoptive T cell-based immunotherapies. Neoantigens, peptides derived from mutated proteins, are absolutely tumor-specific yet mostly patient-specific, and are implicated in the efficacy of checkpoint blockade immunotherapy[1-4]. In contrast to tumor-specific private neoantigens, tumor-associated antigens (TAAs) that are shared across patients may be more attractive for immunotherapy due to the more efficient and rapid treatment of a greater number of patients [5-7]. Recent studies have focused on the discovery of non-canonical antigens, which are antigens derived from the aberrant translation of presumed non-coding transcripts and/or the aberrant or deregulated transcription of non-coding genomic regions, UTR, or genomically altered frames. Such aberrant transcription and translation events lead to the generation of peptide sequences that are missing in conventional protein sequence repositories[8,9]. If such translation events lead to the presentation of tumor-specific and immunogenic HLA ligands, these occurrences could substantially expand the repertoire of targetable epitopes for cancer immunotherapy[8-19]. Currently, approximately 1% of the entire genome is annotated as protein-coding regions, yet 75% of the genome can be transcribed and theoretically translated, potentially offering a pool of previously unexplored peptide targets[20].

To date, mass spectrometry (MS) is the only analytical methodology that allows the direct identification of the HLAp repertoire *in vivo*[21]. Often, MS-based immunopeptidomic discoveries are limited to the standard, available protein sequence database, usually containing only annotated proteome-derived sequences. Recently, several studies have included protein sequences derived from the translation of transcripts identified from RNA-Seq, or ribosome profiling, in MS-based searches[8,22-27]. Overall, these studies warrant further development regarding many key aspects. Importantly, elevated false discovery rates (FDRs) for the non-canonical space can occur when MS reference data are populated with polypeptide sequences derived from all potential three- or six-frame translations of transcribed regions[28]. Several studies did not compute FDRs or apply sample-specific thresholds for FDR calculations[23,27]. Furthermore, rigorous experimental confirmations of such non-canonical sequences by targeted MS is currently lacking. Additionally, current workflows often introduce a risk of bias by pre-filtering peptide identifications based on HLA-binding predictions[23,27]. Due to the above limitations and to an *a priori* restriction of the search space to tumor-specific non-canonical polypeptide sequences[23], the overall biogenesis of non-canonical HLA binding peptides (noncHLAp) remains to date understudied.

Here, we describe a proteogenomic approach that allows identifying tumor-specific noncHLAp derived from the translation of presumed non-coding transcripts, such as from (long) non-coding genes (lncRNAs), pseudogenes, untranslated regions (UTRs) of coding genes, and transposable elements (TEs). We performed immunopeptidomics analyses while integrating tumor exome, bulk and single-cell transcriptome (scRNA-Seq), and whole translatome data. We then implemented NewAnce, a new analytical approach for non-canonical element identification that combines two MS/MS search tools, along with group-specific FDR calculations to identify noncHLAp. Together, this approach unveiled a large number of unique noncHLAp, highlighting the potential of this approach to increase the range of targetable epitopes in cancer immunotherapy.

Results

A comprehensive strategy for noncHLAp identification

MS-based immunopeptidomics was performed on seven patient-derived melanoma cell lines and two pairs of lung cancer samples with matched normal tissues **(Fig. 1a)**, which resulted in the identification of 60,320 unique proteome-derived HLA class-I bound peptides (protHLAIp) and 11,256 proteome-derived HLA class-II bound peptides (protHLAIIp). For the exploration and identification of non-canonical peptides presented naturally *in vivo*, whole exome and RNA-Seq data were generated from all samples **(Fig. 1a and Supplementary**

Data 1). We inferred expression of presumed non-coding genes, such as lncRNAs, pseudogenes and other non-protein-coding genes, from individual samples' RNA-Seq data. In addition, we applied an analytical pipeline to assign TE-derived RNA-Seq reads to single loci (see Methods section for more details), resulting in expression data for transcribed TEs. All three forward open reading frames (ORFs) (stop-to-stop) in the above transcripts were subsequently *in silico* translated into polypeptide sequences. For every sample, the polypeptide sequences were concatenated to personalized canonical proteome references containing allelic variant information from patient tumor exome data. Finally, we searched the MS immunopeptidomics data against these personalized reference databases.

Database size affects false positives in noncHLAp detection

In silico translation of transcripts in three forward reading frames results in a large number of potential polypeptide sequences. In proteogenomics, searching MS data against such inflated protein reference databases may propagate false positives[28,29]. Hence, our first investigative step was to understand the impact of database size on the level of false positives in immunopeptidomics datasets. We searched reference databases containing canonical (i.e., UniProt) and our non-canonical polypeptide sequences with a single search tool (MaxQuant) and at a global 1% FDR. The accuracy was assessed by assigning HLA-binding prediction scores to the MS-identified peptides with MixMHCpred[30]. We reasoned that non-canonical HLA class I bound peptides (noncHLAIp) should follow the same binding rules as protHLAIp[31]. First, we compared a generic non-canonical protein sequence database derived from the three forward frame (three-frame) translation of all non-coding transcripts from ENCODE[32] with a sample-specific protein sequence database derived from the three-frame translation of lncRNAs and pseudogenes from the RNA-Seq data using an expression cut-off value of > 0 fragments per kilobase of transcript per million mapped reads (FPKM). Additional databases of decreasing size were assembled by retaining only the sequences that originated from more highly expressed genes (FPKM > 2, > 5 or > 10). Reducing the size of the database by personalizing and focusing on highly expressed genes led to an increase in the percentage of noncHLAIp that were predicted to bind to their respective HLA alleles (MixMHCpred p-value ≤ 0.05) **(Fig. 1b)**. Restricting the database to polypeptide sequences originating from highly expressed genes should, on the one hand, improve the accuracy of MS-based non-canonical peptide identification, while on the other hand, lead to the potential loss of peptides encoded by lower-expressing transcripts. Hence, in this study, we included all non-canonical transcripts with FPKM > 0 to circumvent the need to exclude polypeptide sequences based on low expressing transcripts.

NewAnce improves the accurate identification of noncHLAp

We developed the computational module NewAnce, which combines the MS search tools MaxQuant[33] and Comet[34], with the implementation of a group-specific strategy for the FDR calculation (see Methods section for more information and **Supplementary Fig. 1a-g** for performance evaluation). All HLAp identified by either of the search tools were consequently matched against an up-to-date UniProt/TrEMBL sequence database (95,106 protein sequences of the human reference proteome (up000005640), with isoforms) to extract noncHLAp that do not map back to known human proteins in UniProt. For every sample, FDRs were calculated separately for protHLAp and noncHLAp **(Fig. 1c and Supplementary Fig. 1a)**. Only consensus (intersection) peptide-spectrum matches (PSMs) from Comet and MaxQuant were retained for further downstream analyses. Estimating the FDR after retaining the intersection is challenging. Nevertheless, most false positive PSMs are specific to one search tool, and the remaining decoys in NewAnce indicated an estimated FDR of < 0.001%.

With NewAnce, the number of protHLAIp identified across 11 samples ranged from 3,490 to 16,672 per sample, and from 817 to 5,777 for protHLAIIp **(Supplementary Data 2)**. Furthermore, up to 148 noncHLAIp per individual sample were identified with NewAnce, with a combined total of 452 unique noncHLAIp **(Supplementary Data 2 and Supplementary Data 3)**. Of note, noncHLAp are defined here as the peptides derived from either non-protein-coding genes, such as lncRNAs and pseudogenes, or TEs. As the majority of the non-protein-coding genes were lncRNAs, these will be henceforth collectively termed lncRNAs. Among the four HLA-II expressing samples investigated, only 4 non-canonical HLA class II bound peptides (noncHLAIIp) were detected out of 11,256 protHLAIIp. Re-searching the 2,597 PSMs of identified noncHLAIp against the human proteome UniProt database concatenated with the list of non-canonical peptide sequences and allowing identification of alternative sequences including six common modifications (see Methods section) revealed a very low level of ambiguity (**Supplementary Data 4 and 5**).

We employed two complementary methods to assess the accuracy of our approach. First, we predicted the binding of peptides to their respective HLA allotypes. Across all 11 samples, 90% of the noncHLAIp and 91% of the protHLAIp identified with NewAnce were predicted to bind the HLA allotypes (median values, **Supplementary Fig. 1h**). As expected, NewAnce detected fewer HLAp than Comet (PSM FDR of 3%) or MaxQuant (PSM FDR of 3%), while with a more routinely applied FDR of 1% using MaxQuant alone, more HLAp were obtained with NewAnce **(Fig. 2a, Supplementary Fig. 1i-l and Supplementary Data 2)**. Importantly, for the noncHLAIp repertoire (lncRNAs and TEs), significantly higher percentages of peptides predicted to bind the HLA allotypes were identified by NewAnce than MaxQuant or Comet alone **(Fig. 2b and Supplementary Fig. 1i-l)**.

In addition, we correlated the observed mean retention time (RT) of a given peptide against the calculated hydrophobicity index (HI), which corresponds to the percentage of acetonitrile at which the peptide elutes from the analytical HPLC system. Calculating the sequence-specific hydrophobicity of peptides identified by NewAnce by SSRCalc[35] showed that the RT distribution of non-canonical peptides was on the diagonal line, and was not significantly different from the distribution of proteome-derived peptides, supporting their correct identification **(Fig. 2c)** (one-sided F-test p-value: 1.0e+0). However, a significant difference in RT distribution was observed when comparing non-canonical peptides identified by NewAnce to those identified by MaxQuant (one-sided F-test p-value: 6.3e−32) or Comet alone (one-sided F-test p-value: 8.4e−20) **(Fig. 2d)**. Similar results were obtained for all investigated samples **(Supplementary Fig. 2)**.

A common approach to boosting non-canonical peptide identifications is searching the MS data with a single tool (or the union of two tools) while applying a permissive FDR followed by an additional step of filtering to include only peptides predicted to bind the relevant HLA allotypes[36]. To evaluate this approach, we compared the correlation between the HIs and RTs of predicted non-canonical HLA binders and non-binders identified at 3% PSM FDR with either MaxQuant **(Fig. 2e)** or Comet **(Fig. 2f)**. Predicted binders showed better correlations between the HI and RT than non-binders (one-sided F-test p-values: 8.4e-6 for MaxQuant and 4.4e-18 for Comet). These correlations were fairly poor for MaxQuant, while a much better correlation was calculated for Comet, likely due to the conservative group-specific FDR control strategy we applied for Comet.

Notably, when examining the source protein sequence origins of all noncHLAIp, we detected an enrichment towards the C-termini of their precursor protein sequences. This effect was also observed for protHLAIp originating from similarly short canonical proteins **(Supplementary Fig. 3a-b)**.

Targeted-MS and Ribo-Seq confirm noncHLAIp detection

To experimentally validate the NewAnce computational pipeline, we investigated a selection of NewAnce-identified HLAp from a melanoma sample (0D5P) with targeted MS-based analyses. All identified noncHLAIp (lncRNAs and TEs, n=93), as well as a similarly sized subset of protHLAIp from clinically relevant TAAs (n=71) detected in 0D5P, were synthesized in their heavy isotope-labelled forms for MS-targeted validation. The selected TAAs were chosen solely based on their interesting tumor-associated biological functions, such as known cancer/testis or melanoma antigens. Here, MS-based targeted confirmation by parallel reaction monitoring (PRM) was directly compared between the non-canonical and proteome-derived peptide groups by spiking the heavy labelled peptides into multiple independent replicates of 0D5P immunopeptidomic samples, revealing that protHLAIp confirmation was superior to that of noncHLAIp (78.5% for TAAs versus 55.2% for lncRNAs and 27.7% for TEs) **(Fig. 3a, Supplementary Data 6 and Supplementary Data 7)**. We also observed that the

PRM validation was dependent on the source RNA expression level **(Supplementary Fig. 4a-d)**, measured peptide intensities **(Supplementary Fig. 4e-h)**, and detectability by MS/MS across multiple 0D5P replicates **(Supplementary Fig. 4i-l)**.

To further validate the noncHLAIp with an additional targeting strategy, we analyzed sample 0D5P also by Ribo-Seq, which involves the sequencing of ribosome protected fragments (RPFs). Periodic RPF distributions (see Methods section) that supported translation from the correct ORFs of the transcripts encoding the identified noncHLAIp were observed for 22.2% of the TE peptides and 21.3% of the lncRNA peptides, compared to 100% of the TAAs **(Fig. 3b)**. Notably, nine lncRNA HLAIp and two TE peptides were validated independently by both the PRM and Ribo-Seq approaches. For example, the noncHLAIp SYLRRHLDF was confirmed by MS **(Fig. 3c)**, and the translated ORF that generated the peptide was mapped back to two non-coding RNA transcripts **(Fig. 3d-e)**.

Low RNA expression limits noncHLAIp presentation

We then characterized the expression levels of source RNAs encoding HLAIp in more depth. For this purpose, we compared all identified source genes of protHLAIp to source genes of noncHLAIp in the 0D5P sample. The protein-coding source genes had a median FPKM value of 9.3, whereas the presumed non-coding source genes showed lower expression overall, with a median FPKM of 2.1 **(Fig. 4a-b)**. Generally, higher numbers of unique peptides identified per gene were correlated with higher expression levels. PRM-validated noncHLAIp covered a large dynamic range of gene expression, and interestingly, a few were confirmed at very low source RNA expression levels **(Fig. 4c-d)**.

The low expression levels of source genes that generated noncHLAp prompted us to investigate the regulation of non-canonical HLA presentation and whether their expression can be induced or enhanced with drug treatments. We treated melanoma cells with either decitabine (DAC), a DNA methyltransferase inhibitor, known to reactivate epigenetically silenced genes, or with interferon gamma (IFNγ), known to upregulate antigen presentation[37-

[40]. As expected, we observed large quantitative changes in the presentation of protHLAIp when T1185B melanoma cells were treated with IFNγ. Specifically, we found enhanced presentation of peptides derived from immune-related genes, likely due to their high gene expression and increased production of HLA-I molecules **(Supplementary Fig. 4m)**. However, no obvious change was observed for the noncHLAIp repertoire, with 60% of the identified noncHLAIp remaining unaltered by IFNγ treatment, suggesting that transcription is the limiting step in the presentation of noncHLAIp or that transcription of noncHLAIp is not affected generally by IFNγ **(Supplementary Fig. 4n)**. Furthermore, we explored the effect of the hypomethylating agent DAC on noncHLAIp in melanoma. Although DAC induced the expression of selected hypomethylating-induced immune genes[41], TAAs and non-coding transcripts **(Supplementary Fig. 4o-q)**, changes in the 0D5P noncHLAIp repertoire were modest. Nonetheless, we identified and confirmed the presence of a unique DAC-induced noncHLAIp derived from a lncRNA **(Supplementary Fig. 4r)**.

Ribo-Seq improves the coverage of protHLAIp and noncHLAIp

Next, we hypothesized that immunopeptidomes would correlate more closely with translatome than transcriptome data. To build the translatome-based database for the MS search, all ORFs showing periodic RPF distribution were extracted for the 0D5P sample, and translated *in silico*. This technique reduced the size of the search space, and we used this independent discovery method in our study to identify additional noncHLAIp, including those derived from previously unexplored ORFs in coding genes.

We investigated the extent by which a protein sequence database inferred by Ribo-Seq could replace or complement the search performed with our personalized references comprising canonical protein sequences concatenated with polypeptide sequences from the three-frame translation of expressed non-coding transcripts. Using 0D5P as a representative immunopeptidomic dataset, we observed a positive correlation between RNA expression and HLAIp-sampling (see Methods section) searched against the personalized protein sequence database (r= 0.392) **(Fig. 4e)**. Then, we searched the same immunopeptidomics MS data against the *de novo*-assembled Ribo-Seq inferred database, and we correlated this HLAIp-sampling with RNA abundance **(Fig. 4f)** or with translation rates based on the spectral coefficient of the 3-periodic signal in the Ribo-Seq data (see Methods section) **(Fig. 4g)**. This approach resulted in a significantly higher positive correlation between HLAIp-sampling searched against the Ribo-Seq inferred database and the translation rate (r=0.574) than the overall RNA abundance (r=0.431, two-sided p-value< 10e-16). Thus, evidence exists that the immunopeptidome, at least for the 0D5P sample, is better captured by the translatome than the transcriptome.

Notably, restricting the database to actual translation products by Ribo-Seq provided a deeper coverage of the immunopeptidome than a canonical protein sequence database **(Fig. 4h)**. This enhanced coverage led to the identification of additional noncHLAIp derived mainly from ORFs that are not included in canonical annotation but still showing periodic footprint of translation, such as those originating in 5' or 3' UTRs, presumed non-coding RNAs, retained introns, and pseudogenes. The majority of those identified were derived from either upstream ORFs or other un-annotated ORFs **(Fig. 4i-j)**. Many of these additional noncHLAIp were missed using the RNA-Seq inferred database. Of note, this method also takes into account products arising from ribosomal frameshifting, which could be relevant in the context of non-canonical antigens[42]. Interestingly, only 16 common lncRNA-derived noncHLAIp were found when comparing both strategies, which likely reflects the limited detection of periodic Ribo-Seq reads in transcripts with low expression or low mappability **(Supplementary Fig. 5a-c)**.

scRNA-Seq reveals heterogeneity in presumed non-coding genes

Tumor cell heterogeneity could be a key factor underlying immune escape, leading to the inefficacy of cancer immunotherapies. To understand the pattern of non-coding gene expression at the single-cell level, we performed scRNA-Seq on the 0D5P melanoma cell line. Overall, 1,400 cells were sequenced at a total depth of 176 million reads, resulting in the detection of a median of 6,261 genes per cell (total of 19,178 detected genes). As expected, clustering of 0D5P cells revealed dependency on **the** cell cycle status **(Fig. 5a),** and thus we explored whether source genes associated with the cell cycle **(Fig. 5b-c)**. Then, we confirmed that the antigen presentation machinery as well as many of the selected TAAs were uniformly expressed in all cells and were thus independent of the cell cycle, as expected **(Fig. 5d)**.

Out of the 71 presumed non-coding source genes identified by bulk transcriptomics, 35 were also detected at the single-cell level **(Fig. 5d)**. HLAIp derived from presumed non-coding source genes detected with higher coverage at the single-cell level were also those confirmed by PRM (6 out of 8 genes confirmed in > 50% cells and 14 out of 27 genes confirmed in < 50% cells) and by Ribo-Seq (37 out of 41 genes confirmed in > 50% cells and 25 out of 46 genes confirmed in < 50% cells). Importantly, source non-coding genes clearly showed expression heterogeneity: nearly none of them were uniformly expressed across cells, although the limited sensitivity of scRNA-Seq could account for this variation. The expression of *LINC00520* was higher than expected given its detection in only 75% of cells, suggesting that it is not uniformly expressed **(Fig. 5d)**. Sufficient expression level in a subset of cells would allow the sampling for HLA presentation of overall lowly expressed genes and eventually their detection in the immunopeptidome.

We thus sought to explore the cell subset by identifying known biomarker genes co-expressed with *LINC00520* **(Fig. 5e-h)**. Interestingly, we found that *LINC00520* was co-expressed with the ATP-binding cassette sub-family B member 5 (*ABCB5*) gene **(Fig. 5g).** The ABCB5 mediates chemotherapy drug resistance in stem-like tumor cell subpopulations in human malignant melanoma and is commonly over-expressed in circulating melanoma tumor cells[43], together with beta-catenin (CTNNB1), a key regulator of melanoma cell growth[44], and with its critical downstream target microphthalmia-associated transcription factor (MITF) which mediates melanocyte differentiation[45] **(Fig. 5h)**. *ABCB5* was detected in 37% of 0D5P cells **(Fig. 5f),** which also coexpressed *LINC005520***.** Importantly, we also detected a noncHLAIp epitope encoded by a previously unknown ORF embedded within the *ABCB5* gene, which as shown below is immunogenic. The detection of such non-canonical neoantigens on subsets of melanoma cells with regenerative or metastatic potential could prove highly interesting in the context of immunotherapy.

Identification of tumor-specific noncHLAIp

As our initial MS search space was not restricted to polypeptide sequences derived from tumor-specific transcripts, we retrospectively investigated the potential of identified noncHLAIp to be classified as tumor-specific. A public database of RNA sequencing data from 30 different healthy tissues (Genotype-Tissue Expression, GTEx[46]) was assessed at a strict 90th percentile, which sets the expression of a gene at the top 10% of its expression across all samples. We identified 335 noncHLAIp from 280 lncRNA genes in the seven melanoma samples, of which 23% were expressed in only any of our tumor samples and not in the healthy tissues (excluding testis due to its immunoprivileged nature) **(Fig. 6)**. Among these genes was the tumor-specific *LINC00518* gene, which has been proposed as a two-gene classifier for melanoma detection together with the TAA *PRAME*[47].

Using an in-house curated inventory of human TE-derived polypeptide sequences (from three-frame translations) as a reference, we found 88 unique TE-HLAIp in our whole dataset. Some were derived from autonomous TEs, such as long tandem repeat (LTR) retrotransposons and long interspersed nuclear elements (LINEs), and others were derived from non-autonomous retrotransposons such as short interspersed nuclear elements (SINE) and SINE-VNTR-Alu (SVA) elements **(Supplementary Fig. 6a)**. Importantly, 60 of the 88 TE-HLAIp were found in presumed non-coding TE regions and therefore represent previously unknownHLA peptides. For example, peptides derived from AluSq2 SINE/Alu and L1PA16 LINE/L1 elements were expressed in only skin and testis. These TE-HLAIp would have been overlooked in canonical MS-based searches.

We next examined whether our approach could identify tumor-specific non-canonical targets in the ideal case in which normal and tumor biopsies are available, i.e., from the two lung cancer patient samples included in the present dataset. For the C3N-02671 lung tumor sample, 21 noncHLAp were detected by MS; however, none of the peptides were tumor-specific. In the C3N-02289 sample, we identified 45 noncHLAIp by MS **(Fig. 7a)**, among which 10 peptides were identified uniquely in the tumor tissue. Four of these source genes were also entirely absent at the RNA level in the adjacent lung **(Fig. 7b)**. Interestingly, the noncHLAIp from *RP11-566H8.3* was also testis-specific in the GTEx database (90th percentile transcripts per million (TPM) ≤ 1) **(Fig. 7b)**, thus qualifying as a non-canonical cancer testis antigen.

The same analyses of TE genes in lung tumor sample C3N-02289 resulted in the identification of one LTR7B LTR/ERV1 TE-HLAIp that was present in the tumor tissue; however, this gene is also expressed in healthy brain. For sample C3N-02671, no TE-derived HLAIp were detected.

To comparatively assess the expression of canonical tumor antigens in the same samples, we investigated select TAAs using the same methodology **(Supplementary Fig. 6b-c)**. We identified six TAA protHLAIp that were exclusively detected in the tumor tissue of C3N-02289 (BIRC5, TERT, FAP, SPAG4, MAGEA9 and BCL2L1). We also detected uniquely in C3N-02671 tumor tissue two protHLAIp TAAs (CCND1 and PXDNL) and five protHLAIIp TAAs (MMP2 and CEACAM5).

NoncHLAIp are shared across patient samples

We investigated the prevalence of shared noncHLAIp among the nine tumor samples analysed. We identified 27 peptides that were detected in at least two patient samples. Seven noncHLAIp, already validated in 0D5P, were confirmed by PRM in at least one other patient sample that expressed HLA allotypes with identical or highly similar binding specificities **(Supplementary Table 1)**, with a total of 15 individually detected PRM events **(Fig. 8a)**. Interestingly, one noncHLAIp, VTDQASHIY, derived from microcephalin-1 antisense RNA (*MCPH1-AS1*), was independently confirmed with PRM in three melanoma or lung cancer patients **(Supplementary Fig. 7a-b)**. Further, the shared presentation of the noncHLAIp AAFDRAVHF, derived from the family of LINEs (LINE/L2) on chromosome 6, was confirmed in two melanoma samples **(Supplementary Fig. 7c-d)**. Interestingly, the corresponding source RNA expression is restricted to the skin and testis.

Next, we assessed a large collection of immunopeptidomic datasets (ipMSDB[48], 91 biological cancer tissue/cell line sources, 35 biological healthy tissues/cell line sources; 1,102 MS raw files in total) and obtained the first large-scale signature of noncHLAIp presentation **(Supplementary Data 8)**. In total, 220,293 peptides were obtained from healthy samples versus 280,385 peptides from cancer samples. We re-identified in ipMSDB 92 tumor-specific noncHLAIp (source genes described above and were expressed at 90th percentile TPM ≤ 1 in a maximum of 3 tissues) **(Fig. 8b)**, 60 of which were only detected in cancer immunopeptidome samples. From those,14 were detected in at least one additional cancer sample in ipMSDB. Overall, noncHLAIp presentation showed a trend of enrichment in cancer samples in ipMSDB **(Fig. 8c)**. Interestingly, two noncHLAIp from the lncRNA *HAGLROS* (KVLAGTVLFK and VLAGTVLFK), identified specifically in the lung cancer tissue in our samples, were exclusively found only in cancer samples in ipMSDB, mainly in ovarian cancer samples, consistent with a previous report[49].

Immunogenicity of noncHLAIp with autologous T cells

The involvement of noncHLAIp in tumor immune recognition was assessed by measuring IFNγ release upon peptide stimulation of autologous tumor-infiltrating lymphocytes (TILs) or peripheral blood mononuclear cells (PBMCs) from the same patients. Out of the 786 peptides screened (94 TEs, 421 lncRNAs, 56 alternative ORFs and 215 TAAs), we confirmed the specific recognition by autologous TILs of TAAs, such as the HYYVSMDAL and RLPSSADVEF peptides from tyrosinase (TYR) and RYNADISTF from tyrosinase-related protein 1 (TYRP1) in melanoma sample 0D5P, and of the YLEPGPVTA peptide from the promelanosome protein (PMEL) in melanoma sample T1015A. One non-canonical peptide, KYKDRTNILF, derived from the downstream ORF (dORF) of the melanoma stem-cell marker *ABCB5* gene in 0D5P, was also found to be immunogenic in both autologous CD8+ TILs and

CD8+ T cells from peripheral blood lymphocytes (PBLs) **(Fig. 9a-c)**. Notably, this peptide was shared across three additional melanoma samples in ipMSDB.

Discussion

Our proteogenomics approach led to the stringent identification of hundreds of noncHLAIp derived from presumed non-coding genes, TEs and alternative ORFs. This feat was achieved with NewAnce, a computational module that overcomes the challenge of reduced sensitivity and specificity when searching against large MS search spaces and it can be applied to any (non-canonical) protein sequence database of interest [28,50]. We rigorously tested the validity of noncHLAIp identifications with HLA-binding predictions, sequence-specific retention characteristics, and targeted MS analyses, and provided evidence of translation in peptide-encoding ORFs by Ribo-Seq. Using all of these strategies together, we confirmed that NewAnce was superior to MaxQuant and Comet alone across all the investigated samples. Taking one patient as an example, we conducted PRM and Ribo-Seq analyses to compare a subset of protHLAIp to non-canonical antigen classes (lncRNAs and TEs), thereby validating the identified noncHLAIp at the experimental level. We found that noncHLAIp had an overall lower confirmation rate than protHLAIp, possibly due to their lower expression, which also led to their stochastic detection by MS. Interestingly, the expression and translation of microproteins derived from presumed non-coding RNAs in the heart were recently discovered using a Ribo-Seq directed proteogenomics approach. Similar to our results, evidence of translation was confirmed for 22.5% of the lncRNAs, while 55.4% of the micropeptides were validated by PRM MS[51]. Importantly, our results additionally demonstrate that the correct identification of noncHLAIp in proteogenomic workflows requires proper FDR control and validation using multiple independent methods.

Combining immunopeptidomics with RNA-Seq and Ribo-Seq datasets enables the comprehensive assessment of how transcription, translation and HLA presentation are correlated. Despite the different methodological challenges[52-54], we observed the expected correlations between HLA presentation level and expression, especially by Ribo-Seq, presumably because translation is biologically closer to antigen processing and presentation than transcription is. In addition, we found that in the melanoma sample 0D5P, most of the noncHLAIp derived from the Ribo-Seq inferred database originated from source genes harbouring upstream ORFs (uORFs). Notably, uORFs can trigger the non-sense-mediated decay of mRNAs and provide a rich source of noncHLAIp[55-57].

While a previous study showed that the presentation of non-canonical peptides was enhanced by inflammatory stimuli, the presentation of only specific HLA peptides was documented[58]. In contrast, our large-scale analysis of both DAC- and IFNγ-treated cells did not detect profound changes in noncHLAIp presentation, although non-coding source genes were induced. Hence, we hypothesize that low copy numbers of such noncHLAIp remain a limiting factor of their presentation. Moreover, corroborating prior research[27], we report the enrichment of noncHLAIp originating from the C-termini of source protein sequences. Translation products of such presumed non-coding regions could be considered defective ribosomal products that are expected to be unstable and rapidly degraded, likely bypassing the proteasome[59].

Given the lack of 'complete' tissue immunopeptidomics reference libraries from healthy donors, we propose a workflow to retrospectively search for tumor-specific non-coding source genes with publicly available RNA-Seq databases (such as GTEx[46]). We observed that 23% of the source non-canonical genes were not expressed in healthy tissues (with our selected thresholds), and could be considered tumor-specific. However, in the ideal situation in which both tumor and matched normal tissues were available, we found that the majority of peptides were detected in both, suggesting that the comparison with GTEx overestimates the fraction of true tumor-specific non-canonical ligands, and that some might be patient-specific. Interestingly, two overlapping epitopes were identified in the lncRNA *HAGLROS*, which were expressed and presented uniquely in the lung tumor tissue. This lncRNA has been implicated in cancer progression[60,61] and should be prioritized for downstream validation. Moreover, while Laumont et al.[23] first proposed the existence of shared noncHLAIp, our work validates that noncHLAIp can be shared across multiple tumor samples, and we anticipate better treatment efficacy with such shared noncHLAIp compared to that achieved with private neoantigens[62,63].

The expression of tumor-specific noncHLAp in a subpopulation of tumor cells suggests a dependency on a molecular or functional state. For example, the immunogenic noncHLAIp derived from the dORF in the *ABCB5* gene was moderately expressed in only 37% of the melanoma cells compared to the expression of the *TYR* and *TYRP1* genes, both of which were highly and uniformly expressed and produced confirmed immunogenic epitopes. Immune pressure on selected tumor cell subsets with particular biological relevance—-such as cancer stem-like cells, tumor cells with epithelial-mesenchymal transition features and proliferating tumor cells-—could greatly impact tumor behavior and be clinically beneficial.

Indeed, we found such an immunogenic noncHLAIp from 0D5P derived from the dORF of the *ABCB5* gene. ABCB5 has been shown to be expressed in malignant melanoma-initiating cells and is thought to be responsible for both the progression and chemotherapeutic refractoriness of advanced malignant melanoma[43]. Through an IL1β/IL8/CXCR1 cytokine signalling circuit, ABCB5 has been shown to control IL1β secretion and maintain slow cycling and chemoresistance[64]. Blockage of ABCB5 reversed resistance to multiple chemotherapeutic agents, induced cellular differentiation and impaired tumor growth *in vivo*[64]. We found that *ABCB5* was differentially co-expressed in a cluster of 0D5P cells with the transcription factor *MITF* and *CTNNB1*, whose expression may be enriched in melanoma stem cell populations[65]. The presence of spontaneous specific T cells recognizing the noncHLAIp derived from the dORF of the *ABCB5* gene in both peripheral blood and TILs suggests no central tolerance and that this target could allow immune targeting of the melanoma stem cell subpopulation to curtail tumor growth.

Out of more than 500 noncHLAIp screened, immune recognition by rapidly expanded TILs and PBMCs was detected for only a single immunogenic noncHLAIp. Various mechanisms could account for such lack of recognition by autologous T cells. First, we were able to screen only autologous TILs that had long propagated in culture. We previously reported that TIL *ex vivo* expansion may lead to depletion of T cell clones that recognize tumor neoantigens[66]. Second, it is possible that the melanoma cells, which had to be expanded considerably in culture for immunopeptidomics analyses, could have undergone an alteration of their HLA peptide repertoire, leading to the identification of noncHLAIp that were not originally present in the freshly extracted cells. However, we also assessed snap-frozen lung cancer tissues and still did not observe the immune recognition of identified non-canonical targets in autologous PBMCs. Alternatively, the ability of noncHLAIp to induce a natural immune response might be inferior to that of protHLAp. Low expression might limit the uptake by professional antigen-presenting cells and thus also limit the priming of naïve T cells *in vivo* through cross presentation. Similarly, the engagement of CD4+ T helper cells through HLA class II presentation might also be limited. Nevertheless, tumor-specific non-canonical targets may still be valuable for immunotherapy, even when no prior immune response against the targets has been detected *ex vivo*, as was previously shown for neoantigens[3,4,67]. More research should be performed to thoroughly assess the ability of noncHLAp to augment the protective immune response *in vivo*. Such approaches are supported by evidence in mouse models demonstrating that peptides derived from non-canonical regions can be spontaneously recognized and leveraged in cancer immunotherapy[23,68].

Remarkably, across tumor types, the potential number of predicted noncHLAp is orders of magnitude larger than that of neoantigens encompassing non-synonymous somatic mutations. As T cell-based screenings currently have limited throughput and are expensive[69], an accurate and cost-effective non-canonical target discovery approach is crucial for their further development and use in cancer immunotherapy. With the renewed interest in cancer vaccines and the constantly growing number of antigens screened for immune recognition, we expect that enough training data will become available to allow the development of accurate predictors of immunogenicity. Combining this approach with our developed module NewAnce to shortlist noncHLAp presented *in vivo* and to rank them according to their predicted immunogenicity will facilitate the comprehensive exploration of non-canonical antigens, their association with immune responses and their potential for building effective cancer immunotherapies.

Methods

Patient material

Melanoma cell lines (0D5P, 0MM745, 0NVC) were generated as follows: patient-derived tumors were cut into small pieces before being transferred into a digestion buffer containing collagenase type I (Sigma Aldrich) and DNase I (Roche) for at least one hour. Dissociated cells were washed and maintained in RPMI 1640 + GlutaMAX medium (Life Technologies) supplemented with 10% heat-inactivated FBS (Dominique Dutscher) and 1% Penicillin/Streptomycin Solution (BioConcept). If fibroblasts appeared, they were selectively eliminated with G418 (Geneticin; Gibco) treatment. The primary melanoma cell lines T1185B, T1015A, Me290 and Me275 were generated at the Ludwig Institute for Cancer Research, Department of Oncology, University of Lausanne[70,71]. All established melanoma cells were subsequently grown to 1 x 10^8 cells, collected by centrifugation at 151 x g for 5 min, washed twice with ice cold PBS and stored as dry cell pellets at -20°C until use. For the i*n vitro* 72 h treatment with IFNγ (100 IU/mL, Miltenyi Biotec), T1185B cells were grown to 2 x 10^8 in triplicate. For the treatment with DAC (Sigma-Aldrich), 2 x 10^8 melanoma cells were grown for 8 days in medium containing 0.5 µM DAC, and the drug was readministered on the 4th day.

Autologous TILs were expanded from fresh melanoma tumor samples from patients 0D5P, 0MM745, 0NVC, LAU1185 (tumor cell line T1185B), LAU1015 (tumor cell line T1015A), LAU203 (tumor cell line Me290) and LAU50 (tumor cell line Me275) at the Ludwig Institute for Cancer Research, Department of Oncology, University of Lausanne. The fresh tissues were manually cut into fragments of one to two mm^3. The tumor fragments were then placed in 24-well plates containing RPMI CTS grade (Life Technologies), 10% human serum (Biowest), 0.025 M HEPES (Life Technologies), 55 µmol/L 2-mercaptoethanol (Life Technologies) and supplemented with IL-2 (6,000 IU/mL, Proleukin) for three to five weeks. Following this pre-rapid expansion protocol (REP), TILs were then expanded with another REP as follows: 5x10^6 TILs were stimulated with irradiated feeder cells (Ratio 1:200), anti-CD3 (OKT3, 30 ng/mL, Miltenyi Biotec) and IL-2 (3,000 IU/mL) for 14 days. After 14 days of REP, approximately 2x10^9 TILs were harvested, washed and cryopreserved until use. The purity (i.e., the percentage of CD3 T cells) was > 95%. As an additional control, one flask with the exact same REP conditions without TILs was cultured in parallel, and no cells were detectable at day 14. REP TILs were thawed in 5 IU/mL DNase I (Sigma Aldrich) and cultured in 3000 IU/mL IL-2 for two days in RPMI 1640 medium with GlutaMAX™ Supplement (Gibco), and 8% human serum (Biowest), 10 mM HEPES (Gibco), 50 μM Beta-mercaptoethanol (Gibco), 100 μM non-essential amino acids (Gibco), 100 IU/mL penicillin, 0.1 mg/mL streptomycin, 2 mM L-glutamine (Gibco), 0.1 mg/mL kanamycin sulfate (Carl Roth) and 1 mM sodium pyruvate (Gibco). The cells were then washed twice in complete medium and subsequently rested overnight in the presence of 150 IU/mL IL-2 prior to peptide stimulation.

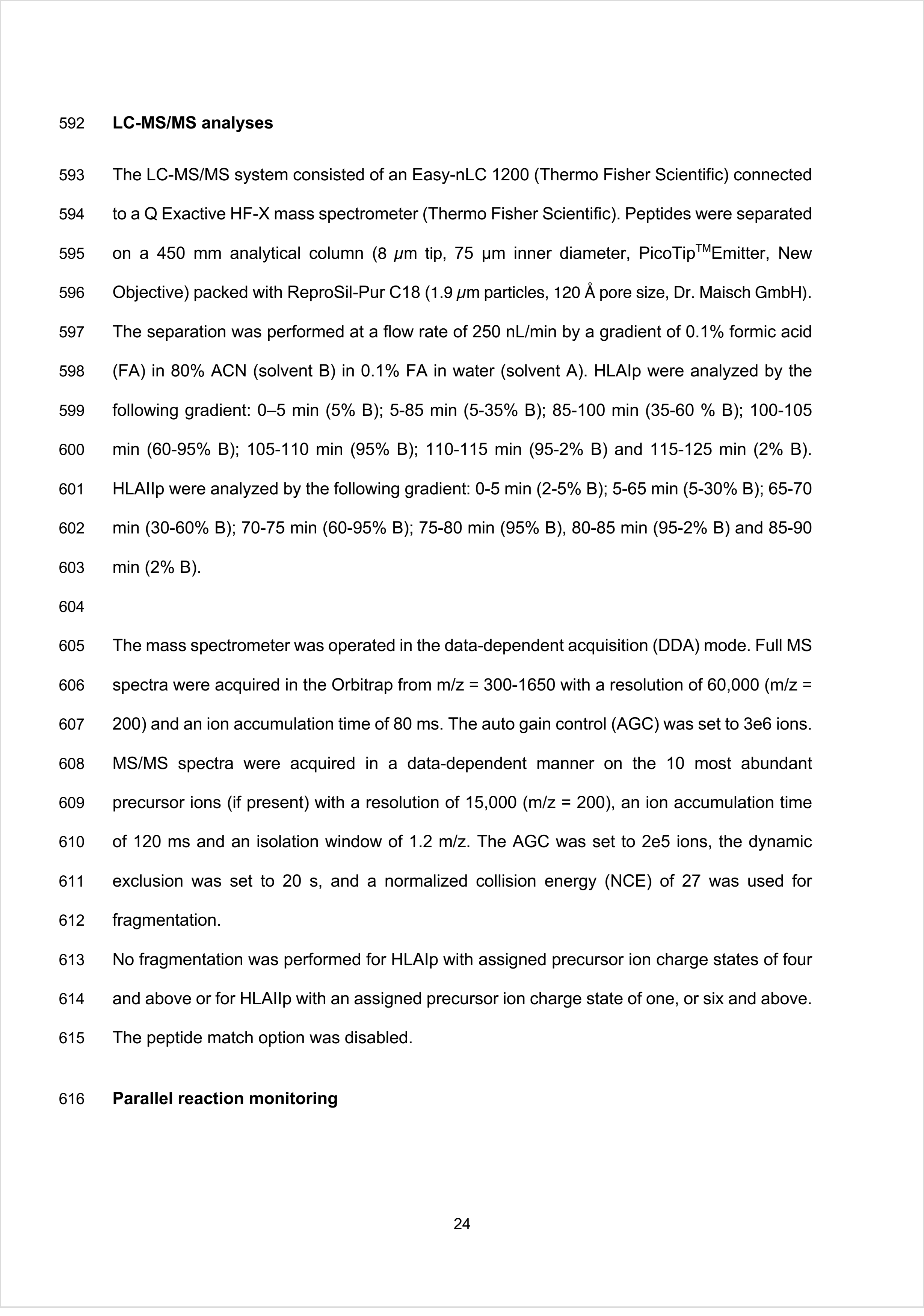
Snap-frozen normal and lung tumor tissue materials from the C3N-02289 (Lung squamous cell carcinoma, grade 2) and C3N-02671 (lung adenocarcinoma, G2) samples were kindly provided by the International Institute of Molecular Oncology. Informed consent was obtained from the participants in accordance with the requirements of the institutional review board (Ethics Commission, CHUV, Bioethics Committee, Poznan University of Medical Sciences, Poznań, Poland).

All cells tested negative for mycoplasma contamination. High-resolution 4-digit HLA-I and HLA-II typing (**Supplementary Data 1**) was performed at either the Laboratory of Diagnostics, Service of Immunology and Allergy, CHUV, Lausanne or in-house using the HLA amplification method with the TruSight HLA v2 Sequencing Panel kit (CareDx) according to the manufacturer's protocol. Sequencing was performed on the Illumina® MiniSeq™ System (Illumina) using a paired end 2x150 bp protocol. The data were analysed with Assign TruSight HLA v2.1 software (CareDx).

Immunoaffinity purification of HLA peptides

We performed HLA immunoaffinity purification according to our previously established protocols[39,72]. W6/32 and HB145 monoclonal antibodies were purified from the supernatants of HB95 (ATCC® HB-95™) and HB145 cells (ATCC® HB-145™) using protein-A sepharose 4B (Pro-A) beads (Invitrogen), and antibodies were then cross-linked to Pro-A beads. Cells were lysed with PBS containing 0.25% sodium deoxycholate (Sigma Aldrich), 0.2 mM iodoacetamide (Sigma Aldrich), 1 mM EDTA, a 1:200 protease inhibitors cocktail (Sigma Aldrich), 1 mM phenylmethylsulfonylfluoride (Roche), and 1% octyl-beta-D glucopyranoside (Sigma Alrich) at 4°C for 1 hour. The lysates were cleared by centrifugation in a table-top centrifuge (Eppendorf) at 4°C for 50 min at 21,191 x g. Snap-frozen tissue samples were homogenized on ice in 3-5 short intervals of 5 s each using an Ultra Turrax homogenizer (IKA) at maximum speed. The lysates were then cleared by centrifugation at 75,600 x g in a high-speed centrifuge (Beckman Coulter, Avanti JXN-26 Series,, JA-25.50 rotor) at 4°C for 50 min. For HLA immunopurification, we employed the Waters Positive Pressure-96 Processor (Waters) and 96-well single-use micro-plates with 3 µm glass fibers and 10 µm polypropylene membranes (Seahorse Bioscience, ref no: 360063). Pan HLA-I and HLA-II antibodies cross-linked to beads were loaded onto separate plates, respectively. For tissue samples, depletion of endogenous antibodies was required with Pro-A beads. The lysates were passed sequentially through the first plate containing pan HLA-I antibody-crosslinked beads, then through the second plate with pan HLA-II antibody-crosslinked beads, at 4°C. The beads in the plates were then washed separately with varying concentrations of salts using the processor. Finally, the beads were washed twice with 2 mL of 20 mM Tris-HCl pH 8.

Sep-Pak tC18 100 mg Sorbent 96-well plates (Waters, ref no: 186002321) were used for the purification and concentration of HLA-I and HLA-II peptides. The C18 sorbents were conditioned, and the HLA complexes and bound peptides were directly eluted from the affinity plate with 1% trifluoroacetic acid (TFA; Sigma-Aldrich). After washing the C18 sorbents with 2 mL of 0.1% TFA, HLA-I peptides were eluted with 28% acetonitrile (ACN; Sigma Aldrich) in 0.1% TFA, and HLA-II peptides were eluted with 32% ACN in 0.1% TFA. Recovered HLA-I and -II peptides were dried using vacuum centrifugation (Concentrator plus, Eppendorf) and stored at -20°C.



LC-MS/MS analyses

The LC-MS/MS system consisted of an Easy-nLC 1200 (Thermo Fisher Scientific) connected to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). Peptides were separated on a 450 mm analytical column (8 *µ*m tip, 75 µm inner diameter, PicoTip^{TM}Emitter, New Objective) packed with ReproSil-Pur C18 (1.9 *µ*m particles, 120 Å pore size, Dr. Maisch GmbH). The separation was performed at a flow rate of 250 nL/min by a gradient of 0.1% formic acid (FA) in 80% ACN (solvent B) in 0.1% FA in water (solvent A). HLAIp were analyzed by the following gradient: 0–5 min (5% B); 5-85 min (5-35% B); 85-100 min (35-60 % B); 100-105 min (60-95% B); 105-110 min (95% B); 110-115 min (95-2% B) and 115-125 min (2% B). HLAIIp were analyzed by the following gradient: 0-5 min (2-5% B); 5-65 min (5-30% B); 65-70 min (30-60% B); 70-75 min (60-95% B); 75-80 min (95% B), 80-85 min (95-2% B) and 85-90 min (2% B).

The mass spectrometer was operated in the data-dependent acquisition (DDA) mode. Full MS spectra were acquired in the Orbitrap from m/z = 300-1650 with a resolution of 60,000 (m/z = 200) and an ion accumulation time of 80 ms. The auto gain control (AGC) was set to 3e6 ions. MS/MS spectra were acquired in a data-dependent manner on the 10 most abundant precursor ions (if present) with a resolution of 15,000 (m/z = 200), an ion accumulation time of 120 ms and an isolation window of 1.2 m/z. The AGC was set to 2e5 ions, the dynamic exclusion was set to 20 s, and a normalized collision energy (NCE) of 27 was used for fragmentation.

No fragmentation was performed for HLAIp with assigned precursor ion charge states of four and above or for HLAIIp with an assigned precursor ion charge state of one, or six and above. The peptide match option was disabled.

Parallel reaction monitoring

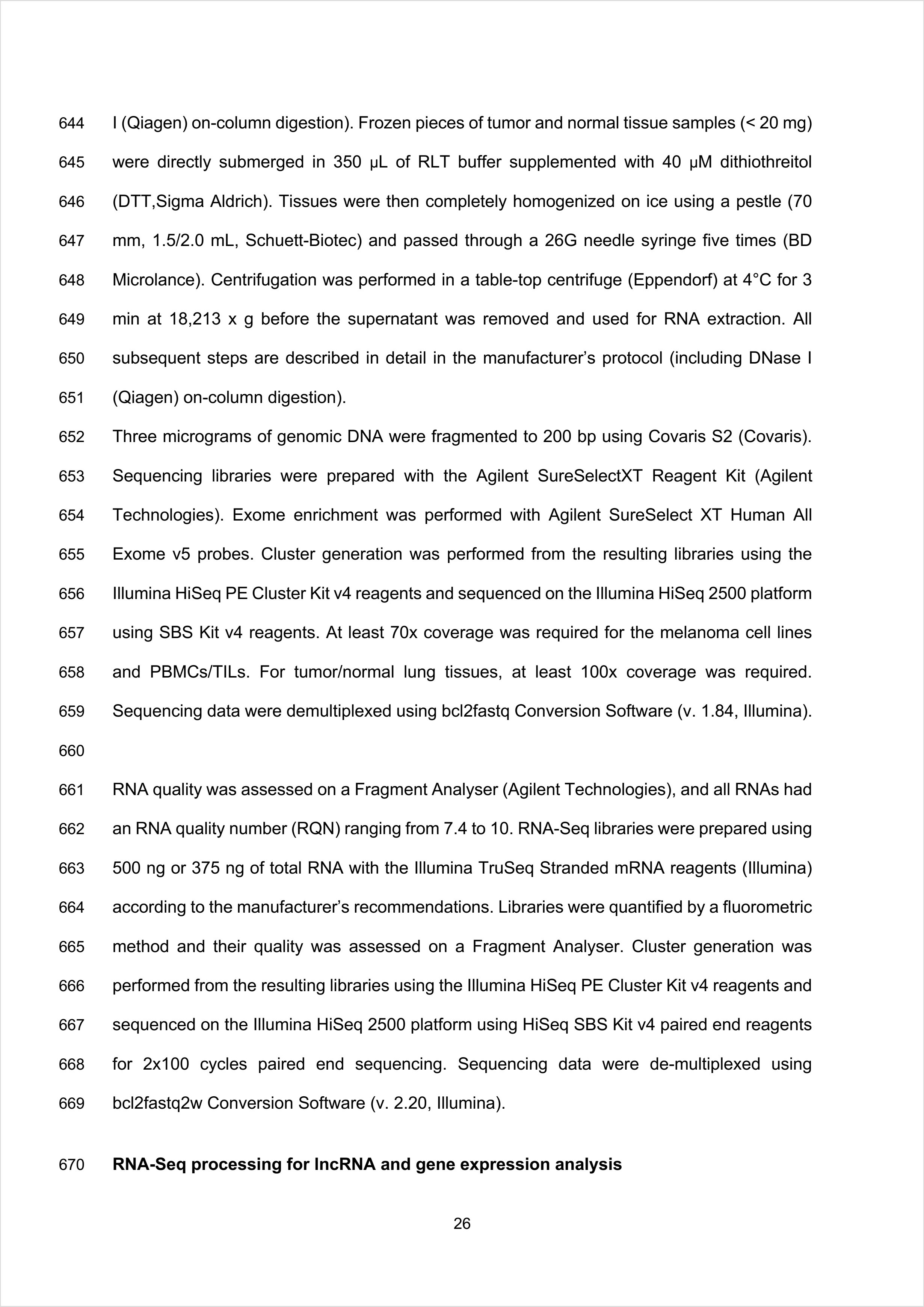
Selected endogenous HLAp that required confirmation by PRM were ordered from Thermo Fisher Scientific as crude (PePotec grade 3) or HPLC grade (purity > 70%) with one stable isotope labelled amino acid. The mass spectrometer was operated at a resolution of 120,000 (at m/z = 200) for the MS1 full scan, scanning a mass range from 300-1650 m/z with an ion injection time of 100 ms and an AGC of 3e6. Then each peptide was isolated with an isolation window of 2.0 m/z prior to ion activation by high-energy collision dissociation (HCD, NCE = 27). Targeted MS/MS spectra were acquired at a resolution of 30,000 (at m/z = 200) with an ion injection time of 60 ms and an AGC of 5e5. Only those peptides that ultimately passed quality control were considered for further downstream analyses by spiking them back into the patient sample.

The PRM data were processed and analysed by Skyline (v4.1.0.18169, MacCoss Lab Software)[73], and an ion mass tolerance of 0.02 m/z was used to extract fragment ion chromatograms. To display MS/MS spectra, raw data were converted into the MGF format by MSConvert (Proteowizard v3.0.18136), and peak lists for the heavy-labelled peptides and light counterparts were extracted. The assessment of MS/MS matching was performed by pLabel (v2.4.0.8, pFind studio, Sci. Ac.) and Skyline.

Exome/RNA sequencing

DNA was extracted for HLA typing and exome sequencing with the commercially available DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturers' protocols. For tissue samples, pelleted DNA was used, which was obtained after lysis of the tissue and centrifugation during HLA immunopurification. The supernatant was used for HLA immunopurification, whereas the pelleted DNA was resuspended in PBS using a pestle (70 mm, 1.5/2.0 mL, Schuett-Biotec) before DNA extraction according to the manufacturer's instructions.

RNA was extracted for RNA sequencing using the Total RNA Isolation RNeasy Mini Kit (Qiagen) according to the manufacturer's protocol for all melanoma cell lines (including DNase



I (Qiagen) on-column digestion). Frozen pieces of tumor and normal tissue samples (< 20 mg) were directly submerged in 350 μL of RLT buffer supplemented with 40 μM dithiothreitol (DTT,Sigma Aldrich). Tissues were then completely homogenized on ice using a pestle (70 mm, 1.5/2.0 mL, Schuett-Biotec) and passed through a 26G needle syringe five times (BD Microlance). Centrifugation was performed in a table-top centrifuge (Eppendorf) at 4°C for 3 min at 18,213 x g before the supernatant was removed and used for RNA extraction. All subsequent steps are described in detail in the manufacturer's protocol (including DNase I (Qiagen) on-column digestion).

Three micrograms of genomic DNA were fragmented to 200 bp using Covaris S2 (Covaris). Sequencing libraries were prepared with the Agilent SureSelectXT Reagent Kit (Agilent Technologies). Exome enrichment was performed with Agilent SureSelect XT Human All Exome v5 probes. Cluster generation was performed from the resulting libraries using the Illumina HiSeq PE Cluster Kit v4 reagents and sequenced on the Illumina HiSeq 2500 platform using SBS Kit v4 reagents. At least 70x coverage was required for the melanoma cell lines and PBMCs/TILs. For tumor/normal lung tissues, at least 100x coverage was required. Sequencing data were demultiplexed using bcl2fastq Conversion Software (v. 1.84, Illumina).

RNA quality was assessed on a Fragment Analyser (Agilent Technologies), and all RNAs had an RNA quality number (RQN) ranging from 7.4 to 10. RNA-Seq libraries were prepared using 500 ng or 375 ng of total RNA with the Illumina TruSeq Stranded mRNA reagents (Illumina) according to the manufacturer's recommendations. Libraries were quantified by a fluorometric method and their quality was assessed on a Fragment Analyser. Cluster generation was performed from the resulting libraries using the Illumina HiSeq PE Cluster Kit v4 reagents and sequenced on the Illumina HiSeq 2500 platform using HiSeq SBS Kit v4 paired end reagents for 2x100 cycles paired end sequencing. Sequencing data were de-multiplexed using bcl2fastq2w Conversion Software (v. 2.20, Illumina).

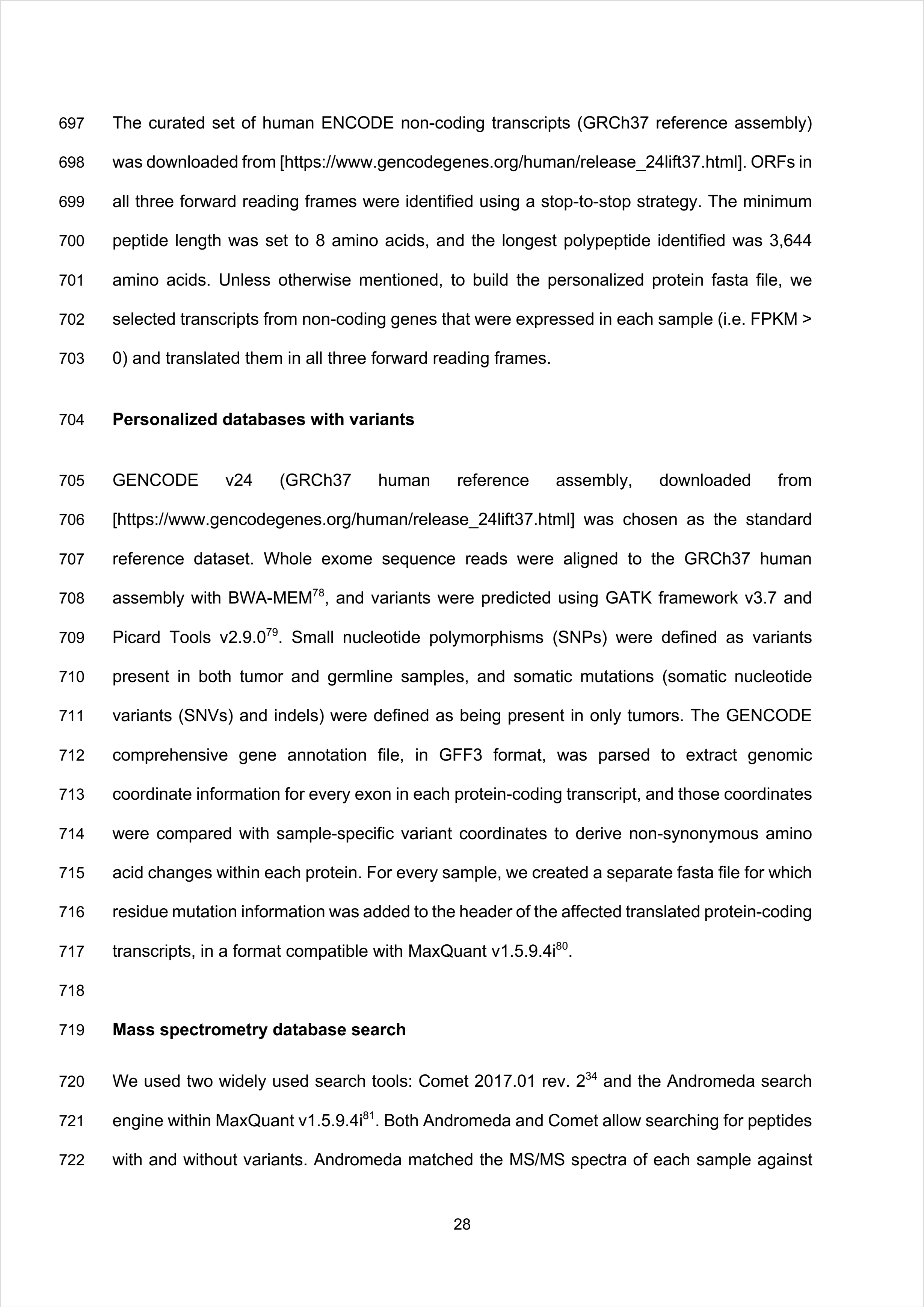
RNA-Seq processing for lncRNA and gene expression analysis

The GENCODE comprehensive gene annotation version 221.2 was downloaded from the GENCODE website [https://www.gencodegenes.org/releases/22.html] and used to define the protein-coding and non-coding gene features, including chromosome position, transcript structure, and transcript and protein sequences. Here, the human reference genome GRCh38/hg38 was downloaded from the UCSC Genome Browser website [http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/] and used as the genome assembly. The RNA-Seq reads were aligned to the GRCh38/hg38 reference genome using RNA-Star (v2.4.2a; [https://github.com/alexdobin/STAR]). Gene expression was normalized and calculated as fragments per kilobase of transcript per million mapped reads (FPKM) values by Cufflinks (v2.2.1) ([http://cole-trapnell-lab.github.io/cufflinks/releases/v2.2.1/]). The gene-level RNA expression data for both protein-coding and non-coding genes were used for downstream gene expression analysis[32,74].

RNA-Seq data processing for TE expression analysis

We developed an analytical pipeline that was capable of assigning TE-derived RNA-Seq reads to single loci in more than 95% of the cases. Reads from the investigated samples and public data from GTEx were mapped to the human (GRCh37) genome using hisat2 v.2.1.0[75]. Counts on genes and TEs were generated using featureCounts 1.6.2[76]. To avoid read assignment ambiguity between genes and TEs, a gtf file containing both was provided to featureCounts. For repetitive sequences, an in-house curated version of the Repbase database was used (fragmented LTR and internal segments belonging to a single integrant were merged). Only uniquely mapped reads were used for counting genes and TEs. Finally, features that did not have at least one sample with 20 reads were discarded from the analysis. Normalization for sequencing depth was performed for both genes and TEs using the Trimmed Mean of M values (TMM) method as implemented in the limma v.3.36.5 package of Bioconductor[77] and using the counts on genes as the library size.

Personalized sequence databases from non-coding transcripts



The curated set of human ENCODE non-coding transcripts (GRCh37 reference assembly) was downloaded from [https://www.gencodegenes.org/human/release_24lift37.html]. ORFs in all three forward reading frames were identified using a stop-to-stop strategy. The minimum peptide length was set to 8 amino acids, and the longest polypeptide identified was 3,644 amino acids. Unless otherwise mentioned, to build the personalized protein fasta file, we selected transcripts from non-coding genes that were expressed in each sample (i.e. FPKM > 0) and translated them in all three forward reading frames.

Personalized databases with variants

GENCODE v24 (GRCh37 human reference assembly, downloaded from [https://www.gencodegenes.org/human/release_24lift37.html] was chosen as the standard reference dataset. Whole exome sequence reads were aligned to the GRCh37 human assembly with BWA-MEM[78], and variants were predicted using GATK framework v3.7 and Picard Tools v2.9.0[79]. Small nucleotide polymorphisms (SNPs) were defined as variants present in both tumor and germline samples, and somatic mutations (somatic nucleotide variants (SNVs) and indels) were defined as being present in only tumors. The GENCODE comprehensive gene annotation file, in GFF3 format, was parsed to extract genomic coordinate information for every exon in each protein-coding transcript, and those coordinates were compared with sample-specific variant coordinates to derive non-synonymous amino acid changes within each protein. For every sample, we created a separate fasta file for which residue mutation information was added to the header of the affected translated protein-coding transcripts, in a format compatible with MaxQuant v1.5.9.4i[80].

Mass spectrometry database search

We used two widely used search tools: Comet 2017.01 rev. 2[34] and the Andromeda search engine within MaxQuant v1.5.9.4i[81]. Both Andromeda and Comet allow searching for peptides with and without variants. Andromeda matched the MS/MS spectra of each sample against the personalized reference libraries (mentioned above). Similarly, the variants were annotated in the PEFF format [http://www.psidev.info/peff] for Comet. Both search tools were run with the same principal search parameters: precursor mass tolerance 20 ppm, MS/MS fragment tolerance of 0.02 Da, peptide length of 8-15 when searching only HLA-I peptides and 8-25 for both HLA-I and HLA-II peptides and no fixed modifications. For samples 0D5P, 0NVC and 0MM745, oxidation (M) and phosphorylation (STY) were set as variable modifications; for the remaining samples only oxidation (M) was included as a variable modification. A PSM FDR of 3% was used for Andromeda as a first filter, and non-canonical reference sequences were loaded into the "proteogenomics fasta files" module for FDR calculations for proteome-derived and non-canonical sequences. For each spectrum the annotated PSMs with the highest score were kept (including the decoy hits calculated by Andromeda from reversed protein sequences) and stored in binary files.

To assure that non-canonical peptide sequences did not match other protein-coding genes, all peptides found by Andromeda or Comet were aligned against an up-to-date UniProt/TrEMBL sequence database (95,106 protein sequences of the human reference proteome up000005640, with isoforms, downloaded 26/09/2018) using an algorithm built in NewAnce. Leucine and iso-leucines were treated as equal since they are not distinguishable by MS. If peptides were found to match standard UniProt sequences, they were assigned as proteome-derived with the UniProt IDs. However, we retained non-canonical TE peptide sequences that matched annotated TEs that were integrated into the human reference in UniProt.

Comet PSMs were read from Comet pep.xml files and all peptides were aligned against the UniProt database as described above. Equivalent to the Andromeda PSMs processing, PSM were annotated and the highest scoring PSMs were stored in binary files. Comet PSM processing was implemented in Java and utilizes the MzJava class library[82]. As described in detail below, it consisted of two main steps: first, three Comet scores *XCorr, deltaCn* and *spScore* and the spectrum charge were combined, and second, the FDR was calculated separately for proteome-derived and non-canonical peptides. The first step boosted the overall number of identified PSMs at a given global FDR, whereas the second step limited the number of false positives in the group of non-canonical peptides at a given global FDR.

All PSMs resulting from the Comet binary files were split into three sublists with PSMs of charge (*Z*) 1 (applicable to HLAIp only), 2, and charge 3 or higher. Further, the three Comet scores *XCorr*, *deltaCn* and *spScore* were considered (the '*expect*' score was left out because it depends on the size of the sequence database). In order to calculate the FDR for 3D vectors $\mathbf{x} = (XCorr, deltaCn, spScore)$, the 3D spaces (one 3D space per charge state *Z*) were partitioned into small cells with 40 intervals in each dimension (**Supplementary Fig. 1a**). The PSMs in the sublist of charge *Z* were then parsed and for each cell, the number of wrong hits (n_0) was set to the number of decoy PSMs in that cell, and the number of true hits (n_1) was set to the number of target (non-decoy) PSMs minus n_0. The 3D probability distributions were estimated by dividing the counts in each cell by the total counts summed over all cells resulting in a distribution for each charge state *Z* for true ($p(\mathbf{x}|Z, H = 1)$) and wrong ($p(\mathbf{x}|Z, H = 0)$) PSMs. In order to obtain smoother distributions, both true (n_1) and decoy (n_0) counts were averaged over a 9-cell nearest neighborhood. This 3D histogram based approach has the advantage that it does not require strong assumptions about the shape of the probability distributions, and in contrast to 1D projection methods, it does take into account the full 3D structure of the score space. On the other hand, it requires fairly large datasets with more than 100'000 PSMs.

The local FDR (*lFDR*) is the probability that a PSM within a given cell is wrong, whereas the global FDR is the probability that a PSM in the final result list from all cells is wrong. It has been shown that *lFDR* calculation provides the most sensitive decision boundaries while controlling the global FDR[83]. Mathematically, $lFDR(\mathbf{x}, Z)$ values for a score vector **x** and charge *Z* can be calculated by Equation (1):

(1)

$$lFDR(\mathbf{x},Z)=\frac{\pi_0 p(\mathbf{x}|Z,H=0)}{\pi_0 p(\mathbf{x}|Z,H=0)+\pi_1 p(\mathbf{x}|Z,H=1)}=\left(1+\frac{\pi_1}{\pi_0}\cdot\frac{p(\mathbf{x}|Z,H=1)}{p(\mathbf{x}|Z,H=0)}\right)^{-1}$$

$$=\left(1+\frac{\pi_1}{\pi_0}\gamma(\mathbf{x},Z)\right)^{-1}$$

where π_0 and π_1 are the class probabilities for true (*H*=1) and wrong (*H*=0) PSMs, and $p(\mathbf{x}|Z,H=0{,}1)$ are the probability distributions as described above. Finally, the *lFDR* threshold was adjusted to yield a global FDR of 3% and all PSMs within cells with *lFDR* values smaller than this threshold were added to the list of PSMs. **Supplementary Fig. 1b** shows a comparison of this 3D histogram approach to a simpler 1D method, where only the *XCorr* score was used, for the 0D5P sample. At the same FDR of 3%, the 3D histogram approach was able to boost the number of unique peptides for both proteome-derived and non-canonical peptides by 22% and 13%, respectively. Importantly, the percentage of predicted HLA binders and the standard error in hydrophobicity index calculation by SSRCalc remained unchanged (**Supplementary Fig. 1c-d**), indicating that the 3D method used in NewAnce did not inflate the error. However, **Supplementary Fig. 1c** also reveals that the percentage of predicted binders is low for the group of non-canonical peptides (only 55% compared to 95% for proteome-derived peptides), indicating a large portion of wrong PSMs in the non-canonical group. This phenomenon has been reported before[28] and is due to a misbalance of true and false hits in the non-canonical sequence databases. Non-canonical sequence databases are typically very large, and they contain mostly sequences that have low probability to contribute to true hits.This causes a strong prevalence for wrong PSMs or a low π_1/π_0 ratio (π_1/π_0 ratio is the total number of true PSMs divided by the total number of wrong PSMs) compared to the proteome-derived database.

In order to tackle this problem, we implemented an approach that estimates the *lFDR* values separately for non-canonical and proteome-derived PSM groups. Since there are only several hundreds of non-canonical PSMs, we could not use the 3D histogram approach directly for the non-canonical PSM group. Instead, we assumed that the probability distributions $p(\mathbf{x}|Z, H = 0,1)$ are the same for non-canonical and proteome-derived PSMs, and that the π_1/π_0 ratios strongly depend on the PSM group. The π_1/π_0 ratios are global measures and can be readily calculated with a few hundred PSMs. Therefore, we first calculated the probability ratios $\gamma(\mathbf{x}, Z)$ for each cell using all PSMs and then calculated the π_1/π_0 ratios separately for the non-canonical and the proteome-derived groups. We then plugged the group specific π_1/π_0 ratios into Equation (1) and obtained a group-specific calculation of the *lFDR* for each cell. The low π_1/π_0 ratio in the non-canonical group will increase the *lFDR* values for this group. When the *lFDR* threshold was adjusted to yield a global FDR of 3%, less but higher quality non-canonical PSMs passed this filter. **Supplementary Fig. 1b** shows that the number of passing non-canonical peptides (3D, 2 Groups) dropped to 28% compared to the number of peptides identified without group adjustment (3D, 1 Group), whereas the number of proteome-derived peptides increased slightly by 8%. However, the percentage of predicted binders among the passing non-canonical peptides (3D, 2 Groups) increased to 85% (**Supplementary Fig. 1c**) and the standard error of HI decreased significantly (**Supplementary Fig. 1d**).

Even if this group specific *lFDR* calculation improved the accuracy of non-canonical PSMs, the fairly low percentage of predicted binders indicated that there was still a larger error in this group. In order to discard more of this residual error, we combined the Comet and Andromeda search results and only the intersection, i.e. PSMs with identical Comet and Andromeda matches (same peptide sequence with the same identification) were retained. As shown in **Supplementary Fig. 1e-g**, this additional filter further reduced the number of non-canonical PSMs, but significantly increased the percentage of predicted binders to 97.3% and decreased the standard error of hydrophobicity index. Without the post-processing of Comet results performed in NewAnce, this improvement would not be possible. When only considering the *XCorr* score and without group specific *lFDR* calculation, combining Comet and MaxQuant would yield more peptides, but with significantly lower percentage of predicted binders (87.3%) and almost double the standard error of HI (1D, 1 Group in green color compared with NewAnce 3% FDR in gray color in **Supplementary Fig. 1e-g**). Using a FDR threshold of 1% instead of 3% for MaxQuant and Comet in NewAnce would only reduce the number of peptides but not increase the percentage of predicted binders, or decrease the standard error of HI, thus justifying our choice of utilizing a 3% FDR threshold (NewAnce 3% FDR in gray color compared with NewAnce 1% FDR in beige color in **Supplementary Fig. 1e-g**).

In order to assign peptides into source protein groups, we implemented a greedy bipartite graph protein grouping algorithm[84]. The total and 'unique' peptide counts were calculated for each protein. To calculate the adjusted peptide counts we sorted the proteins in each group by decreasing number of peptides and for each protein removed the peptides of all proteins higher up in the list.

In order to test the robustness of our approach, the 2,597 PSMs of identified noncHLAp were re-searched against the human reference proteome UniProt database concatenated with the list of non-canonical peptide sequences, including six common modifications. The variable modifications included were 15.9949 Da for oxidation on M, 42.010565 Da for acetylation on the N-terminus, 79.966331 Da for phosphorylation on STY, 119.004099 Da for cysteinylation, 0.98402 Da for deamidation NQ and 57.021464 Da for carbamidomethyl on C. Comet was employed (same parameters as above, but no FDR) to investigate whether PSMs would better fit another possible proteome-derived (modified) sequence based on *XCorr*. The results are reported in **Supplementary Data 4 and 5**.

To build the ipMSDB database, we searched 1,102 immunopeptidomic raw files with Comet (PSM FDR of 1%, as described above), and the Apache Spark cluster computing framework[85] was used to process the results and calculate the FDR. The samples were annotated with basic biological information for further statistical analysis.

Ribo-Seq: experimental protocol

Ribo-Seq was performed according to Calviello et al. 2016[86]. Ribo-Seq libraries were derived from adherent melanoma 0D5P cells that were 80% confluent in 10 cm tissue culture dishes. After washing with ice-cold PBS supplemented with 100 µg/mL cycloheximide (Sigma Aldrich), the cells were immediately snap-frozen by placement in liquid nitrogen followed by placement on wet ice. A lysis buffer containing 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM MgCl2, 1 mM DTT (Sigma Aldrich), 100 µg/mL cycloheximide, 1% (v/v) Triton X-100 (Calbiochem) and 25 U/mL TURBO DNase (Life Tech) in a volume of 400 µL was immediately added to the frozen cells. The cells and buffer were then scraped off, mixed by pipetting, transferred to Eppendorf tubes and lysed on ice for 10 min. The lysate was then titurated by passage through a 26-G needle 10 times with a 1 mL syringe and cleared by centrifugation at 20,000 x g for 10 min at 4°C. The cleared supernatant was then transferred to a pre-cooled tube on ice, and footprinting was performed by adding 1000 U of RNase I (Life Tech. #AM2295) per 400 µL of lysate and incubating in a thermomixer set at 23°C, while shaking at 500 rpm for 45 min. The digestion was stopped by adding 13 µL of SUPERASE-In (Thermo, 20 U/µL) per 400 µL of lysate.

Ribosomes were recovered using two MicroSpin S-400 HR columns (GE Healthcare) per sample. The columns were first equilibrated with a total of 3 mL of buffer containing 20 mM Tris-Cl pH 7.4, 150 mM NaCl, 5 mM MgCl2 and 1 mM DTT by performing 6 rounds of washes with 500 µL of the buffer. The resin was resuspended with the last wash and drained by centrifugation for 4 min at 600 x g. One-half of the sample volume was then filtered per column for 2 min at 600 x g, and the filtered halves were then combined. To the combined flow-through, three volumes of TRIzol LS (Life Tech) were added and RNA was extracted using the Direct-zol RNA Mini-Prep kit (Zymo Research) according to the manufacturer's instructions (including DNase I digestion). RNA was finally eluted in 30 µL of nuclease-free water and quantified using the Qubit RNA Broad Range Assay (Life Tech).

Ribosomal RNA was depleted from up to 5 µg of footprinted RNA using the RiboZero Magnetic Gold kit (Illumina) according to the manufacturer's protocol. Footprinted RNA was precipitated from the supernatant (90 µL) using 1.5 µL of glycoblue (Life Tech), 9 µL of 3 M sodium acetate and 300 µL of ethanol by snap-freezing in liquid nitrogen, incubating for one hour up to overnight at -80°C, and pelleting at 21,000 x g for 30 min at 4°C. The RNA pellet was dissolved in 10 µL of RNase-free water.

Following rRNA depletion, isolation of short fragments and phosphorylation of these fragments by T4 PNK treatment, sequencing libraries were prepared using the NEXTflex Small RNA-Seq Kit v3 (Bioo Scientific). According to the manufacturer's instructions, adapters were diluted 1:2 to decrease adapter dimerization. To determine the optimal number of PCR cycles for library amplification, pilot PCRs with the respective forward and reverse primers were performed for each sample for 12, 14, 16, 18 and 20 cycles. Adapter and primer sequences are published by Bioo Scientific. Products were separated on a native PAGE, and optimal cycle numbers were determined as the threshold cycle of the library product at 160 bp, the expected size for RPFs, with the smallest amount of adapter dimer product (130 bp) possible. After the final PCR, libraries were separated on and excised from an agarose gel, and then cleaned using the Zymoclean Gel DNA Recovery kit (Zymo Research). Library quantification and validation were performed using the Qubit dsDNA HS and Bioanalyzer DNA HS assays, respectively. Three 0D5P control samples and three DAC treated samples (in a pool of 21 libraries) and two 0D5P samples (in a pool of 3 libraries) were sequenced on a NextSeq 500 machine at a loading concentration of 1.6 pM using High Output Kits v2 (Illumina) with 75 cycle single-end reads.

Ribo-Seq: analysis

Ribo-Seq reads were stripped of adaptor sequences using cutdapt, and contaminants such as tRNAs and rRNA were removed by alignment to a contaminants index via Bowtie v 2.3.5, consisting of nucleotide sequences from known human rRNA and tRNA sequences drawn from the GENCODE annotation v24[87]. Unaligned reads from this analysis were then aligned to human genome version hg19 with the STAR v 2.6.1a_08-27[88] splice-aware alignment tool allowing for up to 1 mismatch. The star genome index was built using GENCODE v24 (lift 37).

Reads with up to 20 multi-mapping positions were included, with multi-mapping reads beings separately treated in subsequent periodicity analysis. The RIboseQC pipeline v1.0[89] was used to deduce P-site positions from the Ribo-Seq reads, and the P-site data were then used as input into the SaTAnn pipeline v1.0[90] in combination with custom R scripts[86] for ORF calling. The SaTAnn pipeline searches for the periodic ribosomal footprint pattern characteristic of translated ORFs using a supplied database of transcripts, yielding a set of ORFs corresponding to known coding regions, as well as ORFs originating from UTRs, non-coding RNAs, intron retentions, and read-through events. The 0D5P samples had a median of 2.8 million reads mapped to coding sequences per sample, which constituted a median of 81% of the total reads **(Supplementary Table 2)**. Since the false positive rate of periodicity based ORF calling is thought to be tolerant of non-periodic sources of noise such as genomic contamination, we included all samples for 0D5P. ORFs were called in both individual libraries and in the pooled set of all libraries for 0D5P, and ORFs that were fully contained within ORFs detected in another library were merged. ORFs were tested for periodicity, by a multitaper test[86], and those with a p-value below 0.05 were retained for analyses.

Polypeptide sequences in fasta format were generated from the coordinates of these ORFs and used for both validation of the peptides found using the RNA-Seq-based database and as a *de novo*-assembled database for the subsequent round of peptide detection. Peptides were considered validated by Ribo-Seq if they matched anywhere within the translated ORF sequences.

Ribo-Seq profile plots were plotted with P-site numbers per-base on a log2 (n+1) scale.

The 10x Genomics pipeline and gene expression analyses

For single-cell library preparation on the 10x Genomics platform, the Chromium Single Cell 3′ Library and SingleCell 3’ Reagent v3 were utilized, together with the 10x Chromium single-cell controller instrument in accordance with the official CG000183 RevA user guide. A total of 1,692 0D5P cells were captured for single-cell transcriptomics. The resulting cDNA libraries were sequenced on NextSeq v 2.5 (with Illumina protocol #15048776). Cell Ranger v.3.0.1 software (10x Genomics, [https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines] was used to process data generated using the 10x Chromium platform, with a restriction of including only 1400 cells to avoid cells or debris with low unique molecular identifier (UMI) counts. This approach led to the detection of 19,178 genes with a mean of 125,937 mapped reads. Genes present in at least five cells and cells with at least 200 genes but no more than 50% of mito genes were retained for analysis, resulting in a reduced matrix of 15,710 genes over 1,365 cells.

The raw counts were log-normalized using the NormalizeData implemented in the Seurat R package (Seurat v3). Prior to further processing, we scaled the data to remove cell-cell variations due to cell cycling or a high percentage of mitochondrial genes. For cell cycling correction, we followed the scoring strategy described by Tirosh et al. (2016)[91]: each cell was assigned a “Cell Cycle” score and the difference between G2M and S phase scores was regressed out. Clusters were obtained using a graph-based method implemented in Seurat (FindClusters with a resolution set to 0.5), leading to the identification of 5 clusters. Marker genes for each cluster were identified with FindMarkers from Seurat by setting the logFC threshold parameter to 0.15. Marker genes with an adjusted Bonferroni p-value < 0.05 were considered significantly differentially expressed. Functional analyses of each cluster were performed with STRING-db v11 using their corresponding marker genes as input.

Assessing T cell reactivity

Peptides were synthesized and lyophilized by the Protein and Peptide Chemistry Facility at the Ludwig Institute for Cancer Research (crude, > 80% purity), Department of Oncology, University of Lausanne, or by Thermo Scientific, and resuspended in DMSO at 10 mg/mL. IFNγ ELISpot assays were conducted to assess the reactivity of the REP TILs towards antigens of interest (TAAs, noncHLAIp) using pre-coated 96-well ELISpot plates (Mabtech) according to the manufacturer’s protocol. If necessary, REP TILs were stimulated with a single peptide or a peptide pool at 1 µg/mL *in vitro* for 14 days before re-challenging with the peptide to assess the IFNγ response. For this purpose, REP TILs were plated at 1-2x10^5 cells per well and challenged for 18 h with cognate peptides at a final peptide concentration of 1 µM, in duplicate or triplicate. Medium without peptide was used as a negative control, and 1x Cell Stimulation Cocktail (eBioscience™, Thermo Fisher Scientific) was used as a positive control. Spot-forming units were quantified using the Bioreader-6000-E automated counter (BioSys). Positive hits were identified by having more spots than the negative control wells, which did not contain any peptide, plus 3 times the standard deviation of the negative control. Positivity was confirmed in at least ≥ 2 independent experiments.

The identification of circulating antigen-specific T cells in patient 0D5P was performed as such[66,92]: CD19+ cells were isolated from cryopreserved PBLs using magnetic beads (Miltenyi) and expanded for 14 days with multimeric-CD40L (Adipogen, Epalinges, Switzerland, 1 µg/mL) and IL-4 (Miltenyi, 200 IU/mL). CD8+ T lymphocytes were isolated from cryopreserved PBLs using magnetic beads (Miltenyi) and co-incubated at a 1:1 ratio with irradiated autologous CD40-activated B cells and peptides (single peptides or pools of ≤ 50 peptides, 1 µM each). After 12 days of *in vitro* expansion, CD8+ T cells were re-challenged with cognate peptide and T cell responses were assessed by the ELISpot assay.

Statistical analyses

Statistical analyses were performedwhere appropriate. The following tools were used for statistical analyses: GraphPad Prism 8, Perseus 1.5.5.3, RStudio 3.5.1 and Python 3.6. Specifically, the boxplots in Fig. 8c, Fig. 9a-b, Supplementary Fig. 1c-d and 1f-h, Supplementary Fig. 3a-b and Supplementary Fig. S4a-l were generated using the standard settings in either RStudio or GraphPad Prism. The boxplot settings were: Hinges (25% and 75%), with the median plotted. For Fig. 8c and Supplementary Fig. 1c-d and 1f-g, the notch is additionally shown at +/-1.58 IQR/sqrt(n), where IQR is the interquartile range (difference between 75- and 25-percentile) and n the number of data points. A median at the notch edge corresponds to a 95% significant difference (p-value=0.05). Sample sizes and p-values forSupplementary Fig. 1c-d and 1f-g can be found in the Source Data File. In Fig. 9a-b, Supplementary Fig. 1h, Supplementary Fig. 3 and 4, the whiskers are plotted down to the minimum and up to the maximum value, and each individual value is plotted as a point superimposed on the graph.

HLA-binding predictions

To evaluate the binding affinity of HLAIp, MixMHCpred.v2 prediction software was run on all HLAIp ranging in length from 8-14 amino acids. Peptides with a p-value < 0.05 were considered binders.

Sequence specific HI calculator

Sequence-specific HI was calculated with the SSRCalc vQ.0 tool^{35}, available online at [http://hs2.proteome.ca/SSRCalc/SSRCalcQ.html] . Only unmodified peptides were included and parameters were set to: 100Å C18 column, 0.1% formic acid separation system and without cysteine protection. Observed RTs were obtained from Comet pep.xml files. If a peptide was detected multiple times in the same sample, the mean RT was used. Peptides and their mean RTs were plotted against the calculated HIs. For Fig. 2 c-f, to compare the variances in the differences between the RTs and the regression line, we applied a one-sided F-test.

In order to calculate the standard errors of HI, we regressed the measured RTs against the calculated HI using the lm function in R. This function returns the residuals between the regression line and HI values. The residual absolute errors of the lm-regression were plotted in **Supplementary Fig. 1d and g** (the higher this value, the worse the correlation between predicted and measured values). In this manner, we observe how well the HI calculations correlate the experimentally observed RT.

Correlation analyses

Correlative analyses of the immunopeptidome and transcriptome of 0D5P **(Fig. 3a-d)** were performed by first assigning HLAp to their respective source genes. For noncHLAp, the gene with the highest transcript expression was allocated for further analyses if the peptide map back to more than one non-coding source gene unless otherwise indicated.

Assessing HLAIp sampling

For HLAIp sampling analyses, peptides were assigned to source protein groups as described above. Adjusted peptide counts were taken, summed over a gene, and subsequently matched to their corresponding expression values (either transcriptome or translatome based). Normalized sampling corresponds to the adjusted peptide count per protein, normalized by the protein length. The correlation between gene expression or the spectral coefficients of 3-periodic signals in Ribo-Seq data and HLA presentation were assessed by fitting a polynomial curve of degree 3 to each dataset. Pearson correlation was used to assess the correlation between the fitted curve and the data.

Peptide position analysis

For peptide position analysis within a protein sequence **(Supplementary Fig. 3)**, proteome-derived datasets fitting to the length distribution of the 95% confidence level of the lncRNA dataset were selected. Then, the position of the HLAp, relative to the full protein sequence, was calculated for source lncRNA and proteome-derived sequences. Since the data were not normally distributed, the Wilcoxon test was utilized for statistical analysis.

PRM analyses

For analyses of PRM statistics, MS-based intensities were taken from the initial MaxQuant peptide table output. TAAs for PRM and further comparative analyses were selected from a non-exhaustive list of known and clinically relevant TAAs.

GTEx RNA expression analyses

Tissue-specific gene expression data was downloaded from GTEx, a public resource that contains data from 53 non-diseased tissues across nearly 1000 individuals[46]. We used a custom R script to retrieve gene expression values, based on publicly available GTEx v7 data.

In the case of multiple transcripts matching the same entry, expression data for the most expressed transcript were used. The 90th percentile expression of the gene in the tissue-derived tumor was reported. The FPKM expression units of the investigated sample were converted into TPM units for comparison with the GTEx data. The R package “ComplexHeatmap v1.99.4”[93] from the Bioconductor suite was used to draw heatmaps.

References

1. Schumacher, T.N. & Schreiber, R.D. Neoantigens in cancer immunotherapy. *Science* **348**, 69-74 (2015).
2. Yarchoan, M., Johnson, B.A., 3rd, Lutz, E.R., Laheru, D.A. & Jaffee, E.M. Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* **17**, 569 (2017).
3. Ott, P.A.*, et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217-221 (2017).
4. Sahin, U.*, et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222-226 (2017).
5. Zajac, P.*, et al.* MAGE-A Antigens and Cancer Immunotherapy. *Front Med (Lausanne)* **4**, 18 (2017).
6. Connerotte, T.*, et al.* Functions of Anti-MAGE T-cells induced in melanoma patients under different vaccination modalities. *Cancer Res* **68**, 3931-3940 (2008).
7. Boudousquie, C.*, et al.* Polyfunctional response by ImmTAC (IMCgp100) redirected CD8(+) and CD4(+) T cells. *Immunology* **152**, 425-438 (2017).
8. Ingolia, N.T.*, et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports* **8**, 1365-1379 (2014).
9. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
10. Cardinaud, S.*, et al.* Identification of cryptic MHC I-restricted epitopes encoded by HIV-1 alternative reading frames. *The Journal of experimental medicine* **199**, 1053-1063 (2004).
11. Arun, G., Diermeier, S.D. & Spector, D.L. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends Mol Med* **24**, 257-277 (2018).
12. Jackson, R.*, et al.* The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434-438 (2018).
13. Attermann, A.S., Bjerregaard, A.M., Saini, S.K., Gronbaek, K. & Hadrup, S.R. Human endogenous retroviruses and their implication for immunotherapeutics of cancer. *Annals of Oncology* **29**, 2183-2191 (2018).
14. Khurana, E.*, et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108 (2016).
15. Ho, O. & Green, W.R. Cytolytic CD8+ T cells directed against a cryptic epitope derived from a retroviral alternative reading frame confer disease protection. *J Immunol* **176**, 2470-2475 (2006).
16. Weinzierl, A.O.*, et al.* A cryptic vascular endothelial growth factor T-cell epitope: identification and characterization by mass spectrometry and T-cell assays. *Cancer Res* **68**, 2447-2454 (2008).
17. Probst-Kepper, M.*, et al.* An alternative open reading frame of the human macrophage colony-stimulating factor gene is independently translated and codes for an antigenic peptide of 14 amino acids recognized by tumor-infiltrating CD8 T lymphocytes. *J Exp Med* **193**, 1189-1198 (2001).
18. Gonzalez-Cao, M.*, et al.* Human endogenous retroviruses and cancer. *Cancer Biol Med* **13**, 483-488 (2016).
19. Kassiotis, G. & Stoye, J.P. Immune responses to endogenous retroelements: taking the bad with the good. *Nature reviews. Immunology* **16**, 207-219 (2016).
20. Lander, E.S.*, et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
21. Bassani-Sternberg, M. & Coukos, G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr Opin Immunol* **41**, 9-17 (2016).
22. Erhard, F.*, et al.* Improved Ribo-seq enables identification of cryptic translation events. *Nature methods* (2018).
23. Laumont, C.M.*, et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Science translational medicine* **10**(2018).
24. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223 (2009).
25. Pearson, H.*, et al.* MHC class I-associated peptides derive from selective regions of the human genome. *The Journal of clinical investigation* **126**, 4690-4701 (2016).
26. Granados, D.P.*, et al.* Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nature communications* **5**, 3600 (2014).

27. Laumont, C.M.*, et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nature communications* **7**, 10238 (2016).

28. Nesvizhskii, A.I. Proteogenomics: concepts, applications and computational strategies. *Nature methods* **11**, 1114-1125 (2014).

29. Li, H.*, et al.* Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **17**, 1031 (2016).

30. Bassani-Sternberg, M.*, et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS computational biology* **13**, e1005725 (2017).

31. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular & cellular proteomics : MCP* **14**, 658-673 (2015).

32. Harrow, J.*, et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760-1774 (2012).

33. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372 (2008).

34. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22-24 (2013).

35. Krokhin, O.V. & Spicer, V. Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal Chem* **81**, 9522-9530 (2009).

36. Andreatta, M.*, et al.* MS-Rescue: A Computational Pipeline to Increase the Quality and Yield of Immunopeptidomics Experiments. *Proteomics* **19**, e1800357 (2019).

37. Zhou, F. Molecular mechanisms of IFN-gamma to up-regulate MHC class I antigen processing and presentation. *Int Rev Immunol* **28**, 239-260 (2009).

38. Castro, F., Cardoso, A.P., Goncalves, R.M., Serre, K. & Oliveira, M.J. Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion. *Front Immunol* **9**, 847 (2018).

39. Chong, C.*, et al.* High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferongamma-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. *Molecular & cellular proteomics : MCP* **17**, 533-548 (2018).

40. Zeigerer, A.*, et al.* Regulation of liver metabolism by the endosomal GTPase Rab5. *Cell reports* **11**, 884-892 (2015).

41. Li, H.*, et al.* Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget* **5**, 587-598 (2014).

42. Atkins, J.F., Loughran, G., Bhatt, P.R., Firth, A.E. & Baranov, P.V. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic acids research* **44**, 7007-7078 (2016).

43. Schatton, T.*, et al.* Identification of cells initiating human melanomas. *Nature* **451**, 345-349 (2008).

44. Widlund, H.R.*, et al.* Beta-catenin-induced melanoma growth requires the downstream target Microphthalmia-associated transcription factor. *J Cell Biol* **158**, 1079-1087 (2002).

45. Tachibana, M.*, et al.* Ectopic expression of MITF, a gene for Waardenburg syndrome type 2, converts fibroblasts to cells with melanocyte characteristics. *Nature genetics* **14**, 50-54 (1996).

46. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585 (2013).

47. Gerami, P.*, et al.* Development and validation of a noninvasive 2-gene molecular assay for cutaneous melanoma. *J Am Acad Dermatol* **76**, 114-120 e112 (2017).

48. Muller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization. *Frontiers in immunology* **8**, 1367 (2017).

49. Goode, L.L.*, et al.* A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet* **42**, 874-+ (2010).

50. Blakeley, P., Overton, I.M. & Hubbard, S.J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *Journal of proteome research* **11**, 5221-5234 (2012).

51. van Heesch, S.*, et al.* The Translational Landscape of the Human Heart. *Cell* **178**, 242-260 e229 (2019).

52. Caron, E., Aebersold, R., Banaei-Esfahani, A., Chong, C. & Bassani-Sternberg, M. A Case for a Human Immuno-Peptidome Project Consortium. *Immunity* **47**, 203-208 (2017).

53. Diament, A. & Tuller, T. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol Direct* **11**(2016).

54. Han, Y., Gao, S., Muegge, K., Zhang, W. & Zhou, B. Advanced Applications of RNA Sequencing and Challenges. *Bioinform Biol Insights* **9**, 29-46 (2015).

55. Karousis, E.D. & Muhlemann, O. Nonsense-Mediated mRNA Decay Begins Where Translation Ends. *Cold Spring Harb Perspect Biol* **11**(2019).

56. Apcher, S., *et al.* Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *P Natl Acad Sci USA* **108**, 11572-11577 (2011).

57. Slavoff, S.A., *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**, 59-64 (2013).

58. Prasad, S., Starck, S.R. & Shastri, N. Presentation of Cryptic Peptides by MHC Class I Is Enhanced by Inflammatory Stimuli. *J Immunol* **197**, 2981-2991 (2016).

59. Yewdell, J.W., Anton, L.C. & Bennink, J.R. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *Journal of immunology* **157**, 1823-1826 (1996).

60. Zheng, Y., Tan, K. & Huang, H. Long noncoding RNA HAGLROS regulates apoptosis and autophagy in colorectal cancer cells via sponging miR-100 to target ATG5 expression. *J Cell Biochem* **120**, 3922-3933 (2019).

61. Chen, J.F., *et al.* STAT3-induced lncRNA HAGLROS overexpression contributes to the malignant progression of gastric cancer cells via mTOR signal-mediated inhibition of autophagy. *Mol Cancer* **17**, 6 (2018).

62. Klebanoff, C.A. & Wolchok, J.D. Shared cancer neoantigens: Making private matters public. *J Exp Med* **215**, 5-7 (2018).

63. Dhodapkar, K. & Dhodapkar, M. Harnessing shared antigens and T-cell receptors in cancer: Opportunities and challenges. *Proc Natl Acad Sci U S A* **113**, 7944-7945 (2016).

64. Wilson, B.J., *et al.* ABCB5 Maintains Melanoma-Initiating Cells through a Proinflammatory Cytokine Signaling Circuit. **74**, 4196-4207 (2014).

65. Lang, D., Mascarenhas, J.B. & Shea, C.R. Melanocytes, melanocyte stem cells, and melanoma stem cells. *Clin Dermatol* **31**, 166-178 (2013).

66. Bobisse, S., *et al.* Sensitive and frequent identification of high avidity neo-epitope specific CD8 (+) T cells in immunotherapy-naive ovarian cancer. *Nature communications* **9**, 1092 (2018).

67. Ebrahimi-Nik, H., *et al.* Mass spectrometry driven exploration reveals nuances of neoepitope-driven tumor rejection. *JCI Insight* **5**(2019).

68. Huang, A.Y., *et al.* The immunodominant major histocompatibility complex class I-restricted antigen of a murine colon tumor derives from an endogenous retroviral gene product. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 9730-9735 (1996).

69. Bentzen, A.K., *et al.* Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nature biotechnology* **34**, 1037-1045 (2016).

70. Valmori, D., *et al.* Enhanced generation of specific tumor-reactive CTL in vitro by selected Melan-A/MART-1 immunodominant peptide analogues. *J Immunol* **160**, 1750-1758 (1998).

71. Neubert, N.J., *et al.* A Well-Controlled Experimental System to Study Interactions of Cytotoxic T Lymphocytes with Tumor Cells. *Front Immunol* **7**, 326 (2016).

72. Marino, F., Chong, C., Michaux, J. & Bassani-Sternberg, M. High-Throughput, Fast, and Sensitive Immunopeptidomics Sample Processing for Mass Spectrometry. *Methods in molecular biology* **1913**, 67-79 (2019).

73. MacLean, B., *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010).

74. Djebali, S., *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

75. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).

76. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

77. Gentleman, R.C., *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).

78. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

79. DePristo, M.A., *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).

80. Bassani-Sternberg, M.*, et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature communications* **7**, 13404 (2016).

81. Cox, J.*, et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **10**, 1794-1805 (2011).

82. Horlacher, O.*, et al.* MzJava: An open source library for mass spectrometry data processing. *Journal of proteomics* **129**, 63-70 (2015).

83. Ochoa, A., Storey, J.D., Llinas, M. & Singh, M. Beyond the E-Value: Stratified Statistics for Protein Domain Prediction. *PLoS computational biology* **11**, e1004509 (2015).

84. Zhang, B., Chambers, M.C. & Tabb, D.L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of Proteome Research* **6**, 3549-3557 (2007).

85. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S. & Stoica, I. Spark: cluster computing with working sets. in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing* 10-10 (USENIX Association, Boston, MA, 2010).

86. Calviello, L.*, et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13**, 165-170 (2016).

87. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

88. Dobin, A.*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

89. Calviello, L., Sydow, D., Harnett, D. & Ohler, U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. Preprint at https://www.biorxiv.org/content/10.1101/601468v1 (2019).

90. Calviello, L., Hirsekorn, A. & Ohler, U. SaTAnn quantifies translation on the functionally heterogeneous transcriptome. Preprint at https://www.biorxiv.org/content/10.1101/608794v1(2019).

91. Tirosh, I.*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).

92. Tanyi, J.L.*, et al.* Personalized cancer vaccine effectively mobilizes antitumor T cell immunity in ovarian cancer. *Sci Transl Med* **10**(2018).

93. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

94. Perez-Riverol, Y.*, et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research* **47**, D442-D450 (2019).

Data availability

Sequence data have been deposited into the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession numbers EGAS00001003723 and EGAS00001003724. MS raw files, corresponding fasta reference files and NewAnce outputs have been deposited into the ProteomeXchange Consortium via the PRIDE[94] partner repository with the dataset identifier PXD013649. Databases can be accessed through: The GENCODE v221.2 from https://www.gencodegenes.org/human/release_22.html. The human reference genome GRCh38/hg38 from https://www.ncbi.nlm.nih.gov/assembly/5800238. Human ENCODE non-coding transcripts from https://www.gencodegenes.org/human/release_24lift37.html. GTEx v7 from https://www.gtexportal.org/home/datasets. The UniProt/TrEMBL database from https://www.uniprot.org/proteomes/UP000005640. The source data underlying the Figures and Supplementary Figures, where applicable, are provided as a Source Data file.

Code availability

An executable jar file of NewAnce has been deposited to PRIDE with the dataset identifier PXD013649. The NewAnce code is available on the following GitHub link: [https://github.com/bassanilab/NewAnce.git].

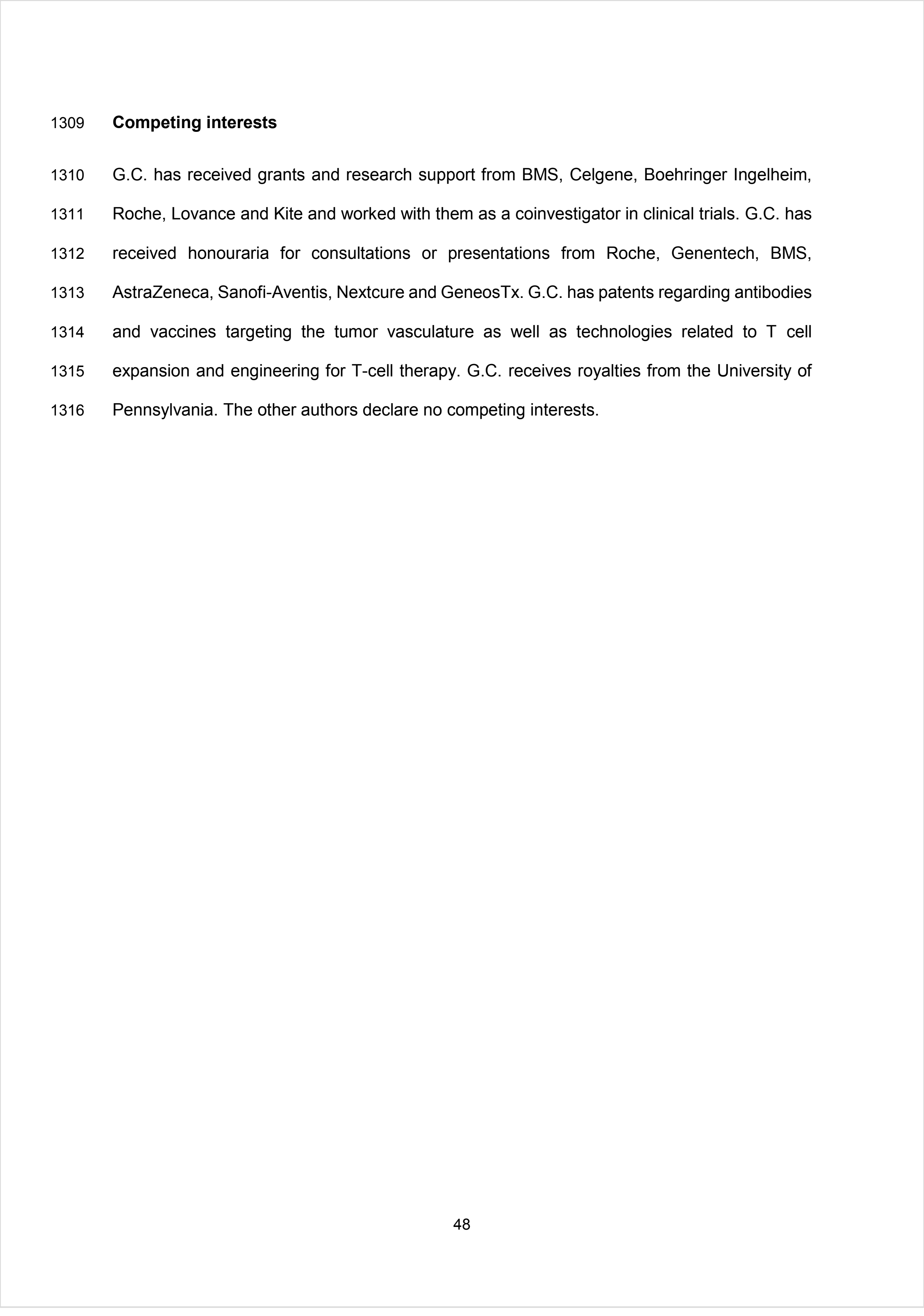
Acknowledgements

We thank the patients for their participation in this study and the Center of Experimental Therapy CHUV, Lausanne, for the collection, generation and allowed use of samples. Library preparation and sequencing was performed at the Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland [https://www.unil.ch/gtf/en/home.html]We thank the Gene Expression Core Facility at the *École polytechnique fédérale* of Lausanne for the scRNA-Seq data. We would also like to express our thanks to Julien Schmidt and Philippe Guillaume at the Protein and Peptide Facility and Tetramer Core Facility, University of Lausanne for the synthetic peptides. Further, we thank Raphael Genolet and Lise Queiroz for the HLA typing of the samples. We thank Fabio Marino for providing a few datasets included in ipMSDB. Last, we are very grateful to Prof. Daniel Speiser and Nicolas Gestermann for kindly providing the melanoma cell lines Me290, Me275, T1015A and T1185B and the autologous REP TILs used in this study.

This work was supported by the Ludwig Institute for Cancer Research and the ISREC Foundation thanks to a donation from the Biltema Foundation. This work was also supported by a grant from the German Federal Ministry of Education and Research (RNA Bioinformatics Center of the German Network for Bioinformatics Infrastructure [de.NBI; BMBF 031 A538C]) awarded to U.O.

Authors' contributions

G.C. and M.B.S. conceived and designed the project and interpreted the results. M.B.S. and C.C. designed the experiments and coordinated, integrated, and interpreted the multi-omics analyses. M.M. conceptualized and implemented the software for MS/MS data processing and FDR calculations in NewAnce and performed the ipMSDB analyses. C.C. J.M. and H.S.P. conducted the immunopeptidomics MS experiments. F.H. assisted in data visualization and GTEx analyses. C.C. and A.A. conducted the TIL experiments, and A.Ha. assisted with the data interpretation. B.J.S. performed the NGS analyses. I.X provided support for the computational infrastructure. D.G. and E.P. performed the bioinformatics quantifications and analysed the of TE expression data and L.S-R. and D.T. provided scientific interpretations of the TE analysis. M.L. analysed the scRNA Seq data. L.Z. analysed the RNA-Seq data for the protein-coding and non-coding genes. I.B. and A.Hi. processed the samples for Ribo-Seq, and D.H., L.C. and U.O. analysed and interpreted the Ribo-Seq data. M.W., J.L and M.B. coordinated the tissue collection and processing from lung cancer patients. J.L. performed the pathological evaluations. C.C., M.M. and M.B.S. wrote the manuscript with contributions from all authors.



Competing interests

G.C. has received grants and research support from BMS, Celgene, Boehringer Ingelheim, Roche, Lovance and Kite and worked with them as a coinvestigator in clinical trials. G.C. has received honouraria for consultations or presentations from Roche, Genentech, BMS, AstraZeneca, Sanofi-Aventis, Nextcure and GeneosTx. G.C. has patents regarding antibodies and vaccines targeting the tumor vasculature as well as technologies related to T cell expansion and engineering for T-cell therapy. G.C. receives royalties from the University of Pennsylvania. The other authors declare no competing interests.

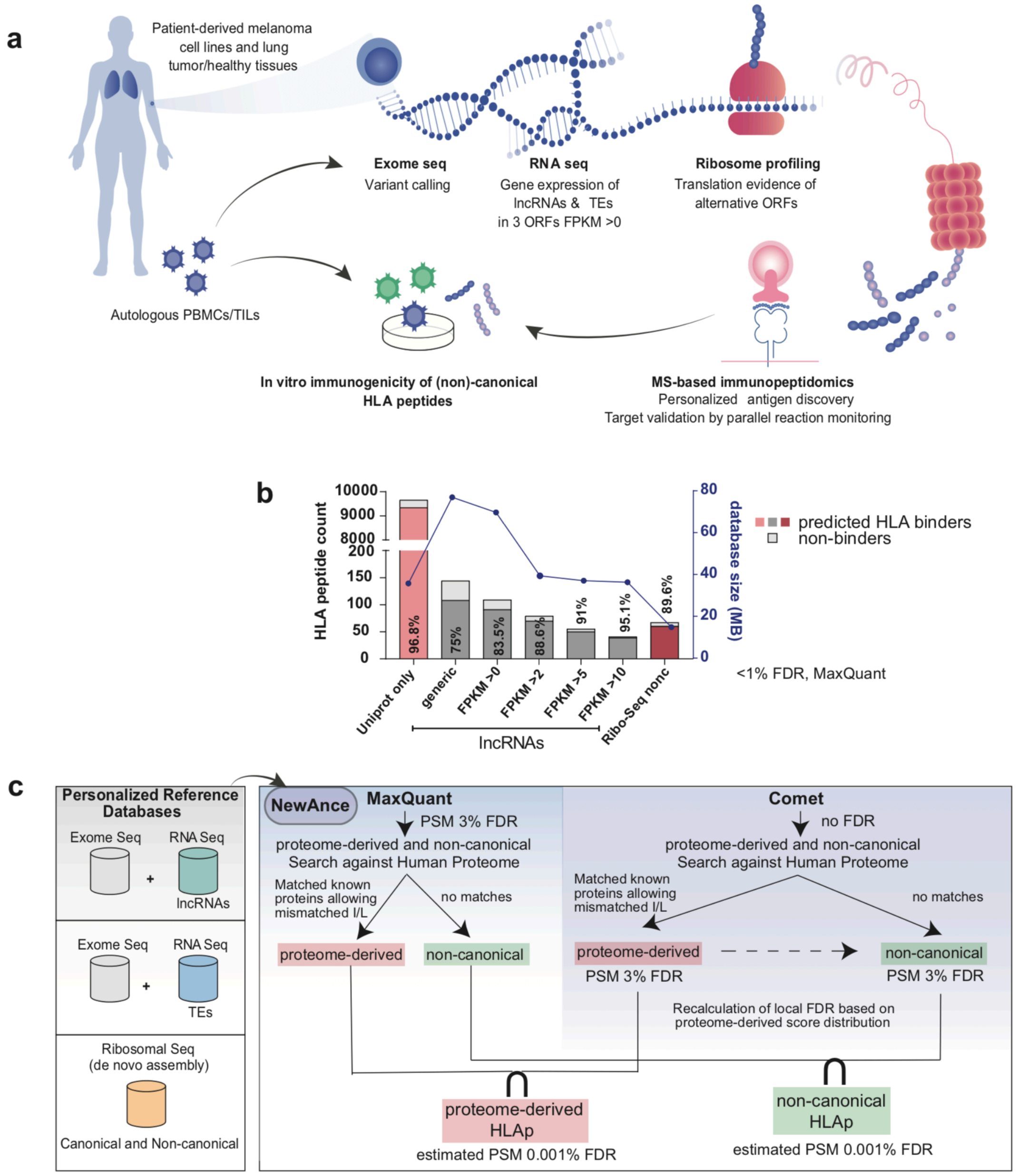

a
Patient-derived melanoma cell lines and lung tumor/healthy tissues
Exome seq
Variant calling
RNA seq
Gene expression of lncRNAs & TEs in 3 ORFs FPKM >0
Ribosome profiling
Translation evidence of alternative ORFs
Autologous PBMCs/TILs
In vitro immunogenicity of (non)-canonical HLA peptides
MS-based immunopeptidomics
Personalized antigen discovery
Target validation by parallel reaction monitoring
b
HLA peptide count
10000
9000
8000
200
150
100
50
0
database size (MB)
80
60
40
20
0
96.8%
75%
83.5%
88.6%
91%
95.1%
89.6%
Uniprot only
generic
FPKM >0
FPKM >2
FPKM >5
FPKM >10
Ribo-Seq nonc
lncRNAs
predicted HLA binders
non-binders
<1% FDR, MaxQuant
c
Personalized Reference Databases
Exome Seq
RNA Seq
+
lncRNAs
Exome Seq
RNA Seq
+
TEs
Ribosomal Seq
(de novo assembly)
Canonical and Non-canonical
NewAnce
MaxQuant
PSM 3% FDR
proteome-derived and non-canonical
Search against Human Proteome
Matched known proteins allowing mismatched I/L
no matches
proteome-derived
non-canonical
Comet
no FDR
proteome-derived and non-canonical
Search against Human Proteome
Matched known proteins allowing mismatched I/L
no matches
proteome-derived
PSM 3% FDR
non-canonical
PSM 3% FDR
Recalculation of local FDR based on proteome-derived score distribution
∩
proteome-derived HLAp
estimated PSM 0.001% FDR
∩
non-canonical HLAp
estimated PSM 0.001% FDR

Fig. 1 A proteogenomics approach for the robust identification of noncHLAp. a A schematic of the entire workflow is shown, where tissue samples or tumor cell lines were obtained from patients, and exome, RNA and Ribo-Seq were performed to provide a framework to assess the non-canonical antigen repertoire. HLAp were immunoaffinity-purified from cancer cell lines and matched tumor/healthy lung tissues and then analysed by MS. Immunopeptidomics spectra were then searched against RNA- and Ribo-Seq-based personalized protein sequence databases that contain non-canonical polypeptide sequences. MS-identified noncHLAlp were validated by targeted MS-based PRM and tested for immunogenicity using autologous T cells or PBMCs. **b** The percentage of predicted HLA binders of length 8-14 mer peptides with a MixMHCpred p-value ≤ 0.05 was used to evaluate the accuracy of the identified HLAlp by MaxQuant at 1% FDR as a function of database size (blue line). The percentage of predicted binders obtained for each condition is shown for each bar for the melanoma cell line 0D5P. **c** Different protein sequence databases combining whole exome sequencing and inferrences from RNA-Seq and Ribo-Seq data were utilized. NewAnce was implemented by retaining the PSM intersection of the two MS search tools MaxQuant and Comet, and applying group-specific FDR calculations for protHLAp and noncHLAp. Source data are provided as a Source Data file.

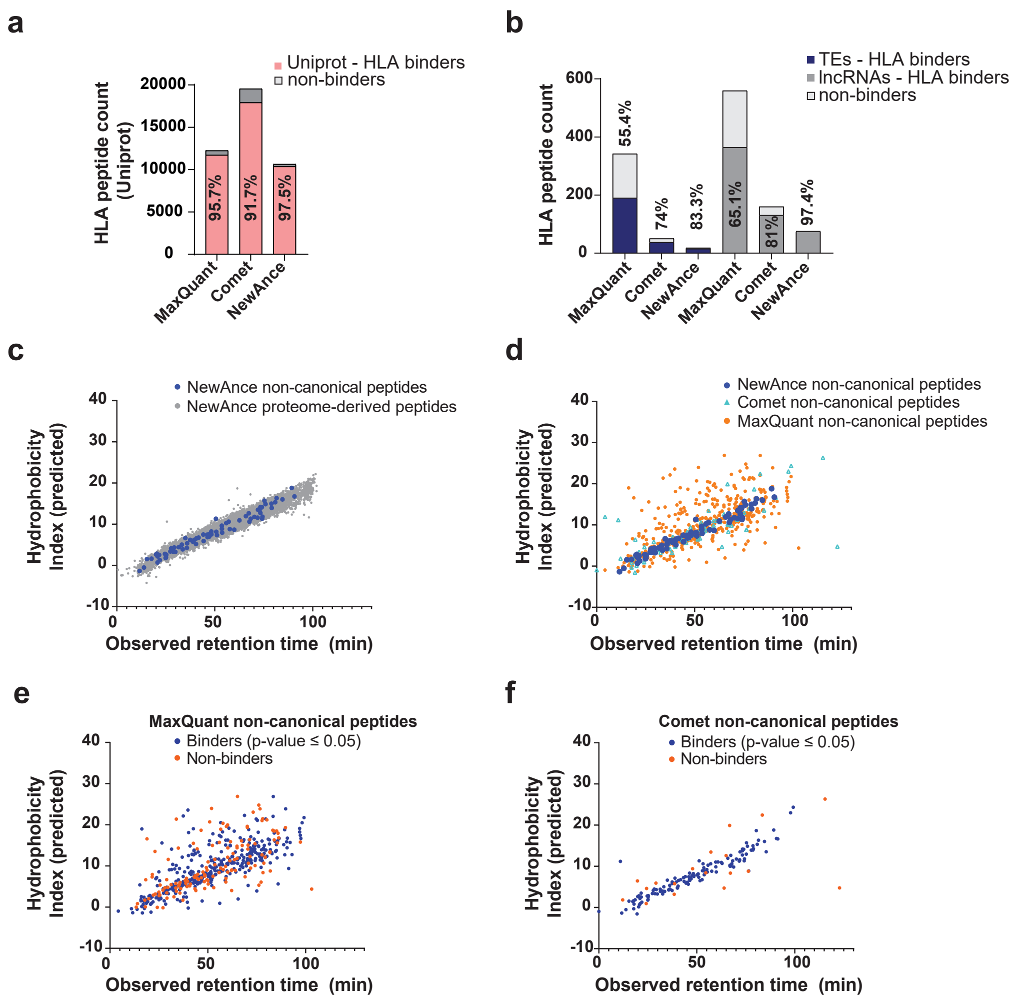


Fig. 2 Two complementary methods to assess the accuracy of NewAnce. a The percentages of predicted proteome-derived HLA-I binders in 0D5P were assessed with each MS search tool (MaxQuant and Comet at FDR 3%) and NewAnce. **b** Similar to a, the comparisons were performed for the different non-canonical antigen classes. **c** Retention predictions by SSRCalc for peptides identified in melanoma 0D5P. The observed mean retention time is plotted against the hydrophobicity indices for NewAnce-identified proteome-derived versus lncRNA-derived non-canonical peptides. **d** All peptides identified with each tool (MaxQuant, Comet, NewAnce) were analysed based on their hydrophobicity indices. **e** Hydrophobicity index calculation for MaxQuant- or **f** Comet-identified 8- to 14-mer peptides, based on predicted HLA binding. Source data are provided as a Source Data file.

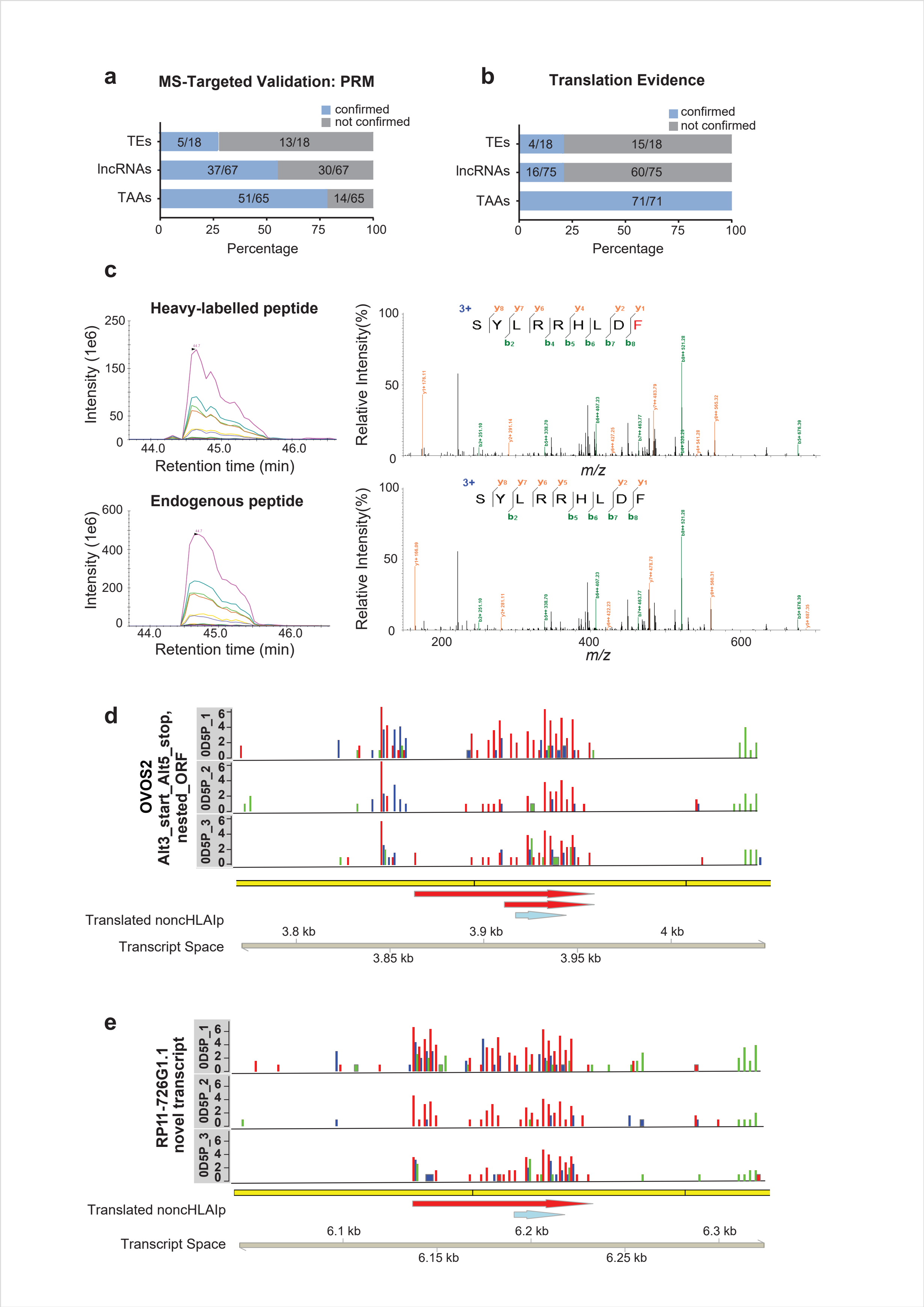

a
MS-Targeted Validation: PRM
confirmed
not confirmed
TEs
lncRNAs
TAAs
5/18
13/18
37/67
30/67
51/65
14/65
0
25
50
75
100
Percentage
b
Translation Evidence
confirmed
not confirmed
TEs
lncRNAs
TAAs
4/18
15/18
16/75
60/75
71/71
0
25
50
75
100
Percentage
c
Heavy-labelled peptide
Intensity (1e6)
250
150
50
0
44.0
45.0
46.0
Retention time (min)
Endogenous peptide
Intensity (1e6)
600
400
200
0
44.0
45.0
46.0
Retention time (min)
Relative Intensity(%)
100
50
0
3+
S Y L R R H L D F
y8 y7 y6 y4 y2 y1
b2 b4 b5 b6 b7 b8
m/z
Relative Intensity(%)
100
50
0
3+
S Y L R R H L D F
y8 y7 y6 y5 y2 y1
b2 b5 b6 b7 b8
200
400
600
m/z
d
OVOS2
Alt3_start_Alt5_stop,
nested_ORF
0D5P_1
0D5P_2
0D5P_3
6
4
2
0
Translated noncHLAIp
Transcript Space
3.8 kb
3.9 kb
4 kb
3.85 kb
3.95 kb
e
RP11-726G1.1
novel transcript
0D5P_1
0D5P_2
0D5P_3
6
4
2
0
Translated noncHLAIp
Transcript Space
6.1 kb
6.2 kb
6.3 kb
6.15 kb
6.25 kb

Fig. 3 MS-based and ribosome footprint evidence of non-canonical peptide generation. A set of proteome-derived tumor-associated antigens and noncHLAIp (lncRNAs and TEs) from melanoma 0D5P were synthesized in their heavy labelled form and spiked back into replicates of HLAIp eluted from 0D5P to confirm the presence of endogenous HLAIp. The proportions of confirmed and non-confirmed HLAIp as determined by **a** PRM and **b** Ribo-Seq-targeted validation are shown for each of the antigen classes. **c** An example of the co-elution profiles of the transitions of heavy labelled and endogenous noncHLAIp (from lncRNA; SYLRRHLDF) from 0D5P (left) is shown. The MS/MS fragmentation pattern further confirms the presence of the endogenous peptide (Δ m=10 Da) (right). **d, e** The Ribo-Seq profiles of two source genes show the frequency of Ribo-Seq reads from the ribosome's P-site in three replicates. Library size-normalized P-sites per basepair are shown on a log2 scale on the Y-axis, with P-sites inferred as a constant offset from the 5' end of the footprint for each read length. The colored bars represent different reading frames. The yellow bars below the plots represent exons. For example, the noncHLAIp SYLRRHLDF in *OVOS2* (blue arrow) falls within two nested, Ribo-Seq-supported ORFs (red arrows), within which most P-sites (red bars) fall in the first reading frame. Source data are provided as a Source Data file.

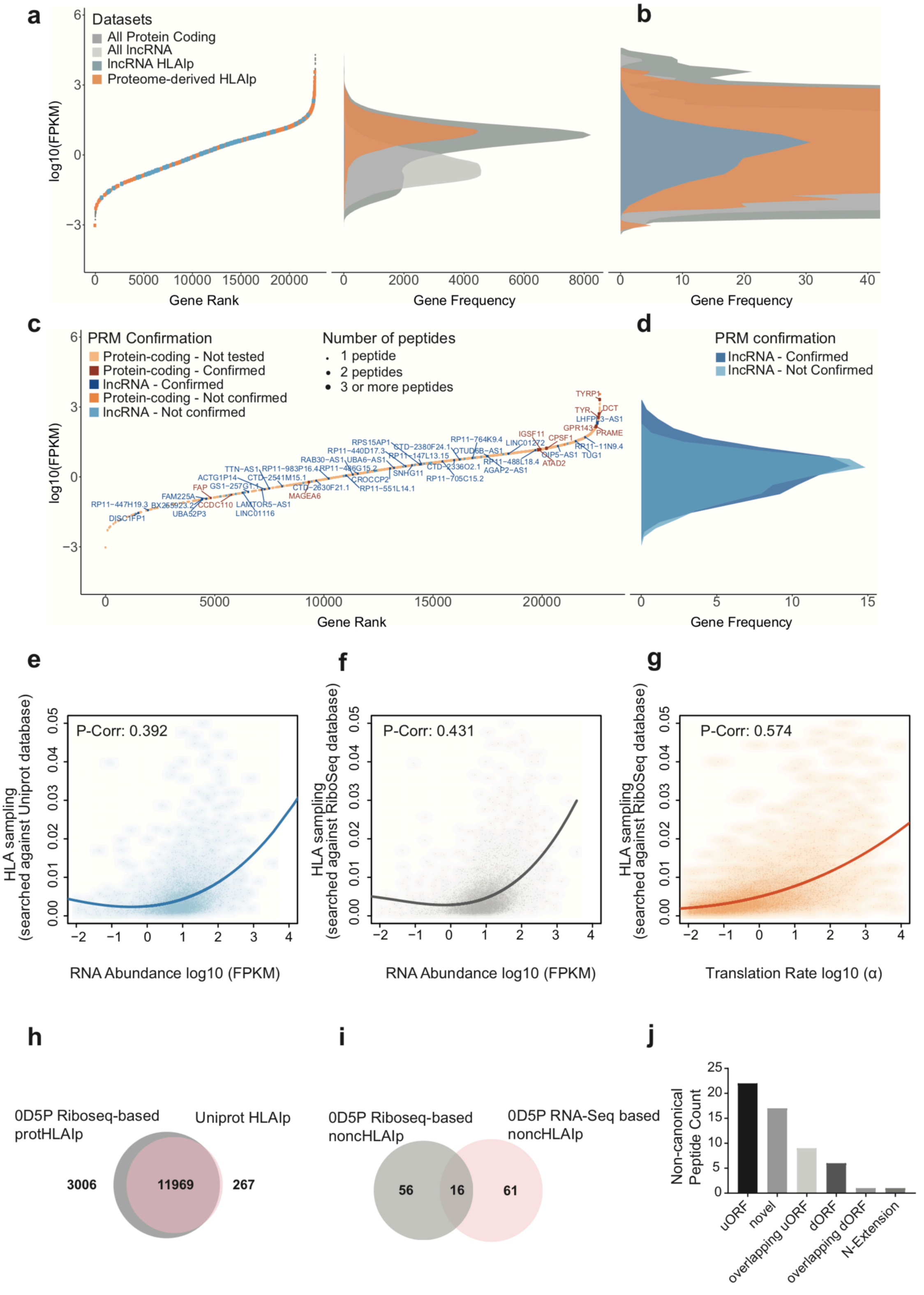

a
Datasets
All Protein Coding
All lncRNA
lncRNA HLAIp
Proteome-derived HLAIp
log10(FPKM)
6
3
0
−3
0
5000
10000
15000
20000
Gene Rank
0
2000
4000
6000
8000
Gene Frequency
b
0
10
20
30
40
Gene Frequency
c
PRM Confirmation
Protein-coding - Not tested
Protein-coding - Confirmed
lncRNA - Confirmed
Protein-coding - Not confirmed
lncRNA - Not confirmed
Number of peptides
1 peptide
2 peptides
3 or more peptides
TYRP1
TYR
DCT
LHFPL3−AS1
GPR143
PRAME
IGSF11
CPSF1
RP11−764K9.4
LINC01272
RP11−11N9.4
RPS15AP1
CTD−2380F24.1
OTUD6B−AS1
OIP5−AS1
TUG1
RP11−440D17.3
RP11−147L13.15
RP11−488L18.4
ATAD2
RAB30−AS1
UBA6−AS1
CTD−2336O2.1
AGAP2−AS1
TTN−AS1
RP11−983P16.4
RP11−486G15.2
SNHG11
RP11−705C15.2
ACTG1P14
CTD−2541M15.1
CROCCP2
FAP
GS1−257G1.1
CTD−2630F21.1
RP11−551L14.1
MAGEA6
FAM225A
RP11−447H19.3
BX255923.2
CCDC110
LAMTOR5−AS1
UBA52P3
LINC01116
DISC1FP1
log10(FPKM)
6
3
0
−3
0
5000
10000
15000
20000
Gene Rank
d
PRM confirmation
lncRNA - Confirmed
lncRNA - Not Confirmed
0
5
10
15
Gene Frequency
e
P-Corr: 0.392
HLA sampling
(searched against Uniprot database)
0.00
0.01
0.02
0.03
0.04
0.05
−2
−1
0
1
2
3
4
RNA Abundance log10 (FPKM)
f
P-Corr: 0.431
HLA sampling
(searched against RiboSeq database)
0.00
0.01
0.02
0.03
0.04
0.05
−2
−1
0
1
2
3
4
RNA Abundance log10 (FPKM)
g
P-Corr: 0.574
HLA sampling
(searched against RiboSeq database)
0.00
0.01
0.02
0.03
0.04
0.05
−2
−1
0
1
2
3
4
Translation Rate log10 (α)
h
0D5P Riboseq-based
protHLAIp
Uniprot HLAIp
3006
11969
267
i
0D5P Riboseq-based
noncHLAIp
0D5P RNA-Seq based
noncHLAIp
56
16
61
j
Non-canonical
Peptide Count
0
5
10
15
20
25
uORF
novel
overlapping uORF
dORF
overlapping dORF
N-Extension

Fig. 4 RNA- and Ribo-Seq-based gene expression analyses from melanoma 0D5P. a Genes are ranked based on their RNA expression levels in 0D5P. P protein-coding (orange) and presumed non-coding (blue) source genes, in which HLAIp were identified. The frequency distributions of the gene expression levels of protein-coding and non-coding (lncRNA) genes are shown. **b** The region of interest is magnified to show the distribution of noncHLAIp source gene expression. **c** Source gene restriction plot. Targeted MS validation was performed, and confirmations are denoted for all identified non-canonical peptides and for a subset of protHLAIp (selected TAAs). Confirmed hits indicate that one or more peptides from that source gene were validated by PRM. Point sizes represent the number of peptides identified per source gene. **d** Frequency distribution of gene expression for MS-confirmed versus non-confirmed (or inconclusive) noncHLAIp. Scatterplots show the correlation between **e** UniProt-based HLA-I sampling and RNA abundance**, f** Ribo-Seq-based HLA-I sampling and RNA abundance, and **g** Ribo-Seq-based HLA-I sampling and translation abundance. HLA-I sampling was calculated from the adjusted peptide counts normalized by protein length. Determination of the correlation between gene expression and HLA-I sampling was assessed by fitting a polynomial curve of degree 3 to each dataset. Pearson correlation values were calculated to assess the correlation between the fitted curve and the corresponding dataset. **h** With data derived from 0D5P, a comparison of the overall overlap in unique HLAIp identified with RNA-Seq-based and Ribo-Seq-based assembled databases for MS search is shown. **i** Overlap of noncHLAIp identified by RNA-Seq- and Ribo-Seq-based searches. **j** The total number of noncHLAIp identified by Ribo-Seq is depicted for each of the respective ORF types. Source data are provided as a Source Data file.

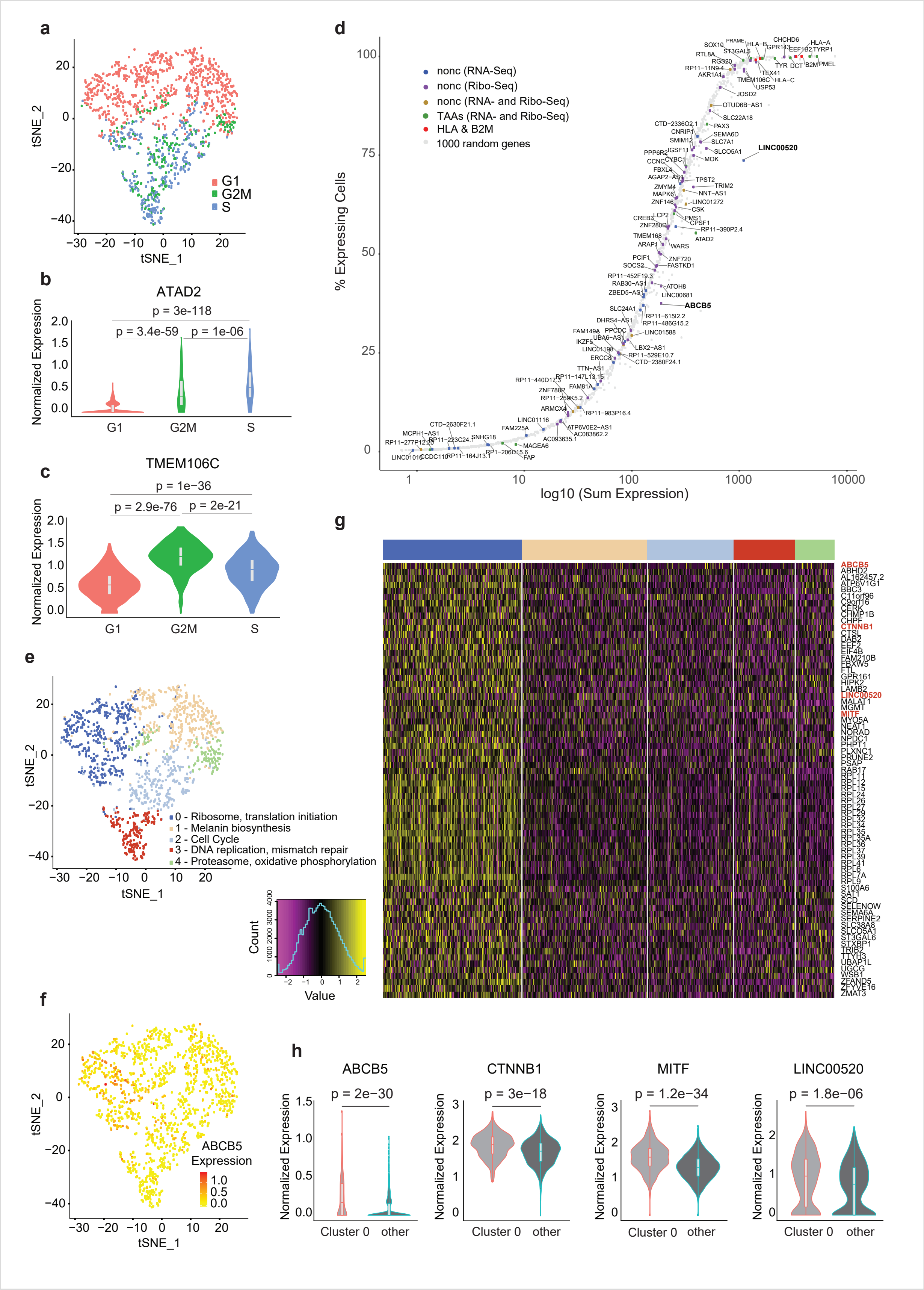

a
tSNE_2
20
0
−20
−40
−30 −20 −10 0 10 20
tSNE_1
G1
G2M
S
b
ATAD2
p = 3e-118
p = 3.4e-59
p = 1e-06
Normalized Expression
2.0
1.5
1.0
0.5
0.0
G1
G2M
S
c
TMEM106C
p = 1e−36
p = 2.9e-76
p = 2e-21
Normalized Expression
2.0
1.5
1.0
0.5
0.0
G1
G2M
S
d
% Expressing Cells
100
75
50
25
0
● nonc (RNA-Seq)
● nonc (Ribo-Seq)
● nonc (RNA- and Ribo-Seq)
● TAAs (RNA- and Ribo-Seq)
● HLA & B2M
● 1000 random genes
SOX10
PRAME
HLA−B
CHCHD6
HLA−A
GPR143
EEF1B2
TYRP1
RTL8A
ST3GAL5
RGS20
RP11−11N9.4
AKR1A1
TYR
DCT
B2M
PMEL
TEX41
TMEM106C
HLA−C
USP53
JOSD2
OTUD6B−AS1
SLC22A18
CTD−2336O2.1
PAX3
CNRIP1
SEMA6D
SMIM12
SLC7A1
LINC00520
PPP6R2
IGSF11
SLCO5A1
CCNC
CYBC1
MOK
FBXL4
AGAP2−AS1
TPST2
TRIM2
ZMYM4
MAPK6
NNT−AS1
ZNF146
LINC01272
CSK
CREB3
LCP2
PMS1
CPSF1
ZNF280D
RP11−390P2.4
ATAD2
TMEM168
ARAP1
WARS
PCIF1
ZNF720
SOCS2
FASTKD1
RP11−452F19.3
RAB30−AS1
ATOH8
ZBED5−AS1
LINC00681
ABCB5
SLC24A1
RP11−615I2.2
RP11−486G15.2
DHRS4−AS1
FAM149A
PPCDC
LINC01588
UBA6−AS1
IKZF5
LBX2−AS1
LINC01198
RP11−529E10.7
ERCC8
CTD−2380F24.1
TTN−AS1
RP11−440D17.3
RP11−147L13.15
FAM81A
ZNF788P
RP11−259K5.2
ARMCX4
RP11−983P16.4
LINC01116
FAM225A
ATP6V0E2−AS1
AC083862.2
CTD−2630F21.1
MCPH1−AS1
AC093635.1
RP11−277P12.20
RP11−223C24.1
SNHG18
MAGEA6
LINC01010
CCDC110
RP11−164J13.1
RP1−206D15.6
FAP
1
10
100
1000
10000
log10 (Sum Expression)
e
tSNE_2
20
0
−20
−40
−30 −20 −10 0 10 20
tSNE_1
0 - Ribosome, translation initiation
1 - Melanin biosynthesis
2 - Cell Cycle
3 - DNA replication, mismatch repair
4 - Proteasome, oxidative phosphorylation
Count
0 1000 2000 3000 4000
−2 −1 0 1 2
Value
g
ABCB5
ABHD2
AL162457.2
ATP6V1G1
BBC3
C11orf96
C9orf16
CERK
CHMP1B
CHPF
CTNNB1
CTSL
DAB2
EEF2
EIF4B
FAM210B
FBXW5
FTL
GPR161
HIPK2
LAMB2
LINC00520
MALAT1
MGMT
MITF
MYO5A
NEAT1
NORAD
NPDC1
PHPT1
PLXNC1
PRUNE2
PSAP
RAB17
RPL11
RPL12
RPL15
RPL24
RPL26
RPL27
RPL29
RPL32
RPL34
RPL35
RPL35A
RPL36
RPL37
RPL39
RPL41
RPL6
RPL7A
RPL9
S100A6
SAT1
SCD
SELENOW
SEMA6A
SERPINE2
SLC38A8
SLCO5A1
ST3GAL6
STXBP1
TRIB2
TTYH3
UBAP1L
UGCG
WSB1
ZFAND5
ZFYVE16
ZMAT3
f
tSNE_2
20
0
−20
−40
−30 −20 −10 0 10 20
tSNE_1
ABCB5
Expression
1.0
0.5
0.0
h
ABCB5
p = 2e−30
Normalized Expression
1.5
1.0
0.5
0.0
Cluster 0
other
CTNNB1
p = 3e−18
Normalized Expression
3
2
1
0
Cluster 0
other
MITF
p = 1.2e−34
Normalized Expression
3
2
1
0
Cluster 0
other
LINC00520
p = 1.8e−06
Normalized Expression
2
1
0
Cluster 0
other

Fig. 5 ScRNA-Seq reveals non-coding transcriptional heterogeneity in melanoma 0D5P. a t-SNE plot of the 1,365 cells colored by their “cell cycle” scores. **b** Examples of cell-cycle dependent genes: *ATAD2*, a tumor-associated antigen, and **c** *TMEM106C*, from which a noncHLAIp originated. **d** Genes of interest were plotted based on their sum normalized expression by scRNA-Seq and ordered based on the percentage of cells that expressed the gene. The color codes denote the type of HLAIp identified from those genes. **e** t-SNE plot of the 1,365 cells colored by the five identified clusters. Clusters were annotated based on functional enrichment analyses of marker genes. **f** t-SNE plot highlighting the expression of the *ABCB5* gene enriched in cluster 0. **g** Heatmap showing the scaled and centred expressions of marker genes in cluster 0. The cluster colors from (e) are represented above the plot. **h** Expression profiles of four marker genes in cluster 0 over all other clusters, including two well-known cancer biomarkers, *MITF* and *CTNNB1*, and two source genes for which noncHLAIp were identified, the *ABCB5* gene with a dORF and *LINC00520*. The p-values represented in (b), (c) and (h) were obtained with Wilcoxon tests. Source data are provided as a Source Data file.

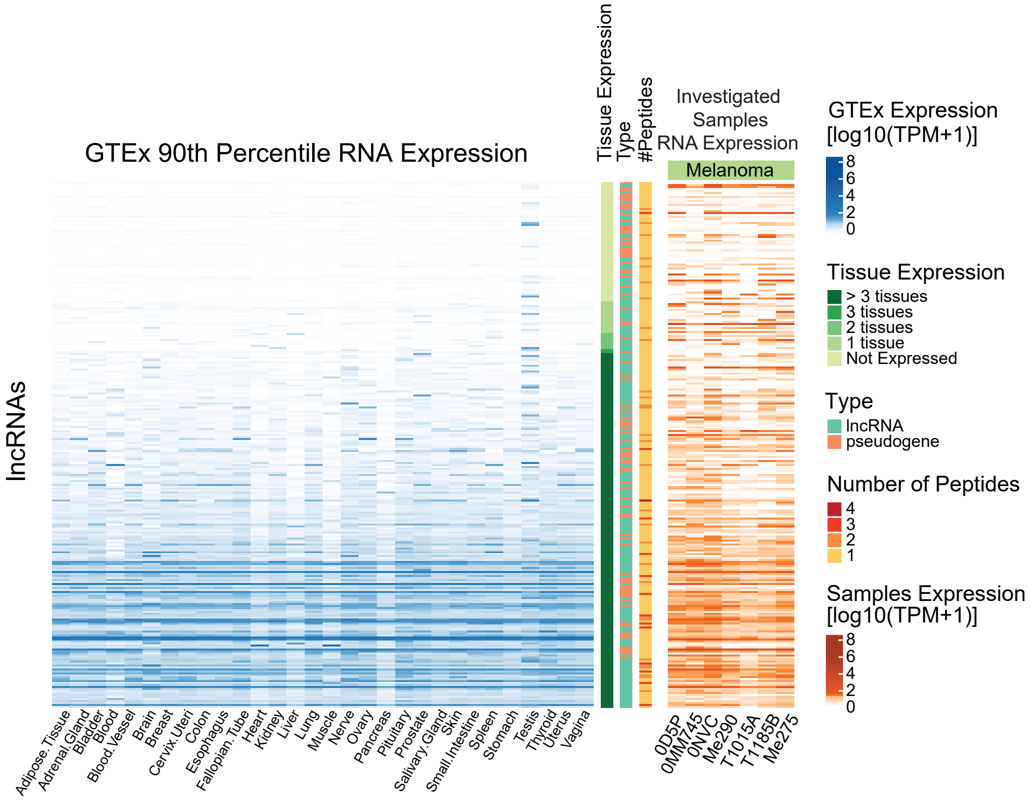


Fig. 6 Non-coding source gene expression in healthy tissues. A comparison of presumed non-coding source gene expression in the investigated samples to that in healthy tissues (GTEx) reveals that a substantial proportion of source non-coding genes are tumor-specific. Heatmap of lncRNA source genes showing the 90th percentile gene expression levels across 30 healthy tissues on the left and the gene expression levels across our investigated melanoma samples on the right. Tissue gene expression was classified as not expressed (90th percentile TPM ≤¬ 1) in any, 1-3, or more than 3 tissues other than testis to assess tumor specificity. Specifically for sample 0D5P, a total of 21.4% of the lncRNA source genes were considered as tumor specific compared to < 1% of the randomly selected protein-coding source genes with similar expression levels (P-value = 1.04 e-33). The number of HLAIp identified per gene is depicted as well as the gene (GENCODE) and sample type. Source data are provided as a Source Data file.

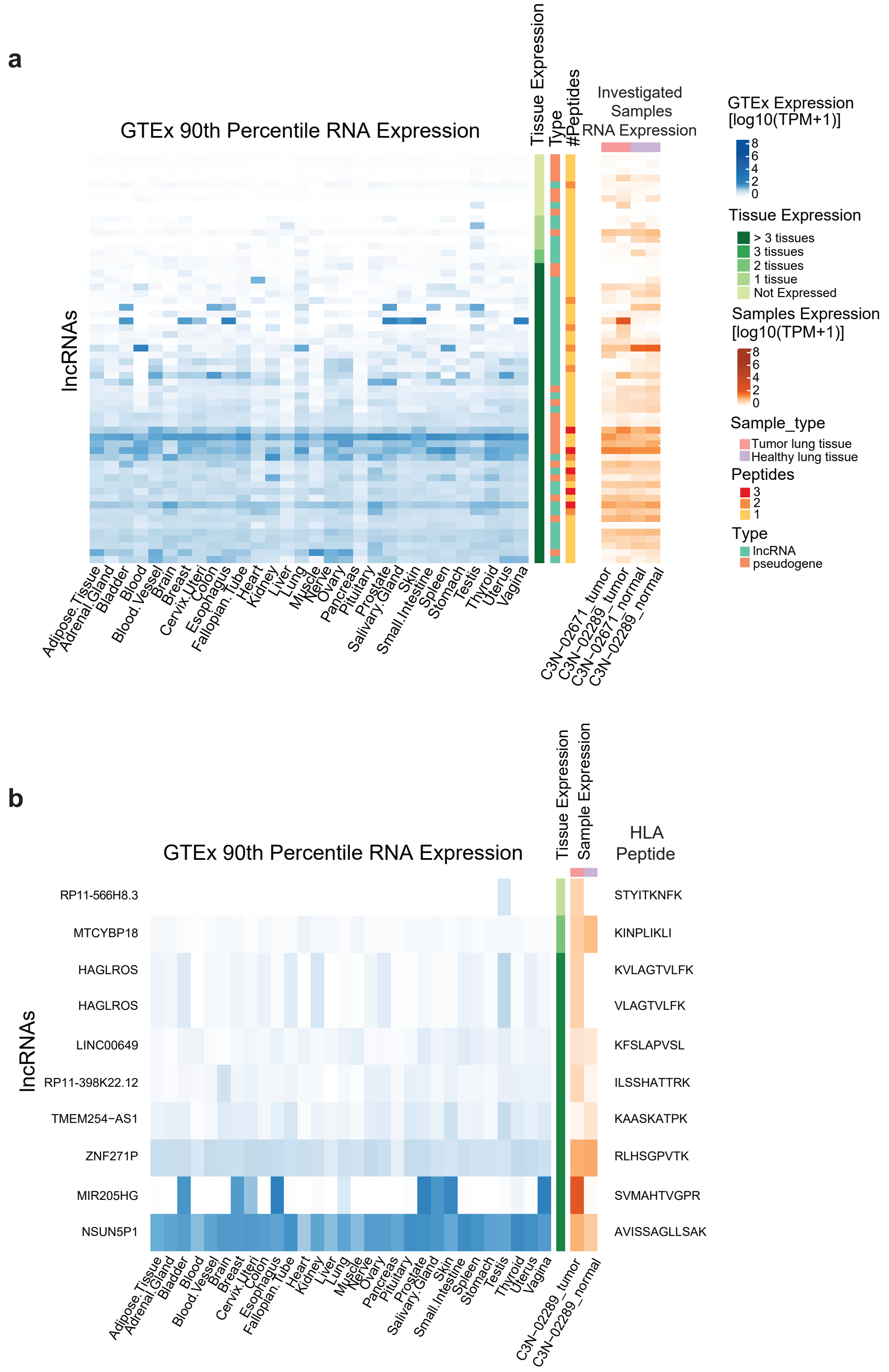

a
GTEx 90th Percentile RNA Expression
Tissue Expression
Type
#Peptides
Investigated Samples RNA Expression
lncRNAs
GTEx Expression [log10(TPM+1)]
8
6
4
2
0
Tissue Expression
> 3 tissues
3 tissues
2 tissues
1 tissue
Not Expressed
Samples Expression [log10(TPM+1)]
8
6
4
2
0
Sample_type
Tumor lung tissue
Healthy lung tissue
Peptides
3
2
1
Type
lncRNA
pseudogene
Adipose.Tissue
Adrenal.Gland
Bladder
Blood
Blood.Vessel
Brain
Breast
Cervix.Uteri
Colon
Esophagus
Fallopian.Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary.Gland
Skin
Small.Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
C3N−02671_tumor
C3N−02289_tumor
C3N−02671_normal
C3N−02289_normal
b
GTEx 90th Percentile RNA Expression
Tissue Expression
Sample Expression
HLA Peptide
lncRNAs
RP11-566H8.3
MTCYBP18
HAGLROS
HAGLROS
LINC00649
RP11-398K22.12
TMEM254−AS1
ZNF271P
MIR205HG
NSUN5P1
STYITKNFK
KINPLIKLI
KVLAGTVLFK
VLAGTVLFK
KFSLAPVSL
ILSSHATTRK
KAASKATPK
RLHSGPVTK
SVMAHTVGPR
AVISSAGLLSAK
Adipose.Tissue
Adrenal.Gland
Bladder
Blood
Blood.Vessel
Brain
Breast
Cervix.Uteri
Colon
Esophagus
Fallopian.Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary.Gland
Skin
Small.Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
C3N−02289_tumor
C3N−02289_normal

Fig. 7 Non-coding source gene expression of lung cancer samples in healthy tissues. A comparison of presumed non-coding source gene expression in the investigated samples to that in healthy tissues (GTEx). **a** Heatmap of lncRNA source genes showing the 90th percentile gene expression levels across 30 healthy tissues on the left and the gene expression levels identified in lung tissue samples on the right. Tissue gene expression was classified as not expressed (90th percentile TPM ≤ 1) in any, 1-3, or more than 3 tissues other than testis to assess tumor specificity. The number of HLAIp identified per gene is depicted as well as the gene (GENCODE) and sample type. **b** Specifically, this was also plotted for the tumor-specific noncHLAIp identified in lung cancer patient C3N-02289. Source data are provided as a Source Data file.

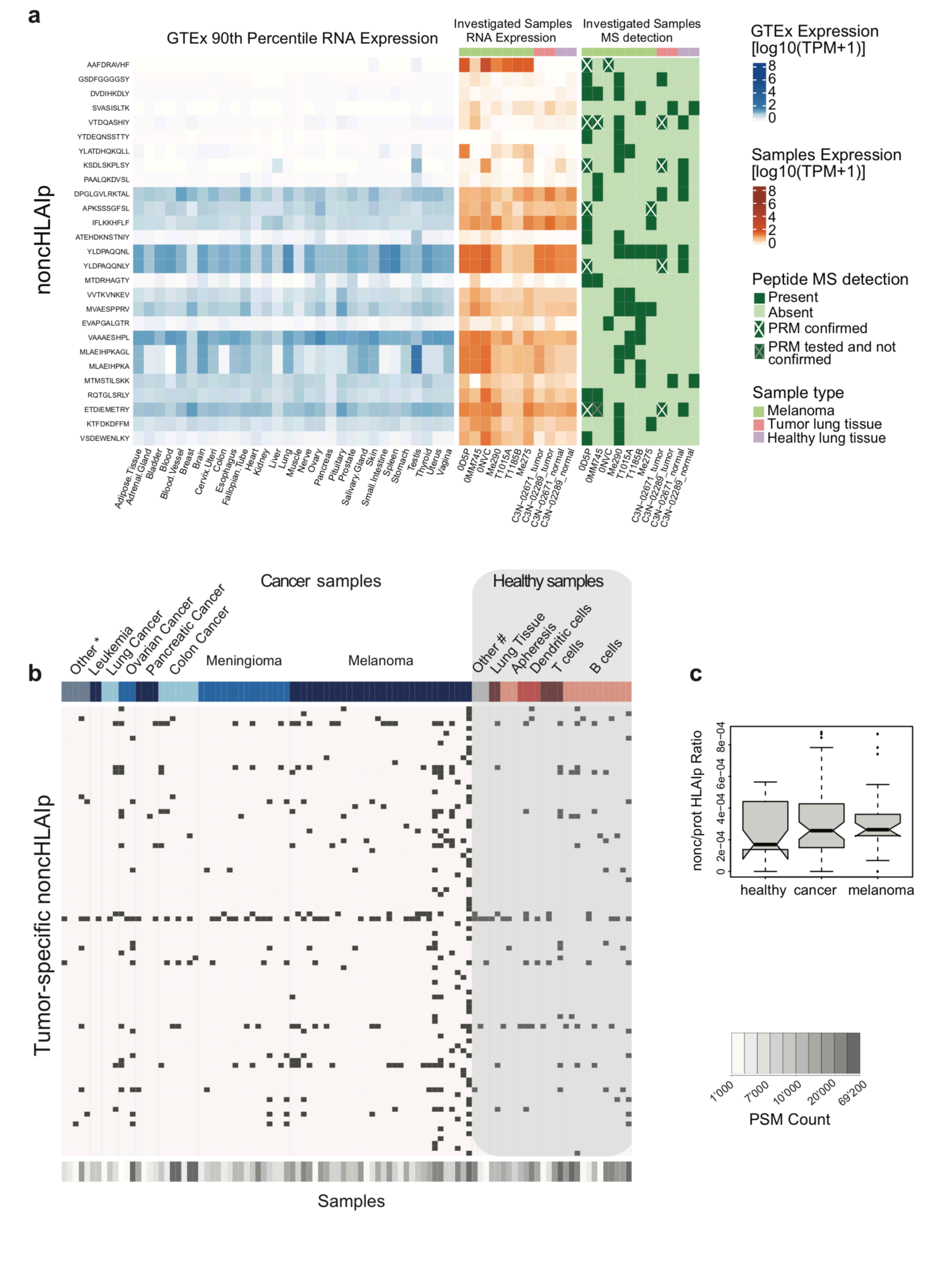

a
GTEx 90th Percentile RNA Expression
Investigated Samples RNA Expression
Investigated Samples MS detection
noncHLAIp
AAFDRAVHF
GSDFGGGGSY
DVDIHKDLY
SVASISLTK
VTDQASHIY
YTDEQNSSTTY
YLATDHQKQLL
KSDLSKPLSY
PAALQKDVSL
DPGLGVLRKTAL
APKSSSGFSL
IFLKKHFLF
ATEHDKNSTNIY
YLDPAQQNL
YLDPAQQNLY
MTDRHAGTY
VVTKVNKEV
MVAESPPRV
EVAPGALGTR
VAAAESHPL
MLAEIHPKAGL
MLAEIHPKA
MTMSTILSKK
RQTGLSRLY
ETDIEMETRY
KTFDKDFFM
VSDEWENLKY
Adipose.Tissue
Adrenal.Gland
Bladder
Blood
Blood.Vessel
Breast
Brain
Cervix.Uteri
Colon
Esophagus
Fallopian.Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary.Gland
Skin
Small.Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
0D5P
0MM745
0NVC
Me290
T1015A
T1185B
Me275
C3N-02671_tumor
C3N-02289_tumor
C3N-02671_normal
C3N-02289_normal
0D5P
0MM745
0NVC
Me290
T1015A
T1185B
Me275
C3N-02671_tumor
C3N-02289_tumor
C3N-02671_normal
C3N-02289_normal
GTEx Expression [log10(TPM+1)]
8
6
4
2
0
Samples Expression [log10(TPM+1)]
8
6
4
2
0
Peptide MS detection
Present
Absent
PRM confirmed
PRM tested and not confirmed
Sample type
Melanoma
Tumor lung tissue
Healthy lung tissue
b
Cancer samples
Healthy samples
Other *
Leukemia
Lung Cancer
Ovarian Cancer
Pancreatic Cancer
Colon Cancer
Meningioma
Melanoma
Other #
Lung Tissue
Apheresis
Dendritic cells
T cells
B cells
Tumor-specific noncHLAIp
Samples
c
nonc/prot HLAIp Ratio
0
2e−04
4e−04
6e−04
8e−04
healthy
cancer
melanoma
1'000
7'000
10'000
20'000
69'200
PSM Count

Fig. 8 NoncHLAIp can be shared across individuals. a The noncHLAIp-centric heatmap (left) shows the corresponding presumed non-coding gene expression (90th percentile) across healthy tissues as well as that in our investigated samples (middle). The peptides that were identified by MS across the investigated samples, and therefore shared, are outlined in the rightmost heatmap. Validation by PRM was performed for multiple noncHLAIp across the corresponding samples and are denoted with cross markings. **b** NoncHLAIp identified across a large collection of immunopeptidomics datasets (ipMSDB) consisting of both cancer and healthy samples. Tumor-specific noncHLAIp were re-identified and a trend of enrichment in cancer samples was observed. The noncHLAIp sequences can be found in the source data file. Cancer samples are labelled in shades of blue, and the star symbol include tumor metastases, myeloma, uterine, brain and liver cancer. Healthy samples are indicated in shades of red, and the hashtag symbol include fibroblast cells and epithelial cells. **c** Boxplot depicting the ratio of noncHLAIp over protHLAIp identified in the different groups of samples derived from ipMSDB (healthy n=27, cancer n=63, melanoma n=25) One-sided T test was performed, without multiple testing correction. Healthy versus cancer p-value=0.17, healthy versus melanoma p-value=0.12. Please refer to the Methods section for boxplot parameters. Source data are provided as a Source Data file.

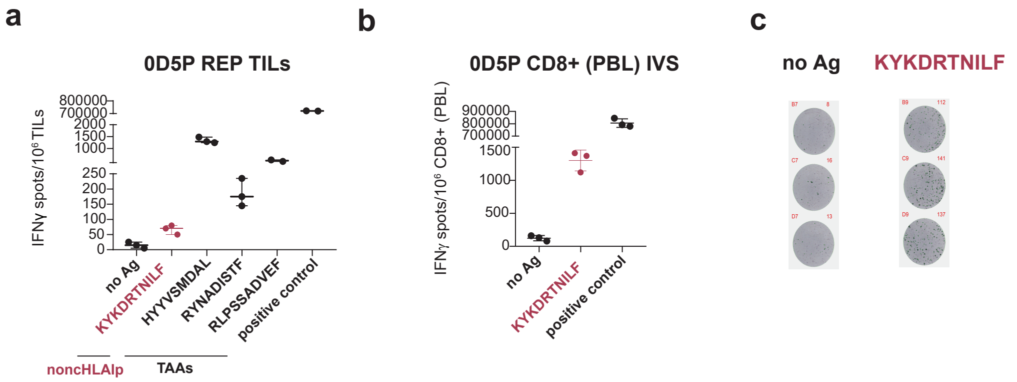


Fig 9 Non-canonical ABCB5 peptide induced an IFNγ response. a Reactivity was measured in melanoma 0D5P by the IFNγ ELISpot assay using autologous REP TILs. Representative example of three TAAs from TYR and TYRP1 and one non-canonical dORF-derived HLAIp from ABCB5 (written in red) that induced an IFNγ response. **b** In addition, a representative example of CD8+ T lymphocytes from PBLs is shown when re-challenged with autologous CD4+ blasts together with 1uM of the non-canonical ABCB5 HLAIp. (No Ag: no peptide, positive control: 1x cell stimulation cocktail). **c** Representative images of the IFNγ ELISpot response against the non-canonical ABCB5 peptide. In (a) and (b), T-cell reactivity for every peptide was validated by ≥2 independent experiments. Please refer to the Methods section for boxplot parameters. Source data are provided as a Source Data file.

Chapter 6

Discussion

Chapter 6 DISCUSSION

The discussion of the presented findings and their interpretation in light of the state-of-the-art is split into two parts. The first section focuses on the development of the HLA immunoaffinity purification method (see Chapter 4), while the second section provides more insight into the proteogenomics pipeline for identification of tumor non-canonical peptides (see Chapter 5).

6.1 The Challenge Of HLA Immunoaffinity Purification

In this first section, the challenges in the area of HLA immunoaffinity purification are covered, and the validation procedures in relation to other methods are outlined. This is followed by an illustration of how the high quality data that was obtained has been used to advance immunopeptidomics studies. The section concludes by describing limitations to the current experimental pipeline, proposing possible improvements, and importantly, detailing where the method will be implemented in the near future.

Previously, TAAs have been identified in a variety of ways, from targeted molecular approaches, to NGS techniques coupled with computational predictions and downstream immunogenicity evaluation with cellular-based assays [185]. Deep and diverse insights into HLA-bound tumor antigens are now constantly being acquired through the development of MS-related technologies [208, 209, 306-309]. The clinical applications of MS in the field of proteomics are already established, and are often based on targeted approaches to identify and quantify protein biomarkers [310]. However, in the immunopeptidomics field, the key question still remains: Will MS represent a clinical technology that can routinely identify tumor-specific neoantigens, and other types of tumor antigens, with high sensitivity?

To date, the field of immunopeptidomics has not reached its full potential, despite the many advances and improvements that have been made over the last years. Thus far, MS has shown promising, although minimal, results in pinpointing tumor antigen targets [209, 213]. This is due to the fact that MS-based technologies often lack in sensitivity, and the associated workflows are highly dependent on practical factors, such as the availability of samples, and prior expertise [215]. For example, experimental HLA immunoaffinity purification workflows suffer from well-acknowledged drawbacks, and remain a bottleneck in the robust and sensitive identification of the antigen repertoire. Specifically, the fundamental issues result from the requirement of obtaining large amounts of initial biological material, often impossible when dealing with patient samples, as well as extensive sample handling during HLA extraction, which leads to significant peptide losses. Lastly, HLA immunoaffinity purification protocols are un-standardized across laboratories, lacking detailed information and step-by-step guides (see Chapter 4).

With these issues in mind, we developed a high-throughput HLA immunoaffinity purification system, described in Chapter 4, with the following features that offer improvements over traditional methods: the possibility for parallel processing and scalability of up to 96 samples using a positive pressure system, the reduction of antibody-crosslinked beads, increased speed, and the elimination of error-prone steps resulting in higher recovery, sensitivity and purity of the measured immunopeptidome. These features are crucial when handling precious clinical samples that often suffer from low availability of material. Increased sensitivity with our pipeline was observed even for single MS measurements of 10^7 B cells. Furthermore, parallel processing allowed high reproducibility to be achieved for label-free comparative analyses. The feasibility of the latter was demonstrated by treating ovarian cells with the inflammatory cytokine IFNγ. For the first time, we uncovered the upregulated presentation of both chymotryptic-like and longer ligands upon IFNγ, likely due to modulation of the antigen processing and presentation machinery.

6.1.1 Method Validation

The purpose of designing a positive pressure high-throughput system was to increase the ease and speed of HLA immunoaffinity purification, and eliminate extensive sample handling. Following this, we performed multiple experiments to validate the reproducibility and robustness of the system. Below, three important validation results are highlighted and discussed.

The first validation strategy employed was to analyze peptide identification and reproducibility of the pipeline using JY, an Epstein-Barr virus (EBV) transformed human B cell line, as a representative sample. This cell line is routinely applied as a quality control in MS processes across different laboratories, and has very recently been used in the FDA-guided validation of technical MS equipment to support clinical trials [104, 118, 260, 311-314]. Here, the values presented from Ghosh and colleagues will serve as rough guidance for evaluating the efficacy of our system [314]. In terms of peptide identification depth across three technical replicates of JY at 1% FDR, the results in this thesis led to the identification of more than 3,000 HLAIp, similar to that reported by Ghosh et al. Next, considering the differences across MS injections of JY, reproducibility and precision with the presented method were demonstrated across both technical and biological replicates, as well as over three different days with Pearson correlations of up to 0.94, in line of what was reported previously. There were minute effects arising from plate manufacture differences, position of wells used, pressure differences and the quality of consumables. Overall, when interpreting the results from this comparison we should consider that different protocols were used and information reported, along with the utilization of different MS instrumentations, computational pipelines and biological material. However, this evaluation alongside the comparable FDA-guided study indicates the robustness of the method presented in this thesis.

Second, a straightforward experiment was performed to demonstrate the adequate peptide recovery of the pipeline. Specifically, 15 endogenous peptides from the B cell line CD165 were selected by considering the distribution of retention times and based on their high intensities, synthesized in their heavy-labelled forms, and spiked back into the sample prior to C18 reverse phase extraction using the positive pressure system. Importantly, all of the peptides were re-identified using this spike-in approach, and no cross contamination was observed. As peptide losses prior to C18 were not directly analyzed, isotopically labelled peptide-MHC monomers could be used in the future to directly quantify the losses from the immunoaffinity purification step [217].

Third, the sensitivity of the system was challenged by assessing HLA peptide yields while decreasing the cell amount input down to 10^7 B cells (CD165). This resulted in the identification of nearly 2,000 HLA-I and 3,000 HLA–II peptides. In contrast, the anticipated results for 10^7 Jurkat cells reported in the latest immunopeptidomics protocol were in the range of a hundred HLA peptides, suggesting the superiority of our presented method to boost peptide yields [216]. However, ultimately, the yield depends on the expression of HLA molecules in the respective sample and the cell type used. Thus, as recently outlined by the immunopeptidomics consortium, it is anticipated that benchmarking processes in multi-laboratory studies using the same sample batch will allow a better head-to-head comparison of HLA immunoaffinity purification protocols [215].

6.1.2 Results From The HLA Immunoaffinity Purification System

The search for mutated neoantigens by discovery-based MS is largely linked to peptide measurement depth [213]. Thus, acquiring high quality immunopeptidome data remains a key pre-requisite, and is especially important to enable robust sequence assignment, deep data-mining, and label-free comparative studies. Below, a selection of findings and insights that were achieved with the high quality data obtained from the method are discussed.

Specifically, in Chapter 4, a label-free comparative study was performed by interrogating the modulation of HLA peptide presentation in ovarian cells upon IFNγ treatment. Importantly, although multiple reports observed overall mediocre correlations between protein levels and presented HLA peptides [96, 240, 260-262], we showed the enhanced presentation of peptides derived from proteins of origin that were over-expressed after IFNγ treatment, such as STAT1 and 2, TAP1 and 2, b2m, OAS3, WARS, IFI16, and IRF. This implies that with sufficient peptide identification depth, along with reproducible methods, the correlation between the immunopeptidome and the proteome can be observed. Additionally, previously un-reported features were identified, such as the upregulated presentation of chymotryptic-like HLA ligands upon IFNγ treatment, likely due to proteasome to immunoproteasome switch confirmed at the proteome level. IFNγ also led to the significantly increased presentation of longer peptides, as well as chymotryptic-like extended peptides, when compared to shorter tryptic-like counterparts. Importantly, IFNγ is an inflammatory agent that is often found highly expressed in the tumor bed [315, 316]. Several studies have now shown that IFNγ can indeed remodel the immunopeptidome extensively. For example, it was very recently shown that IFNγ led to an increase in HLA-B levels, and that this in turn positively correlated with immune cell infiltration in the tumor using TCGA lung cancer data [262]. Importantly, the increase in HLA-B expression upon IFNγ was also observed in our study, leading to the upregulated presentation of HLA-B binding peptides. Together, these findings have an important implication on the routine use of HLA binding predictions for the selection of tumor antigens. Specifically, one could imagine that IFNγ, or other stimuli, could represent critical factors in shaping the antigen repertoire, a feature currently overlooked by HLA prediction algorithms. The phenomenon displayed by IFNγ should be investigated over multiple biological systems and, if universally found, could be used to adapt antigen prioritization strategies.

Using this novel system, we are routinely interrogating the presentation of TAAs in cancer samples based on their known potential to mediate anti-tumor responses [206, 207, 317, 318]. In a collaboration study with the group of Svane, we discovered two overlapping immunogenic cancer-testis GAGE peptides using MS in ovarian cancer sample OC.TIL.11, with one of them previously unreported [319]. GAGE-specific TILs were sorted from bulk TILs and were shown to kill autologous tumor cells. This study highlights the potential of MS-based discovery to uncover immunogenic epitopes in ovarian cancer. Furthermore, as epigenetic modulators in combination with checkpoint blockade has been shown to increase the immunogenicity of the tumor, we sought to investigate the upregulation of TAAs with DAC treatment [320]. Specifically, in Chapter 5, immunopeptidome changes upon DAC treatment in melanoma samples were analyzed. TAAs were indeed upregulated upon treatment, albeit to a lower extent than previously reported [292, 320, 321]. This discrepancy could be due to the previously observed global hypomethylation status at baseline in melanoma cell lines [322-325]. Alternatively, although the effect of DAC was observed at the RNA level, low copy numbers of antigens could have rendered them un-detectable in the immunopeptidome.

In addition, the generation of high quality data not only facilitated label-free comparisons, but has also enabled the significant improvement of existing HLA binding prediction algorithms. The immunopeptidome data collected in the lab of Bassani-Sternberg was applied by the Gfeller group to train algorithms predicting the binding of HLA-I, -II, as well as PTM peptides [175, 178, 212]. Combined, these studies underline the power of using multi-allelic samples from rich, diverse and naturally processed datasets to perform motif deconvolution. In contrast, extensive *in vitro* generation of mono-allelic cell lines can be used to resolve a HLA binding motif without the need for prior de-convolution [177, 179]. However, the multi-allelic approach used by the Gfeller group showed high similarities with methods using mono-allelic cell lines, both in terms of motifs, and the improvement of predictors. Furthermore, motif deconvolution guided by HLA-II peptidomics showed that HLA-DR motifs were sufficiently resolved with pan-HLA-II samples, suggesting that mono-allelic samples may be dispensable. Ultimately, only the MS-based HLA peptide profiling of unmodified biological samples can capture the complexity and co-dependencies of HLA alleles, as well as variations of peptide processing. Thus, the rich information from this strategy will remain the driving force in the evolution of prediction tools [326].

6.1.3 Limitations

A current limitation of the presented HLA extraction system is a high dependency on the strong binding affinities of the antibodies used for immunoaffinity purification. Antibodies that are of lower affinity require longer incubation times with the respective lysate. When employing low-affinity antibodies, the interactions between antibody and HLA complexes do not have time to properly occur, as the positive pressure system results in the continuous flow of samples through the wells. Thus, lower peptide yields are obtained. Furthermore, the plate format is only partially compatible with large sample volumes, such as when applying 5-10 mL plasma samples for immunopeptidomics studies [327, 328]. Therefore, if this pipeline is to be more generally and universally used, then optimizations will be needed for these two limitations. For example, this could include more tightly meshed well filters that lead to the slower flow of the lysate, or the incorporation of an additional plate-compatible incubation step at 4°C. Furthermore, custom made plates could be designed to integrate lysates larger than 2 mL.

6.1.4 Future Perspectives

While the presented system already drastically reduced the amount of time and manual handling needed for HLA immunoaffinity purification, there are further ways to increase the peptide yield. For example, current separation approaches using the common C18 media during desalting or LC lead to the lack of binding of certain peptides that harbor specific chemical properties [216]. Therefore, implementation of other types of separation protocols, such as high pH, or strong cation exchange, could help overcome this bias. Furthermore, peptide yield can be increased through pre-fractionation of immunopeptidomics samples to reduce the complexity of the peptide mixture prior to MS injection [218]. Moreover, to accommodate analyses of small tissue amounts, one could envision the interesting option of miniaturization through the use of digital microfluidic devices for immunoaffinity purification, where processes are easily controlled and lower sample volumes are needed [329, 330]. Lastly, robotic systems could take over the currently semi-manual handling of samples, which would undoubtedly allow for more efficient parallel processing and improve overall peptide yield.

This workflow focuses on primarily on the optimization of experimental steps prior to injection into the MS. Thus, improvements to the methods surrounding MS could be further explored to enhance peptide identification. For example, combining the different MS fragmentation techniques ETD and HCD has been shown to facilitate sequence assignment and increase peptide identification by three-fold in a human B cell line at 1% FDR [268]. In addition, boosting peptide identification and reproducibility over multiple experiments can be achieved by applying DIA techniques with compatible comprehensive spectral libraries [205, 241], currently being optimized in the lab of Bassani-Sternberg. The latter approach is indeed highly attractive to overcome the requirement of large sample quantities, a severe bottleneck in immunopeptidomics. Using libraries generated from DDA analyses with 3e8 cells, the DIA method has been reported to identify more than 3,000 peptides using only 1 million human cells [241]. Further, high resolution accurate mass analyzers recognize a vast amount of peptide features, however, only a small fraction can currently be identified, thus there remains the need for faster and more sensitive measurement options [331]. These options include exciting developments in MS instrumentations. For example, the use of MS techniques that measure the mobility of ions, based on the charge, size and shape of the ion, adds another dimension to ion separation [332, 333]. High Field Asymmetric waveform Ion Mobility Spectrometry (FAIMS) incorporates such an ion mobility device that excels in removing interfering background ions [334]. This has been observed to enhance the limit of detection by nearly an order of magnitude with proteomics samples, leading to higher sensitivity for identification of low abundance peptides. Furthermore, the use of Trapped Ion Mobility Spectrometry (TIMS) allows for increased sequencing speed when coupled to a TOF analyzer, and therefore enables the generation of more complete datasets [335, 336]. We envision that these approaches, alongside the steady improvement of instrument sensitivity, would together allow for even higher identification rates and greater reproducibility across immunopeptidomics studies.

In summary, in the first part of this thesis, a comprehensive experimental toolset was provided to robustly map the immunopeptidome, with the step-by-step protocol published separately [337]. A pilot cancer immunotherapy clinical trial at the University Hospital of Lausanne (CHUV), Switzerland, was recently performed using the presented HLA-immunoaffinity purification system [338]. While no mutated antigens in pancreatic ductal adenocarcinoma (PDAC) patients were found directly by MS-based discovery, the HLA immunoaffinity purification pipeline performed well, with a depth of up to approximately 11,000 HLA-I and 3,000 HLA-II reached. This also led the identification of multiple TAAs, such as mesothelin, mucin-1 and the cellular tumor antigen p53. Importantly, the high quality immunopeptidomics data from the Bassani-Sternberg lab was capitalized to gain insights into the hotspots of antigen presentation [305]. This knowledge benefitted the prioritization of predicted mutated neoantigens for the pilot trial. Overall, vaccine companies and research labs worldwide are working on both improving HLA predictions leveraging MS immunopeptidomics data, as well as designing vaccination strategies through direct MS-based identification of tumor antigens [200, 339-341]. Excitingly, and following on from the success of the prior pilot clinical trial, the presented method will be integrated into clinical trials this year at the CHUV, with the hope of pinpointing actionable neoantigens and elucidating their role in anti-tumor control in PDAC patients, as well as across other cancers.

6.2 Challenges In Non-Canonical Immunopeptidome Discovery

In this second section, the challenges of a proteogenomics-directed immunopeptidomics workflow for identification of non-canonical antigens are discussed, and the rationale of the current pipeline strategy is outlined. Thereafter, the most relevant validation approaches are highlighted, followed by insights into the overall findings. Finally, limitations and potential improvements of the approach are considered, as well as the prospective incorporation of the pipeline into clinical trials.

The discovery of mutated neoantigens, found either through prediction methodologies or by MS, have accelerated the rate of advances in cancer immunotherapy. However, for these mutated neoantigens, a personalized approach is indispensable due to the typically strict patient specificity. As the interest in mutated neoantigens grows [101], both patient-matched exome and RNA expression data, and tumor immunopeptidomics data are being accumulated. Recently, researchers are further exploring these and other new datasets to find common tumor targets that are likely shared across patients, or that could explain tumor editing and escape mechanisms [55, 160, 342-344]. As presumed non-coding genes, as well as TEs, could represent shared targets and have been shown to be tumor-specific and sometimes translated, there is a surge of interest in exploring the non-canonical space in the immunopeptidome [145].

This interest extends to the investigation of the clinical potential of MS-based proteogenomics, and has led to multiple different non-canonical antigen identification approaches being published over the last five years [250, 276, 290, 292, 345]. The resulting identification of noncHLAp were derived from proteasome-generated spliced variants, retroviral elements and novel ORFs. However, such studies are often accompanied with statistical and validity concerns, that can propagate the level of false positive identifications unless handled carefully. Of note, several commentaries and reviews regarding proteogenomics raise these concerns, and recommend thorough evaluation by any researcher in the field [271, 272, 274].

As such, in this work, a comprehensive MS-based proteogenomics and analytical workflow was developed and assessed, leading to the discovery and validation of tumor non-canonical antigens. We incorporated WES, bulk and single-cell transcriptomics, Ribo-Seq, and developed the computational tool NewAnce, that implements two MS/MS search tools in combination. Hundreds of noncHLAp were identified, with a selection being shared across patients and tumor-specific. Importantly, an immunogenic peptide derived from an alternative ORF of the melanoma stem cell marker gene ABCB5 was found.

6.2.1 Rationale Of The Pipeline Strategy

As there is currently no standardized approach to interrogate the non-canonical HLA repertoire, we set out to design an extensive analytical workflow, and discuss below the reasoning behind the choices implemented in the strategy. Specifically, the reasons are outlined for the use of certain patient-derived tumor samples, as well as the generation of customized reference databases for MS searches. Thereafter, the rationale leading to the exploration of specific non-canonical antigens, and the motivation behind the implementation of a novel computational pipeline, are described.

The use of Patient-derived Tumor Samples And Data

At the start of the study to identify tumor non-canonical peptides in the immunopeptidome, we reasoned that the reliability of any pipeline should be tested appropriately to determine the relevance of the discovered peptides in cancer immunotherapy. Therefore, the study was performed on samples with available autologous immune cells to allow for immunogenicity screening. Furthermore, in our proof-of-concept study, with the anticipation that experiments should be repeated, patient-derived cancer cell lines were sought out where appropriate. In order to heighten the chance of finding a non-canonical peptide, high mutational load tumors, such as melanoma and lung cancer samples, were chosen [126, 346]. We first compared whether a per-sample personalized reference based on RNA expression would improve the reliability of MS-based searches when compared to a generic non-personalized database of presumed non-coding regions. Indeed, it was observed that the non-personalized database was larger, and therefore led to an increase in the number of false positives, as indicated by the poorer percentage of non-canonical HLA binders. Consequently, the un-necessary exploration of "junk" was limited by personalizing protein reference databases to include only expressed non-coding regions of the patient. Specifically, WES, paired-end stranded RNA-Seq, and Ribo-Seq (for one sample) were performed to extract the information on variants, gene expression, and translation rate to increase the comprehensiveness of protein sequence databases. This setup allowed the broad exploration of tumor non-canonical antigens, and iterations to be performed where necessary.

The selection of Non-canonical antigens

The different origins of non-canonical antigens interrogated over the recent years have varied from one research group to the other, and most groups have focused on single categories in isolation. For example, the group of Van Allen focused on intron-retained epitopes, the group of Hwu on RNA edited epitopes, and the group of Chen-Harris on TEs (see Section 1.6.2) [290-292]. However, it is currently not possible to obtain a clear picture on which category has the greatest clinical potential, especially given the comparison issues between studies. Thus, to gain a comprehensive view in this work, as many non-protein-coding genes and potentially translated genomic regions were incorporated as possible. Specifically, this included expressed non-protein-coding genes, TEs, as well as non-canonical alternative ORFs, inferred from RNA and Ribo-Seq data. The search for phosphorylated peptides was additionally included, as these could represent interesting tumor-associated targets [104]. We believe that this is a first attempt at interrogating such a broad range of non-canonical antigens in one study.

Our MS-based discovery approach was heavily dependent on data generated by RNA-Seq. For this purpose, we chose to implement a stranded, paired-end and poly-A enrichment protocol for RNA-Seq for the following reasons. First, the stranded approach allowed the database size to be reduced by 3-fold (3-frame instead of 6-frame translation), as directional information could be obtained [277]. Second, a paired-end method increased the confidence in mappability of repetitive regions for TEs at the locus-specific level [347]. Third, Poly-A enrichment typically results in the identification of the majority of non-protein-coding genes [348]. However, a drawback to keep in mind is that some lncRNAs lack poly-A tails, and hence, their potential peptide products might have been missed with the presented approach. This challenge could be overcome in the future by employing ribosomal RNA depletion in place of poly-A enrichment protocols [349].

Performance Evaluation of MS search tools

Using the customized reference databases of expressed non-protein-coding genes and TEs from above, the routinely employed MaxQuant tool was initially used at 1% FDR, that led to the identification of a significant number of non-canonical peptides. However, the majority were determined to likely be false positives, demonstrating a low percentage of HLA binders, as well as poor correlation between observed RT and calculated HI. Furthermore, we systematically tested the generic and personalized database sizes based on

RNA expression (FPKM >0), as well as at medium and higher expression cutoffs (FPKM>2, >5, and >10). This showed that the level of HLA binders increased with smaller databases, improving the quality of the non-canonical peptide data obtained. Therefore, the trade-off between the level of false positives, and completeness of the database was re-considered. Importantly, when mapping back non-canonical sequences to source non-protein-coding genes, it was observed that many sequences were derived from lowly expressed genes, both in bulk and scRNA-Seq. Therefore, with regards to the aforementioned trade-off, we chose not to reduce the database size based on a gene expression threshold. This decision is in line with studies showing that RNA expression does not always dictate antigen presentation [266, 350]. Moreover, while various studies have filtered databases based on HLA binding predictions or tumor specificity [250, 285], these approaches were disregarded in this study so as to obtain more comprehensive insights into non-canonical antigen presentation.

Given that the reference database was left as complete as possible, and the false positive issues an individual search tool faces in this setting, a novel computational tool NewAnce was developed by Markus Müller in the group of Bassani-Sternberg by combining the results of two independent MS-based search tools, MaxQuant and Comet. To first guarantee that identified non-canonical peptides from both MaxQuant and Comet did not match any other proteome-derived sequences, peptides were aligned against a UniProt/TrEMBL database to categorize the sequences into either non-canonical or proteome-derived. NewAnce then considers the intersection of results from both tools, and implements separate FDR calculations for the proteome-derived and non-canonical peptide groups. This specifically limits the number of false identifications for the non-canonical peptide group at a given global FDR. Ultimately, considering common results from two search tools that have different scoring algorithms naturally increases the confidence that the identified novel peptide is not a false positive [351, 352].

Thus, the stringency of identifying novel peptides was handled by the utilized computational strategy, rather than limiting the database size. Overall, NewAnce reduced the number of identified non-canonical peptides by an order of magnitude (from 100s to 10s), while prioritizing superior specificity. This combination of a low number of peptides, that were identified with high confidence, brought operational benefits and enabled all of the peptides across the nine samples to be further interrogated downstream *in vitro*.

6.2.2 Pipeline Validation

After NewAnce was implemented for patient-derived tumor immunopeptidomics samples, the existence of the identified non-canonical peptides were evaluated. This is especially important when reporting novel, yet un-annotated peptide sequences [274]. To first ensure that the identified novel peptides could not be mapped back to other modified proteome-derived peptides, non-canonical peptide-spectrum-matches were re-searched by allowing the matching against six common variable modifications. This resulted in the majority of our peptides (at least 97%) to be correctly identified as novel. Thereafter, multiple complementary validation approaches were conducted, and a brief interpretation of these are presented here, followed by a more detailed discussion on targeted MS-based validation.

Assessment of HLA binding and peptide hydrophobicity

HLA binding is widely accepted to be an estimation of inherent quality in immunopeptidomics data, and should always be incorporated in any purity assessment. In our study, HLA binding predictions showed that the vast majority (median of 91.7%) of NewAnce-identified non-canonical peptides were predicted HLA binders. This is in comparison to the much lower levels of HLA binders identified with the MS search tools MaxQuant (median of 53.3%) and Comet (median of 76.9%) when used individually.

However, HLA peptides that are not predicted to, or weakly, bind have been shown to mediate important anti-tumor responses [251, 353-356]. Therefore, sequence-specific RT prediction, an approach that is more dependent on the overall sequence than the HLA motif, was used as a complementary assessment strategy. This approach has been used to evaluate non-canonical peptides. For example, regarding the post-translationally spliced peptides identified by Faridi et al. [284], a poor correlation between calculated HI and observed RT was reported and indicates false sequence assignment [272]. When NewAnce-identified non-canonical peptides were assessed with sequence-specific RT prediction, a highly positive correlation was found between the calculated HI and observed RT. Importantly, the distribution of the non-canonical peptides was not significantly different from that of the proteome-derived peptides, signifying the correct nature of the identifications across all samples. Importantly, the correlation between RT and HI was significantly better for the non-canonical peptides identified by NewAnce than those from either MaxQuant or Comet alone, providing greater confidence of correct sequence assignment and underlining the superiority of NewAnce.

Targeted PRM validation of non-canonical peptides

Commonly, researchers use synthetic peptides to separately confirm a previously identified sequence, through the correlation of similarities between endogenous and synthetic peptide MS/MS spectra [276, 357]. However, when such an approach is employed, the peptide effects in the original sample matrix is lost, and RTs of endogenous and synthetic peptides cannot be compared. Due to the loss of this information, it is possible that a higher level of false positive identifications can still be obtained.

In contrast, the chosen method was to validate peptides-of-interest using isotopically heavy-labelled synthetic peptides spiked back into the original sample. PRM-based validation of a particular sequence in this way is a two-step process that comes with advantages over the more common approach [358]. First, spike-in enables the heavy-labelled synthetic peptide to be analyzed in the original matrix, and RT between the endogenous and synthetic peptide should be equivalent due to their same physico-chemical properties. Second, the fragmentation patterns should be nearly identical, resulting in their unambiguous validation. Naturally, this approach is only possible if enough of the original sample material is retrospectively available.

In Manuscript 2 (Chapter 5), for the first time to our knowledge, a head-to-head comparison of validation rates was performed by PRM using a set of protHLAIp as a control (n= 71), along with lncRNA- and TE-derived HLAIp (n= 93) from melanoma sample 0D5P. While it was observed that the protHLAIp had higher validation rates than the non-canonical peptide groups, targeted MS was only able to validate approximately 70% of the protHLAIp. The lower percentage than theoretically expected could be due to several reasons. The un-validated peptides might represent false positives identified through the database-dependent search, or remain un-detected due to poor reproducibility between sample injections. Additionally, this could result from the poor ionization efficiency of peptides, low amounts of peptides being injected, and ultimately, instrument sensitivity [238]. In particular, in the presented study, PRM confirmation was shown to be largely dependent on the intensity of the precursor ion and detection over multiple injections. This indicates that the rate of peptide validation could potentially be raised by injecting a higher amount of peptides into the mass spectrometer.

Ultimately, PRM validation was approximately 50% for the non-canonical space. However, when the results from the complementary validation approaches described above are taken into consideration, we believe that most non-canonical peptide identifications from the discovery approach are correct.

6.2.3 Results From The Pipeline

Following the validation performed on the non-canonical peptides, two key sets of analyses, and their corresponding results, are further discussed in the following subsections. First, a Ribo-Seq approach was conducted to both validate the translation of RNA-Seq identified non-canonical peptides in the correct ORF, and acted as an additional discovery approach for non-canonical peptides. Second, tumor specificity and immunogenicity read-outs of the identified non-canonical peptides provide insights into clinical relevance, and are crucial in advancing this pipeline for the treatment of patients.

Ribo-Seq as a proxy to define the immunopeptidome

The sequencing of ribosome-protected mRNA fragments, known as Ribo-Seq, represents a very exciting approach giving direct evidence of translation of an ORF [288]. This approach can therefore drastically reduce RNA-Seq inferred protein sequence reference databases to filter for true translation products. Ribo-Seq was performed in the melanoma sample 0D5P, and compared to the RNA-Seq-based proteogenomics approach to determine if additional insights can be generated with this technique. Matching against a smaller database increases the number of identifications at a given FDR threshold. Thus, results showed that a Ribo-Seq inferred reference increased the number of identifications for the proteome-derived space, and led to the discovery of 56 additional non-canonical peptides for 0D5P. The results were further assessed by comparing whether HLA peptide sampling better correlated with the RNA expression levels or translation rates from Ribo-Seq. The correlation between gene expression and HLA presentation was indeed found to be weaker than the translation of proteins.

Furthermore, it was observed that Ribo-Seq skewed the identification towards more highly expressed source non-coding genes. This stems from the fact that for a translated ORF to be defined in our approach, it needs to have sufficient ribosome coverage and 3- nucleotide periodicity pattern across the transcript. As such, to detect the translation of specific low expressed genes, deeper sequencing with sufficient mapped reads covering that region could confirm the 3-nucleotide periodicity of a translated ORF [286, 359]. Ultimately, the approach validated the previously identified non-canonical peptides derived from the highly expressed non-coding genes (16 out of 77). However, unfiltered Ribo-Seq results showed that there was evidence of at least some mapped reads for the majority of the novel peptides identified by the RNA-Seq-inferred approach.

As the above results were obtained from a database inferred by Ribo-Seq, the method becomes an intriguing option to fully interpret immunopeptidomics data. However, the use of Ribo-Seq in the wider research field remains limited. Adoption of the method is hindered by extensive, complicated, and time-consuming protocols, a requirement for large amounts of initial cells, as well as the need for high sequencing throughput due to the short fragment sizes of ribosome-protected mRNAs [360]. With further protocol improvements that address these challenges, it is anticipated that Ribo-Seq will become an important strategy for proteogenomics analyses, bridging the gap between transcription and HLA presentation.

Tumor specificity of non-canonical peptides

An expressed antigen is tumor-specific when it is not observed anywhere in normal tissues, and thus restricted to the tumor. Therefore, targeting such antigens would not trigger on-target off-tumor effects and toxicity issues. With mutated neoantigens, the molecular changes in the tumor are inherently tumor-specific. In contrast, while non-mutated non-canonical peptides derived from a specific transcript and translation product might be aberrantly expressed in the tumor, there is a risk that these peptides are also expressed in healthy tissues [107, 361].

One approach to narrow down non-canonical tumor-specific targets in MS-based proteogenomics is accomplished by canceling out healthy RNA, thus resulting in tumor-specific data to search against [250]. For example, Laumont et al. used TECs as the healthy "control" and only considered RNA k-mers to be tumor-specific when detected in the tumor and not in the RNA of TECs. This approach could have resulted in missing information on the global presentation of non-canonical antigens, and potentially led to the discarding of relevant novel TAAs, depending on the thresholds applied. Importantly, researchers are still looking into exploiting self or shared TAAs for immunotherapy, due to their inherent potential in rapidly reaching large cohorts of patients, especially when on-target off-tumor effects are considered low risk. For example, the deletion of normal B cells is well tolerated by non-Hodgkin's lymphoma patients when targeting the TAA CD20 [362, 363].

Moreover, immunopeptidomics healthy tissue controls are currently still lacking. Therefore, in order to retain information on potential baseline and tumor-associated non-canonical peptides, we chose to interrogate tumor specificity retrospectively by comparing the source non-coding genes to expression profiles of publicly available healthy tissue RNA-Seq datasets (GTEx)[304]. Approximately 20% of lncRNA source genes were found to be tumor-associated, when compared to selected protein-coding source genes at similar expression levels. However, when looking into the ideal situation where healthy and matched tumor lung tissue from two patients were available, we found that the majority of the identified non-canonical peptides were patient-specific, while not necessarily tumor-specific. Notably, lncRNAs and TEs are usually expressed in a tissue- and cell type-specific manner [164, 348, 364], implying that the expression of non-canonical peptides should be ideally evaluated in the matched healthy patient tissue. Ultimately, despite this, the mechanisms of expression regulation for the non-coding genes are largely unknown and it remains challenging to ensure tumor specificity.

Immunogenicity of non-canonical peptides

Immune responses were detected against several canonical TAAs, including peptides derived from tyrosinase (TYR) and tyrosinase-related protein 1 (TYRP1) in melanoma patient 0D5P, and a melanocyte protein PMEL peptide in melanoma patient T1015A. Ultimately, we detected one non-canonical immunogenic peptide from 0D5P through the Ribo-Seq inferred approach, which was derived from an alternative ORF of the ABCB5 gene. The pre-existing responses were found in autologous TIL products and peripheral blood after *in vitro* stimulation. Alternative ORFs of coding genes have previously been shown to generate novel antigens that are immunogenic, especially in the context of TAAs [149, 151, 365, 366], as well as in the viral context, for example in HIV-I [367, 368].

Interestingly, the link between ABCB5 expression and melanoma occurrence has long been established, with ABCB5 acting as a stem cell marker and conferring chemotherapeutic resistance in malignant melanomas [301, 369]. Hotspot mutations in the canonical ABCB5 gene increase the proliferative and invasive properties of melanoma cells, thereby potentially acting as a tumor suppressor gene [370]. At the single-cell level, we found that ABCB5 was expressed in a specific subpopulation of cells, that included the co-expression of genes CTNNB1 and MITF. Together, there are indications that these three genes have an important role in maintaining a melanoma stem cell niche [301, 371]. The non-canonical ABCB5 peptide is shared across patients, specifically in three melanoma samples from our immunopeptidomics database ipMSDB. These findings highlight the potential of inhibiting tumor growth by immune targeting this epitope in melanoma stem cell subpopulations.

The lack of immunogenicity for the remainder of the identified 451 non-canonical peptides could be due to multiple reasons. For example, T cells may not respond because of their exhausted state after rapid expansion that in turn could lead to a loss of antigen-specific frequencies [372]. Furthermore, the prolonged propagation in culture could have had an altered antigen repertoire of the interrogated samples. Additionally, the lower expression of non-coding genes found in our study (when compared to protein-coding genes) would likely lead to the presentation of non-canonical peptides in low copy numbers. The overall low expression could also explain the scarcity of non-canonical HLAIIp found in our samples (four peptides in total). These low copy numbers may present a limiting factor for cross presentation and for the generation of an appropriate T cell repertoire to initiate a robust anti-tumor response. As observed in viral systems, the relative abundance of specific epitopes, as well as antigen copy numbers, have been shown to correlate with the magnitude of antigen-specific cytotoxic T cell responses [373, 374]. In line with this, pre-existing T cell repertoires that confer protection in mice were correlated with the cognate antigen copy number on the target cell surface [250].

Fortunately, the conducted *in vitro* immunogenicity assays do not fully exclude the possibility that these non-canonical antigens represent good targets for cancer immunotherapy. In some cases, previous studies have reported promising mutated neoantigens that mediate tumor rejection, despite their lack of measurable T cell reactivity *in vitro* [251, 375-377]. Therefore, the currently employed T cell-based assays might not comprehensively characterize all traits of rejection antigens, and further research is needed to evaluate the possibility of using noncHLAp to induce a protective immune response *in vivo*.

6.2.4 Limitations

Despite the encouraging results obtained from this study, the conceptually developed pipeline should continue to be regarded as a work in progress, requiring adaptation depending on the research questions asked. While several limitations have already been discussed throughout this chapter, special consideration is given below to the challenges of building a comprehensive database and the work needed to validate non-canonical peptides.

In view of developing a protein sequence database, our combined results revealed that less is in fact more. However, the comprehensiveness of the database might be improved when considering additional data sources. For example, Ribo-Seq presented an attractive approach to specifically pinpoint translated genomic regions. Nevertheless, the presented Ribo-Seq method excluded information on TEs. This is due to the fact that TEs are mostly repetitive, and the very short Ribo-Seq reads tend to map back to multiple regions in the genome, creating ambiguity. One viable option to include this potentially promising pool of non-canonical antigens would be to concatenate Ribo-Seq data with the three-frame translation of expressed TEs, pre-extracted from an RNA-Seq-based approach. Furthermore, another interesting category of non-canonical peptides could be those derived from alternative spliced variants, with a plethora of studies showing their increased RNA levels in cancer [160]. While these were not investigated in the research at hand, their identification may be possible through the implementation of sophisticated methods capable of computing splicing diversity with RNA-Seq data [378]. As 99% of the cancer specific somatic mutations are found outside of the exon-coding genome [379], whole genome sequencing (WGS) can offer opportunities to search for personalized mutated non-canonical peptides. However, in this study only approximately 1% of the variants identified through WES mapped back to non-protein-coding regions. Although resource intensive, WGS could be performed and integrated in order to customize the non-proteome-derived space for every patient based on their variants.

Regarding the experimental setup for MS-based targeted validation, PRM was performed in order to evaluate the robustness and report the confidence level when using the computational pipeline NewAnce. However, when taking into account caveats that would apply in the clinic, such as limited timeframe, costs, resources and the effort required for multidisciplinary strategies, the question regarding the most suitable validation strategy remains. PRM techniques require a significant amount of time for method development and execution, as well as extensive downstream analyses [358]. Thus, this extensive validation approach is not expected to be routinely feasible in the clinic, except in instances where the routine-targeting of a few (common) peptides is carried out. Nonetheless, given the robust and reliable performance of NewAnce and our confidence in the resulting output, we believe that the implementation of PRM-based validation might not be required in a clinical setting.

6.2.5 Future Perspectives

While a mutated neoantigen discovery pipeline from the Bassani-Sternberg lab is being integrated into the clinical trials at the CHUV in the near future [338], the incorporation of the work presented in this thesis is still pending for the routine evaluation of non-canonical peptides. In order to enable integration into a streamlined approach and support large-scale analyses, the independent steps of the current workflow will need to be combined. These currently distinct steps include the processing of MS data by two different MS search tools and the implementation of NewAnce, as well as the established TE identification method and Ribo-Seq method. Furthermore, additional refinements to the approach could include RT prediction [380], post-filtering via HLA binding, and proper integration with RNA- and Ribo-Seq results. As discussed in Section 6.2.1, the stringency in the current NewAnce method reduced non-canonical peptide identification from 100s to 10s for individual samples, thus, further development should be conducted to increase sensitivity without compromising specificity and ensure that interesting targets are not missed. When taken together, this future work should enable both rich insights and data-mining on tumor non-canonical peptides, bringing more knowledge in terms of potential stratifications in cancer and their immunogenic potential. For example, hotspots of non-canonical antigen presentation in the genome are yet to be uncovered. If data generated by the large-scale use of the described pipeline is brought together, these hotspots could be extracted, as seen in similar analyses for proteome-derived HLA peptides [305]. Ultimately, if correlated with immunogenicity information, this will provide an important aspect for non-canonical antigen prioritization.

Future work regarding the presented results is not limited to the incorporation of the pipeline into clinical trials. Importantly, we identified an immunogenic antigen from a novel alternative ORF of the melanoma stem cell marker ABCB5 **(Figure 8)**. Excitingly, there is an indication from our unpublished Ribo-Seq results that this novel ORF is also found across two other melanoma samples, underlining its potential to be shared. Given the critical role that this marker may play in cancer progression, it is vital for future work to evaluate its relevance across both cancer types and patients. For example, this could be explored by predicting the peptides that bind to the most common HLA alleles from the alternative ORF. These peptides should be synthesized in their heavy-labelled form and spiked back into a range of samples for targeted identification by MS. Furthermore, while our preliminary attempts have thus far been unable to sort the non-canonical ABCB5-peptide-specific TILs with multimers, other peptides originating from this novel ORF could be tested for immunogenicity and used to identify antigen-specific T cell populations. This would help explore whether ABCB5 non-canonical antigen-specific TCRs are shared among individuals by extracting information from further cellular analyses, and could shed light on its potential as a biomarker for diagnostic and prognostic applications.

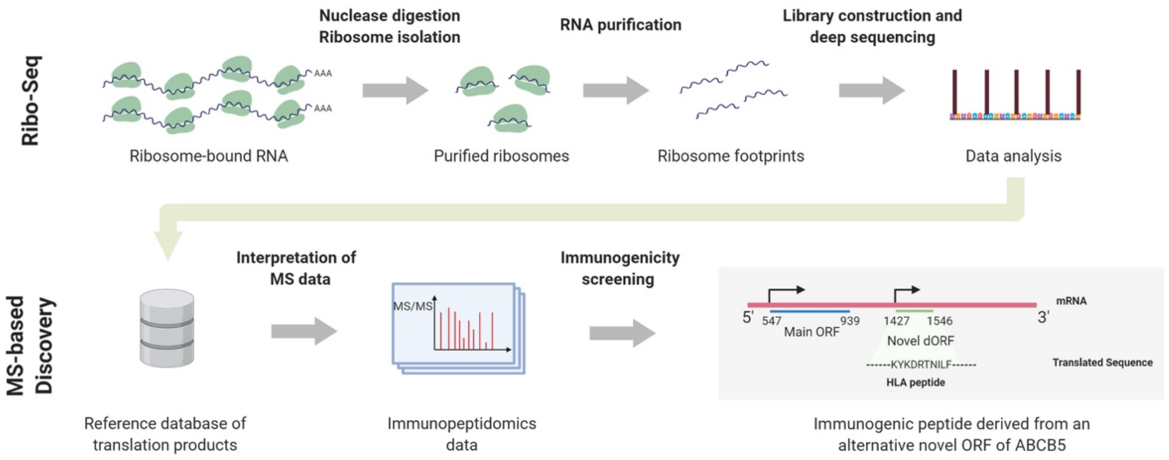


Figure 8 – Ribo-Seq directed immunopeptidomics. *The potential of Ribo-Seq to support interpretation of MS data is illustrated. Translating ribosomes are halted, for example, with elongation inhibitors. The mRNAs with occupied ribosomes are digested with nucleases to enrich and isolate for ribosome-protected fragments. The fragments are thereafter purified, resulting in ribosome footprints. Libraries are constructed and deep sequencing performed. Sophisticated computational analysis allows the mapping of ribosome footprints and the identification of regions with triplet periodicity. Statistical calculations are performed to define translated loci with high confidence. These translated ORFs are used to build a protein reference database. The database is utilized as input to interpret immunopeptidomics data in a database-dependent search. We identified with Ribo-Seq a novel ORF in ABCB5 in three melanoma samples, that led to the MS-based discovery of a non-canonical HLA peptide. Importantly, this peptide was found to be immunogenic in autologous immune cells, and thus represents a promising cancer immunotherapy target that warrants further exploration. Inspired by Hsu et al., PNAS, 2016 [381].*

Chapter 7

Concluding Remarks

Chapter 7 CONCLUDING REMARKS

Ultimately, the question remains: are non-canonical peptides relevant targets for cancer immunotherapy? To answer this, the characteristics of an ideal tumor antigen are considered. Specifically, both the antigen and source protein should be implicated in cancer, with their expression unique to tumors. Ideally, the antigen would be a HLA binder, recognized by autologous T cells, and shared across tumor types and patients. As discussed, our exploration of non-canonical peptides has found that they can be presented on tumors, are HLA binders, and could be immunogenic. One of the most intriguing findings presented here is the discovery that non-canonical peptides can indeed be shared across multiple tumor samples, and more frequently than nonsynonymous somatic mutations. This was validated by PRM and through analyses in ipMSDB, and indicates the potential to advance off-the-shelf treatment options. That being said, unlike mutated neoantigens, which are genuinely restricted to tumors, it is currently uncertain to what extent these non-canonical peptides are tumor-specific, and further research is needed on this topic.

Overall, through the extensive development of experimental and analytical methods that interrogate the tumor HLA repertoire, this thesis provides a comprehensive proteogenomics-directed immunopeptidomics framework **(Figure 9)**. These methodologies were optimized for the prioritization of actionable antigens as targets for cancer immunotherapy, for both pre-clinical and clinical pipelines. Consequently, meticulous and stringent assessments of the reported workflows were performed in order to affirm their robustness and validity, while concurrently providing deeper insights into the field of non-canonical antigen presentation. In conclusion, this work should help guide us towards the more informed prioritization of tumor antigens for personalized immunotherapy, an objective that we will continue working towards.

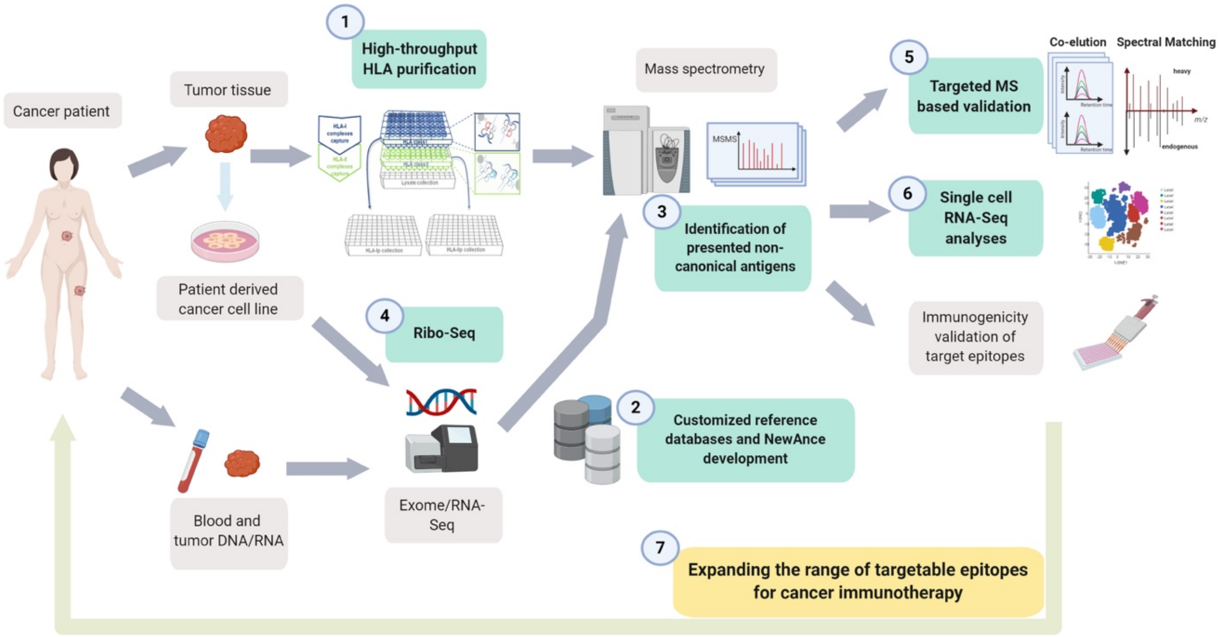

1
High-throughput HLA purification
Cancer patient
Tumor tissue
Patient derived cancer cell line
Mass spectrometry
MSMS
3
Identification of presented non-canonical antigens
4
Ribo-Seq
2
Customized reference databases and NewAnce development
Blood and tumor DNA/RNA
Exome/RNA-Seq
5
Targeted MS based validation
Co-elution
Spectral Matching
heavy
endogenous
m/z
6
Single cell RNA-Seq analyses
Immunogenicity validation of target epitopes
7
Expanding the range of targetable epitopes for cancer immunotherapy

Figure 9 - Contributions towards the MS-based proteogenomics workflow for cancer immunotherapy. *My research work presented in this thesis focused on (1) significantly improving the HLA peptide purification method, and (2) enabling deep antigen discovery with our in-house developed NewAnce computational tool. We generated patient customized databases to (3) identify tumor non-canonical antigens by MS. (4) Ribo-Seq was performed to gain deeper insights into the translation potential of source genomic regions, and additionally applied for MS-based non-canonical peptide discovery. (5) A substantial portion of non-canonical peptides were validated by targeted MS, and (6) their expression explored with scRNA-Seq. Combined, the integration of the steps in this workflow led to the development of a robust pipeline and (7) expanded the range of targetable epitopes for cancer immunotherapy.*

Chapter 8

Acknowledgments

Chapter 8 ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank Dr. Michal Bassani-Sternberg for providing me the opportunity to perform my PhD in her group for the last 4 years. Thank you very much for your supervision, and for your trust in me, which I highly appreciate. Thank you for helping me to grow into an independent scientist, and for all the confidence that you have given me during this time. I am very grateful to have been able to undertake my thesis in your group, and have learned a lot from you.

With the deepest gratitude, I would also like to thank Prof. George Coukos for giving me this unique opportunity to work in his lab and for the support that was always available when needed. I greatly appreciate the inspiration that you've provided to us all over the years to strive towards our scientific goals. Thank you very much for believing in me and in our work.

I would like to also acknowledge my PhD jury, with Prof. Fabio Martinon as my committee representative and president, and Dr. Manfredo Quadroni, Prof. Didier Trono and Prof Uwe Ohler who were present during my first-year and mid-thesis meetings. Thank you for all the advice and suggestions you provided to our scientific research along the way. In addition, I would like to warmly thank the experts for my PhD thesis, Prof. Yardena Samuels, Dr. Peter van Veelen and Dr. Marco Gerlinger, for travelling all the way to my defense and for taking the time to read, discuss and evaluate my PhD research. I would also like to sincerely thank all of the internal and external collaborators and Coukos lab members for their invaluable help and expertise over the years, which has certainly helped make this whole project a reality!

Next, I would like to warmly thank all of the past and present members of Michal's group. In particular, I would like to thank Justine Michaux for not only always being a fantastic partner-in-crime and an amazing source of encouragement in both the tough and not-so-tough moments, but for her constant scientific support throughout these years! We had a great many laughs and tears together, inside and outside of the lab, and I couldn't have done it without you (along with Daniel and Mystique)! I can always count on you guys, so thank you!

I'd also like to warmly thank Dr. Hui Song Pak for flawlessly taking care of both of the mass spectrometers and my sample injections over the years, and for always answering my (sometimes odd) questions on mass spectrometry! Next, I would like to thank Dr. Fabio Marino, from whom I have learnt a great deal about both life and work-related matters. I had many great times with you, and have missed you in the last year of my PhD. I couldn't have asked for a better partner to write a manuscript with, and thank you for always answering all my questions without judgement and for always having my back!

Also, a big thank you to Dr. Markus Müller, Florian Huber and Dr. Brian Stevenson, who have all been incredibly reliable and significant team members during my PhD, and with whom I have had many great discussions with and learnt a lot from. Thank you for always patiently providing me with support, and for all the great work we've done together! Lastly, I'd like to greatly thank Dr. Elodie Lauret Marie Joseph and Dr. Humberto Gomes Ferreira, whom, despite having only recently joined the group, I already cannot imagine not knowing. We've have all had many great times together as a group, and I hope that we will continue to do so!

I'd like to take this opportunity to thank some very special people that I have had the joy to meet during my PhD. Big hugs to the girl group with Julie Fierle, Marthe Solleder, Valentina Bianchi and Mariia Bilous. Together,

you girls have pulled me through this challenge and supported me fully. We shared so many fun times both in and out of the lab, and had some amazing adventures, sport sessions, brunches, lunches, and dinners that I will never forget. Without a doubt, I know that we will stay fast friends! Special thanks go to Julie Fierle and Fabian Sesterhenn, who have always been supportive in both my personal and private life, sharing many hobbies, ramen dinners, and our love for Whisky! I hope that we stay great friends over the many years to come.

Second to last, I'd like to express the greatest gratitude to my family, Mum, Papi and my little brother Sean, who have always unconditionally believed in me and supported me along the way. Spending time with them has always been incredibly joyful, leaving me relaxed and energized again when I went back to work. I couldn't have done it without you all. Also, I would like to thank all of my friends and relatives in Zurich and abroad, who have always been there for me whenever I needed them, and with whom I have shared many lovely moments between Zurich, Lausanne and Kuala Lumpur.

Lastly and most importantly, I'd like to thank Dr. Benjamin Ryder, who has been my rock during these past four years. You have supported me unfailingly all the way, believed in me when I didn't believe in myself, and gave me motivating pep talks on how to lean in, to feel confident about myself and to never give up. Your optimistic and joyful view on life is infectious, and this has always managed to make me smile and laugh in every situation. Thank you so much for your love, generosity, and support, and I couldn't have asked for a better person to have shared this journey with. I can't wait for our many more adventures together with Whisky!

Chapter 9

Appendix

Co-Author Manuscripts

Supplementary Information

Curriculum Vitae

Chapter 9 APPENDIX

9.1 Co-Author Manuscripts

During my time as a PhD student at the Ludwig Institute for Cancer Research at the University of Lausanne, I have contributed to many successful research projects that employed immunopeptidomics and proteomics approaches. The published manuscripts, along with those that are currently being prepared, are listed below. These were a result of intensive and rewarding collaborations, and my contributions to each are further detailed.

In Marino and Chong et al., 2019 [337], I co-developed a book chapter outlining the detailed methodology of our high-throughput HLA immunoaffinity purification pipeline for basic and translational applications.

In Racle et al., 2019 [178], I generated a significant amount of immunopeptidomics data, which were used for training prediction algorithms for HLA-II epitopes.

In Westergaard et al., 2019 [319], I significantly contributed to this work both experimentally and through subsequent downstream analyses, leading to the successful identification of immunogenic tumor-associated antigens in a patient-derived ovarian cancer cell line.

In Mylonas et al., 2018 [271], I assisted in the underlying research and analyzing the associated data, leading to the development of a workflow that accurately estimates the fraction of proteasome-generated spliced peptides found in the immunopeptidome.

In Bassani-Sternberg et al., 2017 [175], I generated a significant amount of immunopeptidomics data, which were used for training prediction algorithms for HLA-I ligands.

In Caron et al., 2017 [215], I documented the discussions and presentations from the immunopeptidomics community workshop in Zurich. This material enabled the highlighted gaps in immunopeptidomics research to be accurately conveyed, and was used for the Meeting Report.

In Bruand et al., manuscript in revision, I experimentally validated the differential signatures in the proteome that was found for BRCA1 knockdown versus control samples.

In Semilietof et al., manuscript in preparation, I performed immunopeptidomics experiments and analyzed the data obtained.

9.2 Supplementary Information

9.2.1 Manuscript 1

Supplemental Data for:

High-throughput and sensitive immunopeptidomics platform reveals profound IFNγ-mediated remodeling of the HLA ligandome

Chloe Chong[1,2*], Fabio Marino[1,2*], HuiSong Pak[1,2], Julien Racle[1,2,5], Roy T. Daniel[3], Markus Müller[4], David Gfeller[1,2,5], George Coukos[1,2], Michal Bassani-Sternberg[1,2#]

[1] Ludwig Institute for Cancer Research, University of Lausanne, 1066 Epalinges, Switzerland.

[2] Department of Oncology, University of Lausanne, 1015 Lausanne, Switzerland.

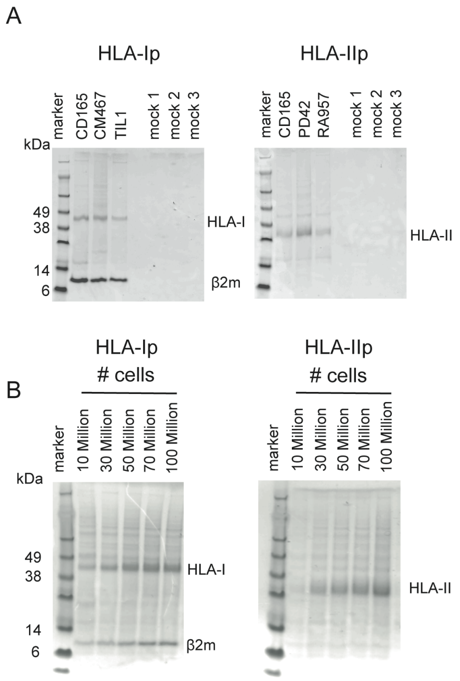
[3] Service of Neurosurgery, University Hospital of Lausanne, 1015 Lausanne, Switzerland.

[4] Vital IT, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

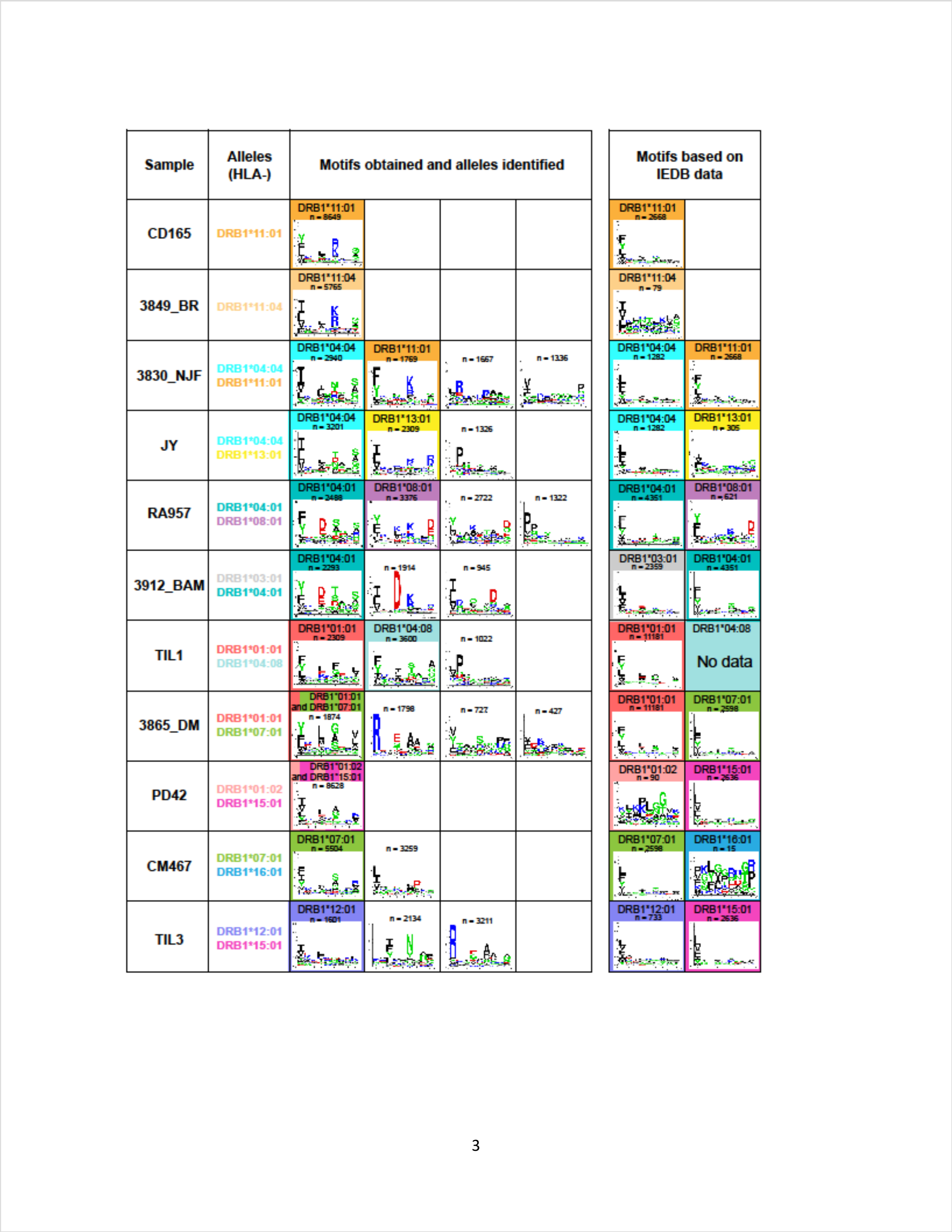
[5] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

[#] Correspondence author: Michal Bassani-Sternberg (Michal.bassani@chuv.ch)

*Equally contributing authors

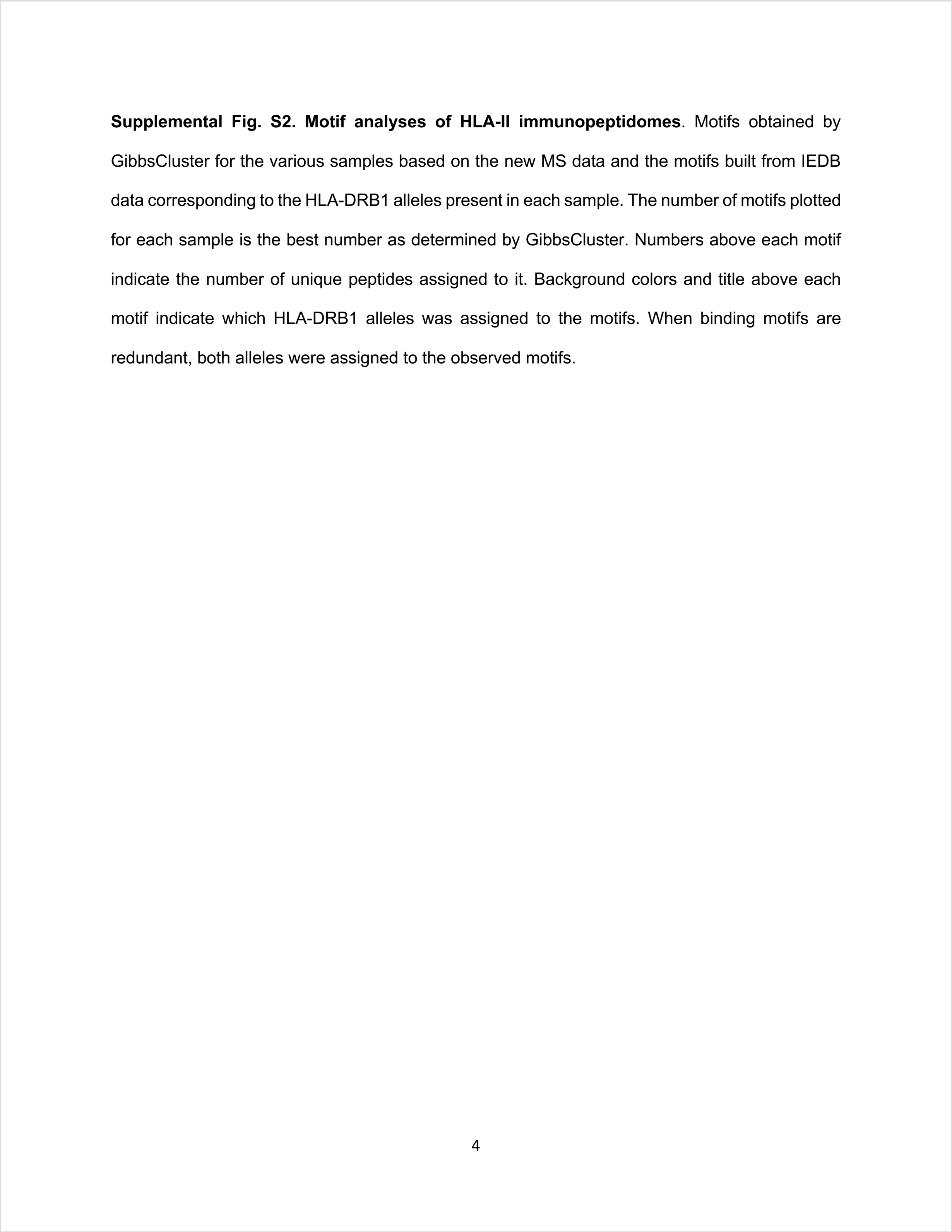


Supplemental Fig. S1. SDS-gel semi-quantification of recovered subunits of the HLA complexes. (A) Eluted HLA-I heavy chains and β2m molecules, and HLA-II heavy chains from lysates of selected cell line samples and the three mock samples, respectively. (B) Eluted HLA-I heavy chains and β2m molecules, and HLA-II heavy chains for lysate volumes corresponding to 10, 30, 50, 70 and 100 million CD 165 B-cells, respectively.



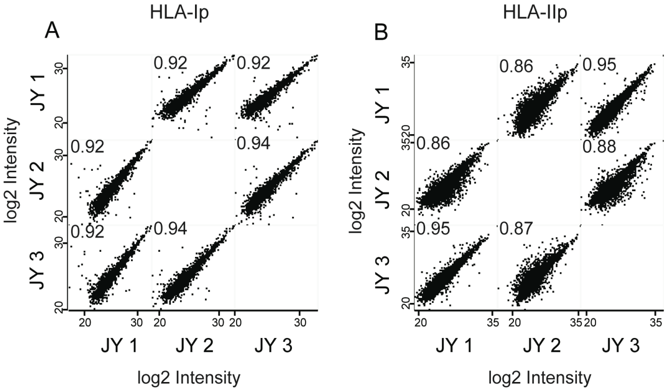
Sample	Alleles (HLA-)	Motifs obtained and alleles identified				Motifs based on IEDB data	
CD165	DRB1*11:01	DRB1*11:01 n = 8649				DRB1*11:01 n = 2668	
3849_BR	DRB1*11:04	DRB1*11:04 n = 5765				DRB1*11:04 n = 79	
3830_NJF	DRB1*04:04 DRB1*11:01	DRB1*04:04 n = 2940	DRB1*11:01 n = 1769	n = 1667	n = 1336	DRB1*04:04 n = 1282	DRB1*11:01 n = 2668
JY	DRB1*04:04 DRB1*13:01	DRB1*04:04 n = 3201	DRB1*13:01 n = 2309	n = 1326		DRB1*04:04 n = 1282	DRB1*13:01 n = 305
RA957	DRB1*04:01 DRB1*08:01	DRB1*04:01 n = 2488	DRB1*08:01 n = 3376	n = 2722	n = 1322	DRB1*04:01 n = 4351	DRB1*08:01 n = 621
3912_BAM	DRB1*03:01 DRB1*04:01	DRB1*04:01 n = 2293	n = 1914	n = 945		DRB1*03:01 n = 2359	DRB1*04:01 n = 4351
TIL1	DRB1*01:01 DRB1*04:08	DRB1*01:01 n = 2309	DRB1*04:08 n = 3600	n = 1022		DRB1*01:01 n = 11181	DRB1*04:08 No data
3865_DM	DRB1*01:01 DRB1*07:01	DRB1*01:01 and DRB1*07:01 n = 1874	n = 1798	n = 727	n = 427	DRB1*01:01 n = 11181	DRB1*07:01 n = 2598
PD42	DRB1*01:02 DRB1*15:01	DRB1*01:02 and DRB1*15:01 n = 8628				DRB1*01:02 n = 90	DRB1*15:01 n = 2636
CM467	DRB1*07:01 DRB1*16:01	DRB1*07:01 n = 5504	n = 3259			DRB1*07:01 n = 2598	DRB1*16:01 n = 15
TIL3	DRB1*12:01 DRB1*15:01	DRB1*12:01 n = 1601	n = 2134	n = 3211		DRB1*12:01 n = 733	DRB1*15:01 n = 2636

3

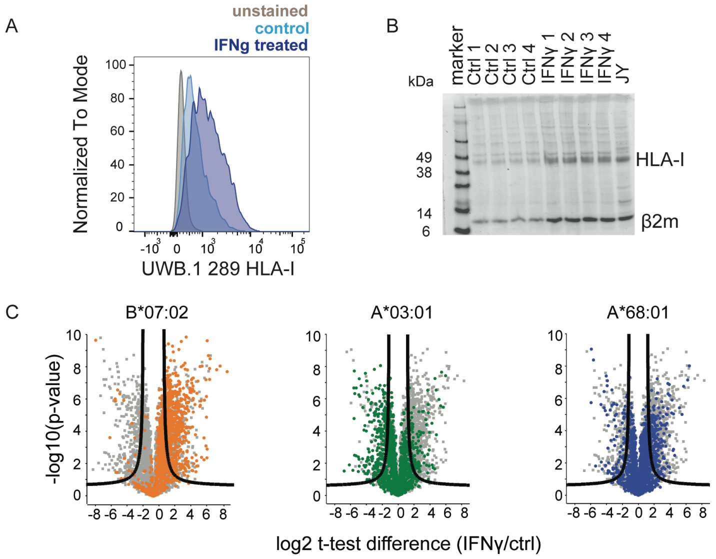


Supplemental Fig. S2. Motif analyses of HLA-II immunopeptidomes. Motifs obtained by GibbsCluster for the various samples based on the new MS data and the motifs built from IEDB data corresponding to the HLA-DRB1 alleles present in each sample. The number of motifs plotted for each sample is the best number as determined by GibbsCluster. Numbers above each motif indicate the number of unique peptides assigned to it. Background colors and title above each motif indicate which HLA-DRB1 alleles was assigned to the motifs. When binding motifs are redundant, both alleles were assigned to the observed motifs.

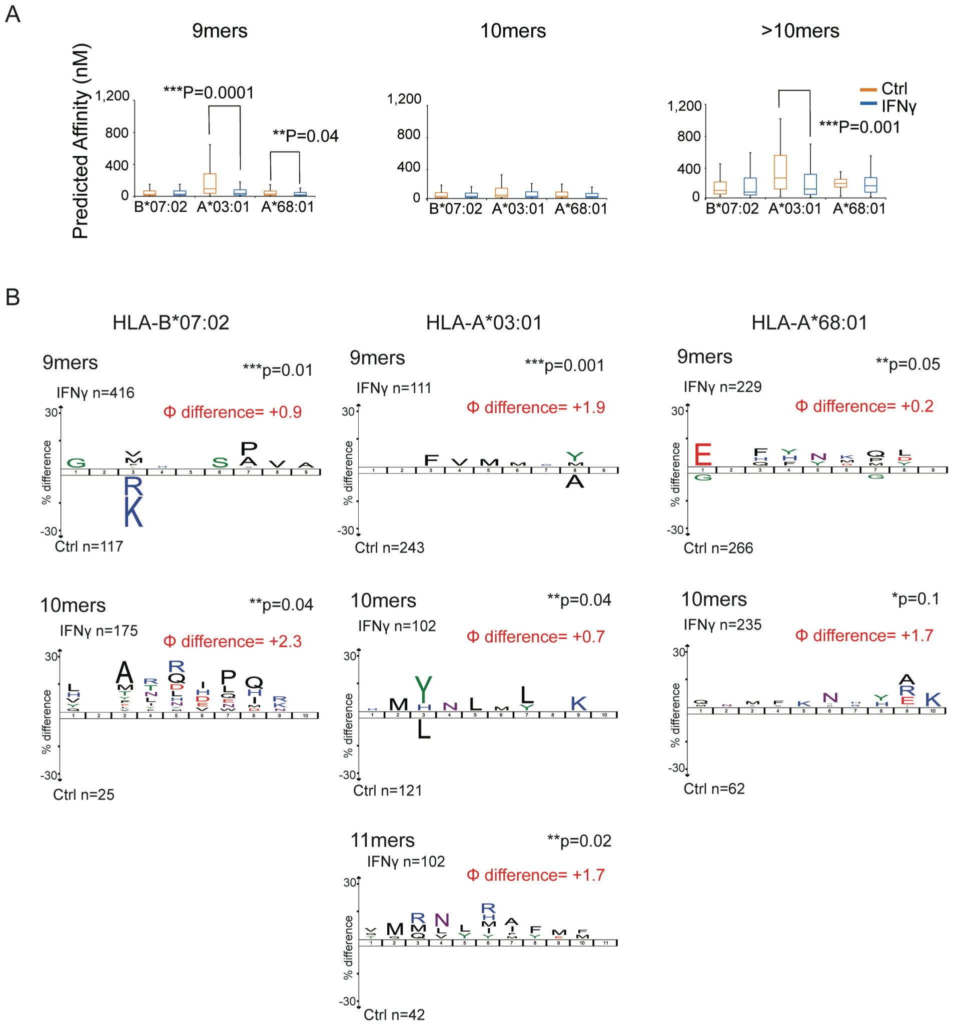
4



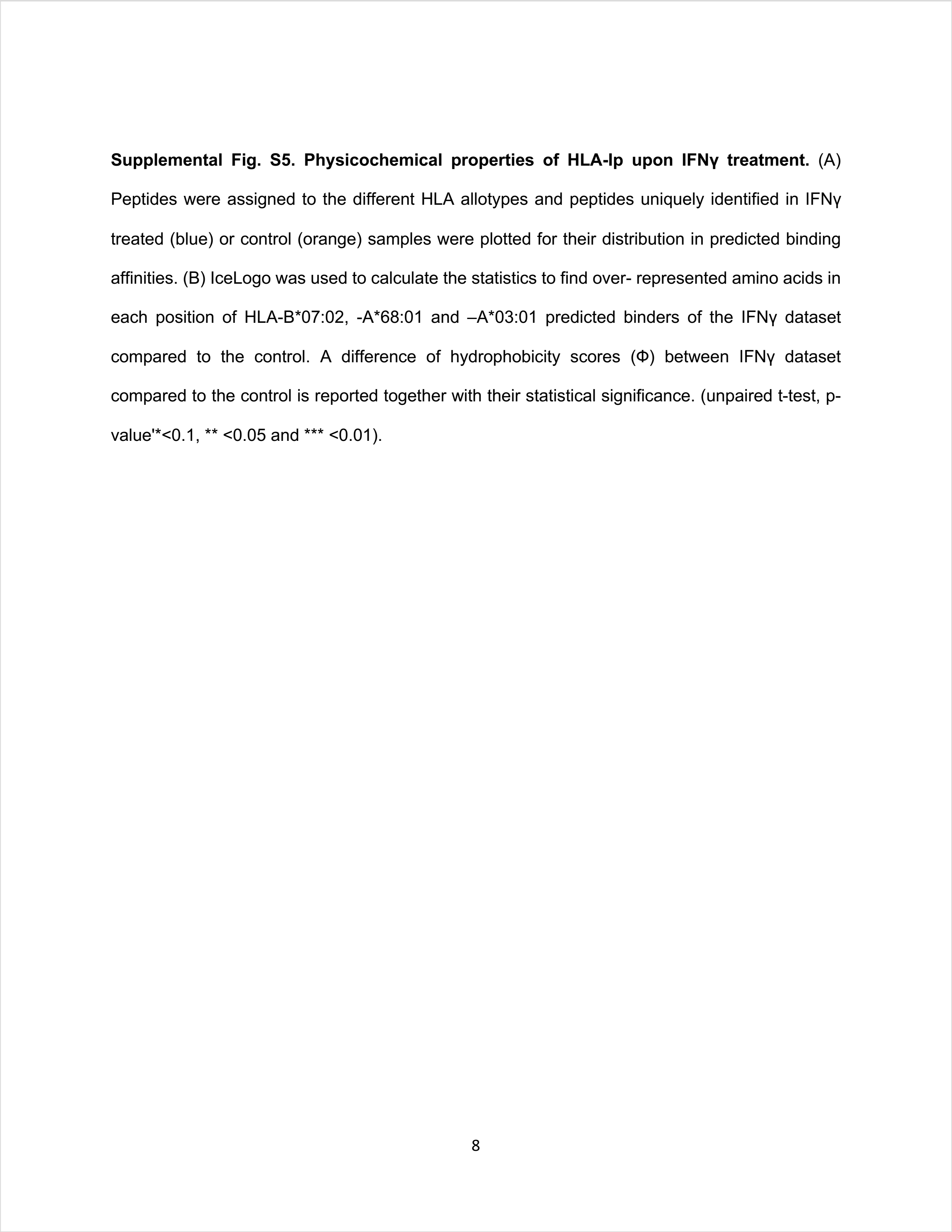
Supplemental Fig. S3. Excellent inter-plate reproducibility. (A) Inter-plate reproducibility calculated by Pearson correlations of Log2 transformed intensities of HLA-Ip and (B) HLA-IIp purified from JY cells on different days and with different stocks of reagents and plates.



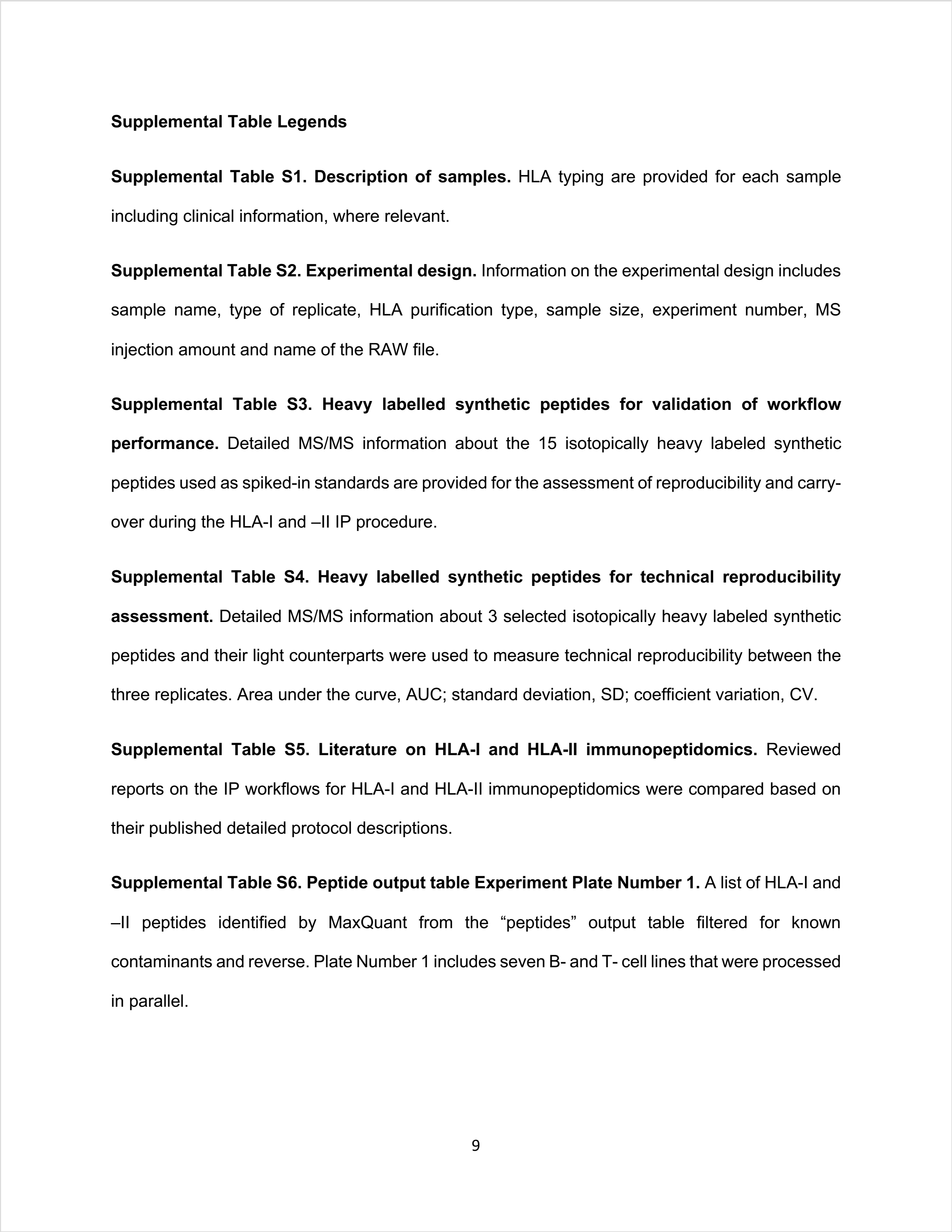
Supplemental Fig. S4. Increased expression of HLA-I complexes and allele-specific changes upon IFNγ treatment. (A) Increased expression of HLA-I on the surface of UWB.1 289 cells upon IFNγ treatment detected by FACS analysis. (B) SDS-gel semi-quantitative analysis confirmed global increase in intensities of HLA class I heavy chains and β2m molecules after IFNγ treatment. (C) Volcano plot summarizing unpaired t-test analysis of the immunopeptidome of IFNγ treated versus untreated cells. Peptides located above the lines are statistically significantly modulated in their level of presentation (FDR=0.01, S0=1). Peptides were colored on the exact volcano plot as follows: peptides predicted to bind the HLA-B*07:02 in orange, to the HLA-A*03:01 in green and to the HLA-A*68:01 in blue.


A
9mers
10mers
>10mers
Predicted Affinity (nM)
1,200
800
400
0
***P=0.0001
**P=0.04
***P=0.001
Ctrl
IFNγ
B*07:02
A*03:01
A*68:01
B
HLA-B*07:02
HLA-A*03:01
HLA-A*68:01
9mers
***p=0.01
IFNγ n=416
Φ difference= +0.9
Ctrl n=117
***p=0.001
IFNγ n=111
Φ difference= +1.9
Ctrl n=243
**p=0.05
IFNγ n=229
Φ difference= +0.2
Ctrl n=266
10mers
**p=0.04
IFNγ n=175
Φ difference= +2.3
Ctrl n=25
**p=0.04
IFNγ n=102
Φ difference= +0.7
Ctrl n=121
*p=0.1
IFNγ n=235
Φ difference= +1.7
Ctrl n=62
11mers
**p=0.02
IFNγ n=102
Φ difference= +1.7
Ctrl n=42
% difference
30
-30

7



Supplemental Fig. S5. Physicochemical properties of HLA-Ip upon IFNγ treatment. (A) Peptides were assigned to the different HLA allotypes and peptides uniquely identified in IFNγ treated (blue) or control (orange) samples were plotted for their distribution in predicted binding affinities. (B) IceLogo was used to calculate the statistics to find over- represented amino acids in each position of HLA-B*07:02, -A*68:01 and –A*03:01 predicted binders of the IFNγ dataset compared to the control. A difference of hydrophobicity scores (Φ) between IFNγ dataset compared to the control is reported together with their statistical significance. (unpaired t-test, p-value'*<0.1, ** <0.05 and *** <0.01).



Supplemental Table Legends

Supplemental Table S1. Description of samples. HLA typing are provided for each sample including clinical information, where relevant.

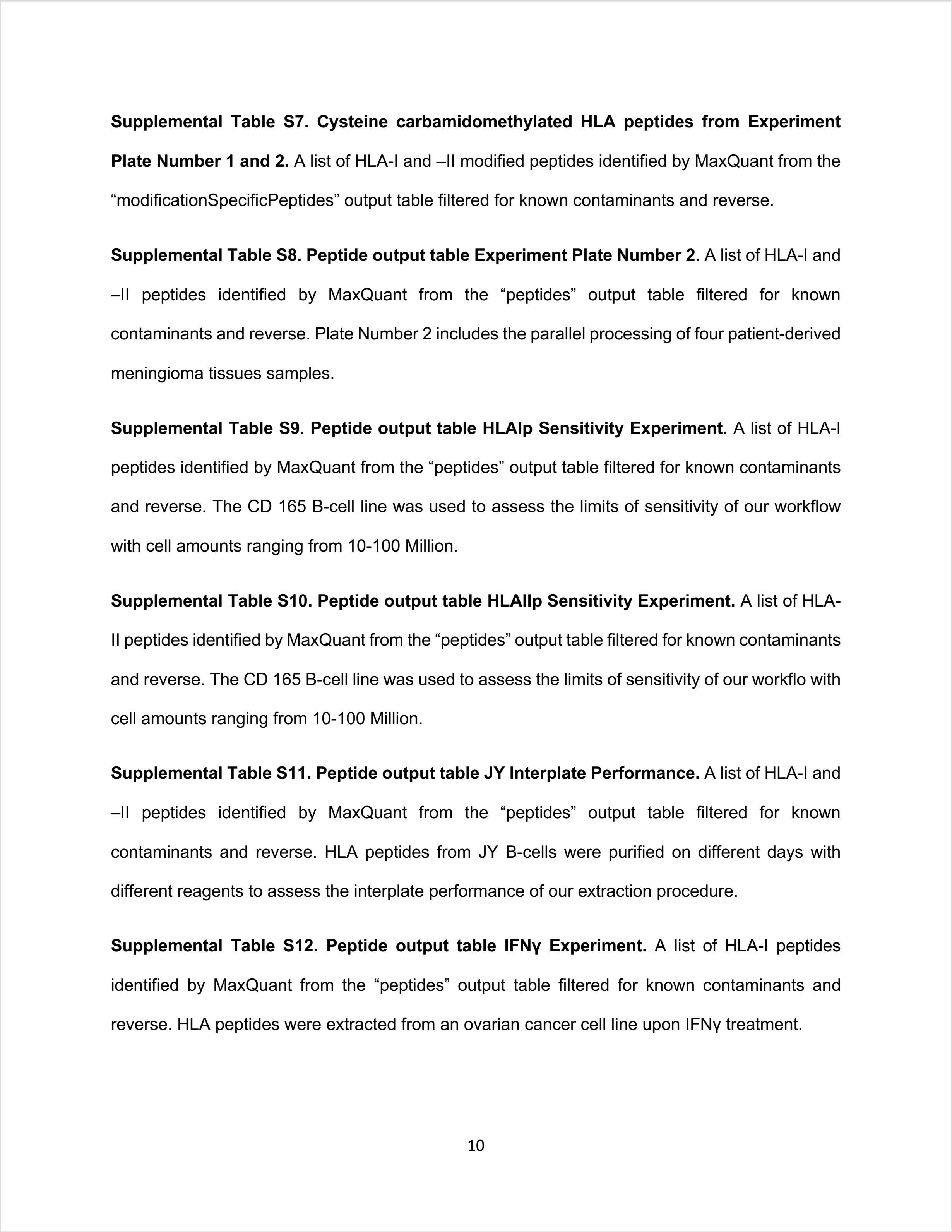
Supplemental Table S2. Experimental design. Information on the experimental design includes sample name, type of replicate, HLA purification type, sample size, experiment number, MS injection amount and name of the RAW file.

Supplemental Table S3. Heavy labelled synthetic peptides for validation of workflow performance. Detailed MS/MS information about the 15 isotopically heavy labeled synthetic peptides used as spiked-in standards are provided for the assessment of reproducibility and carry-over during the HLA-I and –II IP procedure.

Supplemental Table S4. Heavy labelled synthetic peptides for technical reproducibility assessment. Detailed MS/MS information about 3 selected isotopically heavy labeled synthetic peptides and their light counterparts were used to measure technical reproducibility between the three replicates. Area under the curve, AUC; standard deviation, SD; coefficient variation, CV.

Supplemental Table S5. Literature on HLA-I and HLA-II immunopeptidomics. Reviewed reports on the IP workflows for HLA-I and HLA-II immunopeptidomics were compared based on their published detailed protocol descriptions.

Supplemental Table S6. Peptide output table Experiment Plate Number 1. A list of HLA-I and –II peptides identified by MaxQuant from the "peptides" output table filtered for known contaminants and reverse. Plate Number 1 includes seven B- and T- cell lines that were processed in parallel.



Supplemental Table S7. Cysteine carbamidomethylated HLA peptides from Experiment Plate Number 1 and 2. A list of HLA-I and –II modified peptides identified by MaxQuant from the “modificationSpecificPeptides” output table filtered for known contaminants and reverse.

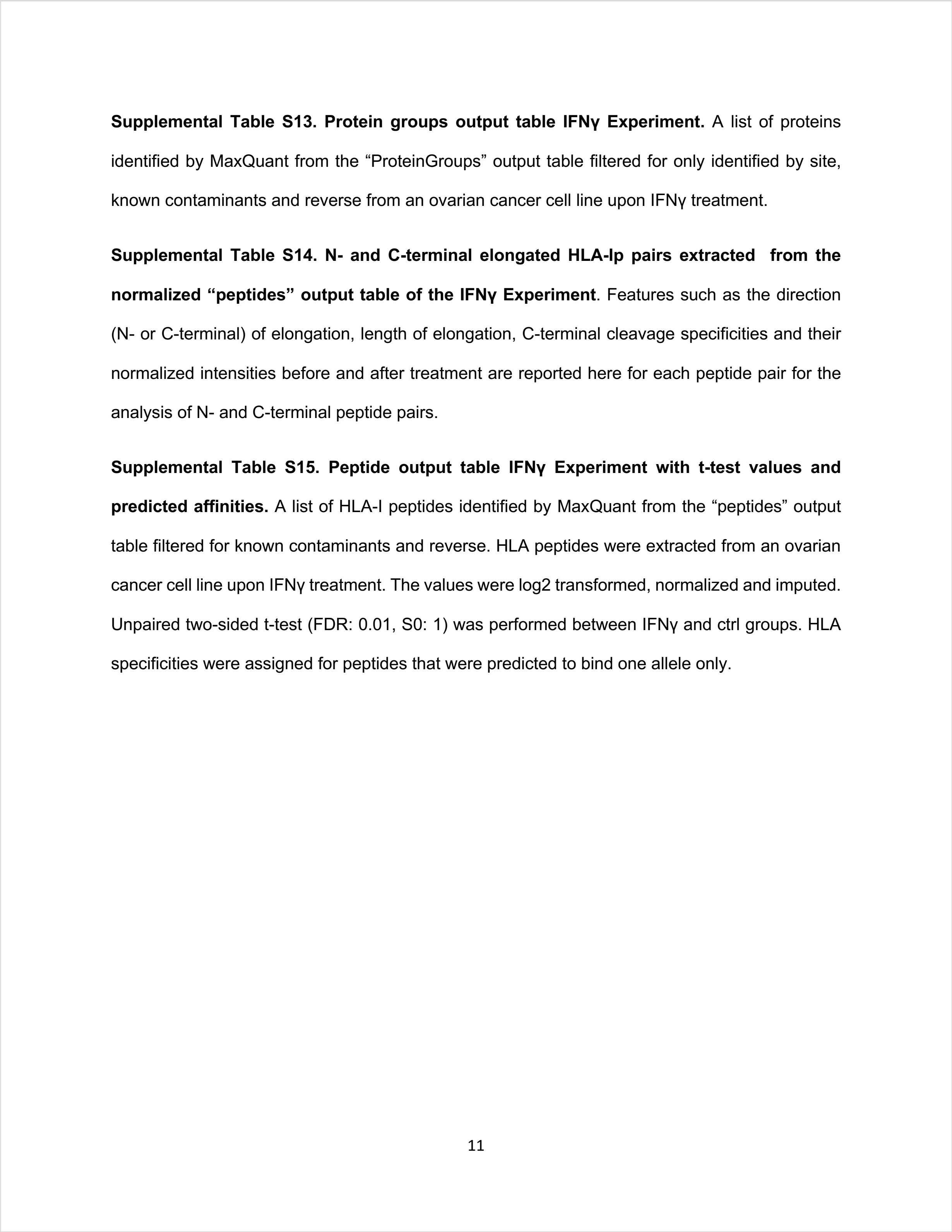
Supplemental Table S8. Peptide output table Experiment Plate Number 2. A list of HLA-I and –II peptides identified by MaxQuant from the “peptides” output table filtered for known contaminants and reverse. Plate Number 2 includes the parallel processing of four patient-derived meningioma tissues samples.

Supplemental Table S9. Peptide output table HLAIp Sensitivity Experiment. A list of HLA-I peptides identified by MaxQuant from the “peptides” output table filtered for known contaminants and reverse. The CD 165 B-cell line was used to assess the limits of sensitivity of our workflow with cell amounts ranging from 10-100 Million.

Supplemental Table S10. Peptide output table HLAIIp Sensitivity Experiment. A list of HLA-II peptides identified by MaxQuant from the “peptides” output table filtered for known contaminants and reverse. The CD 165 B-cell line was used to assess the limits of sensitivity of our workflo with cell amounts ranging from 10-100 Million.

Supplemental Table S11. Peptide output table JY Interplate Performance. A list of HLA-I and –II peptides identified by MaxQuant from the “peptides” output table filtered for known contaminants and reverse. HLA peptides from JY B-cells were purified on different days with different reagents to assess the interplate performance of our extraction procedure.

Supplemental Table S12. Peptide output table IFNγ Experiment. A list of HLA-I peptides identified by MaxQuant from the “peptides” output table filtered for known contaminants and reverse. HLA peptides were extracted from an ovarian cancer cell line upon IFNγ treatment.



Supplemental Table S13. Protein groups output table IFNγ Experiment. A list of proteins identified by MaxQuant from the “ProteinGroups” output table filtered for only identified by site, known contaminants and reverse from an ovarian cancer cell line upon IFNγ treatment.

Supplemental Table S14. N- and C-terminal elongated HLA-Ip pairs extracted from the normalized “peptides” output table of the IFNγ Experiment. Features such as the direction (N- or C-terminal) of elongation, length of elongation, C-terminal cleavage specificities and their normalized intensities before and after treatment are reported here for each peptide pair for the analysis of N- and C-terminal peptide pairs.

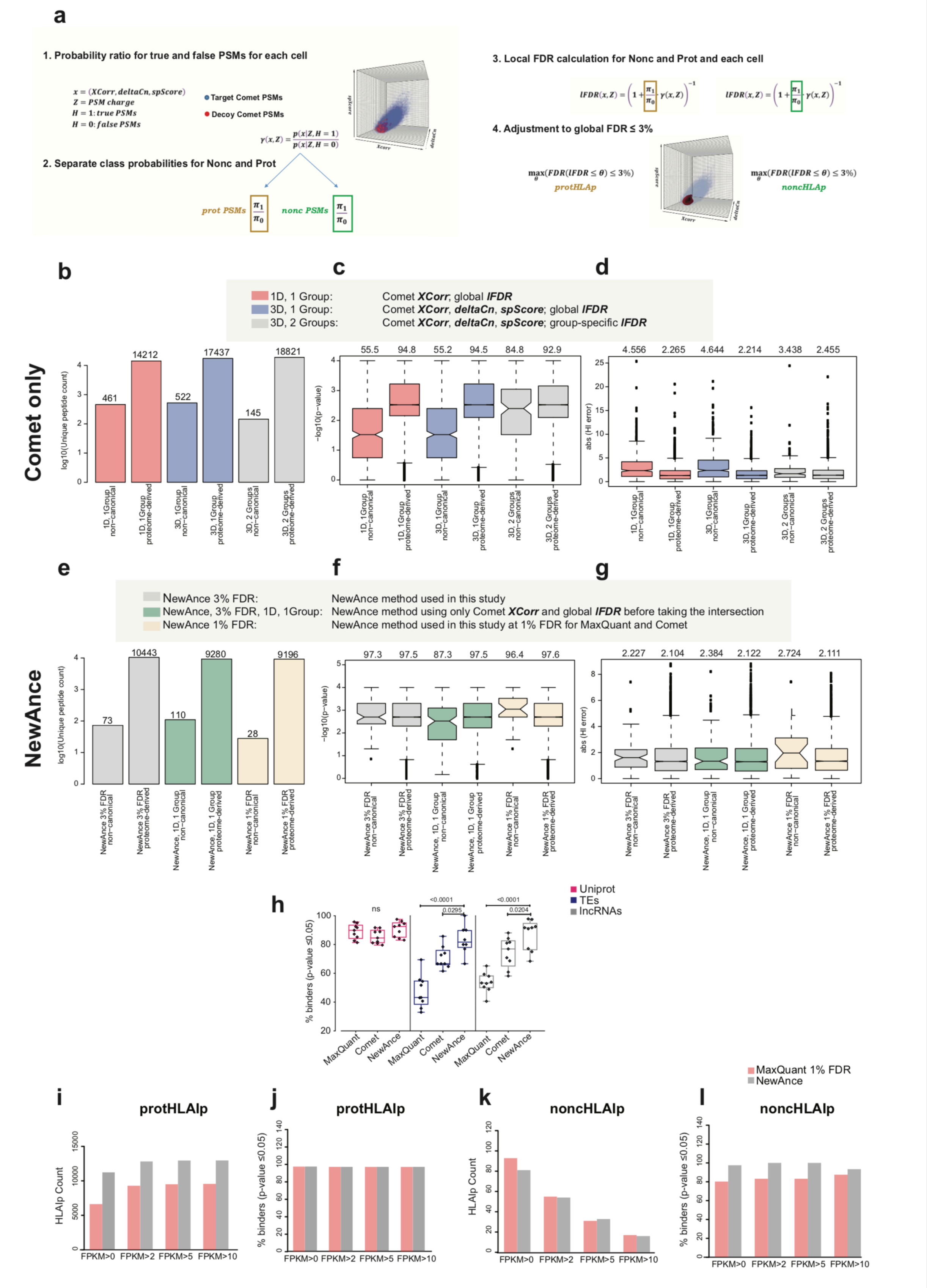
Supplemental Table S15. Peptide output table IFNγ Experiment with t-test values and predicted affinities. A list of HLA-I peptides identified by MaxQuant from the “peptides” output table filtered for known contaminants and reverse. HLA peptides were extracted from an ovarian cancer cell line upon IFNγ treatment. The values were log2 transformed, normalized and imputed. Unpaired two-sided t-test (FDR: 0.01, S0: 1) was performed between IFNγ and ctrl groups. HLA specificities were assigned for peptides that were predicted to bind one allele only.

9.2.2 Manuscript 2

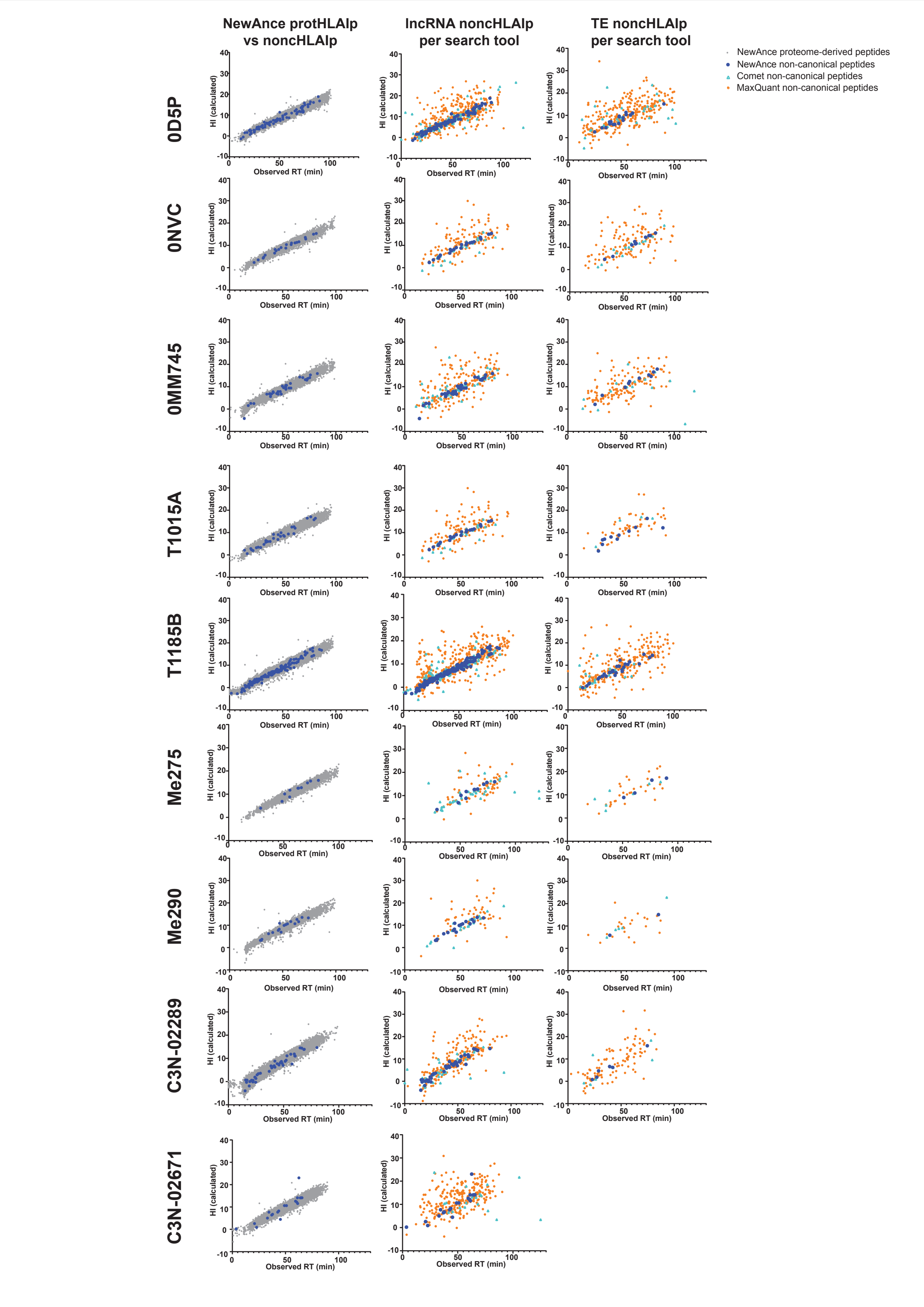
Supplementary Information

Integrated Proteogenomic Deep Sequencing and Analytics Accurately Identify Non-Canonical Peptides in Tumor Immunopeptidomes

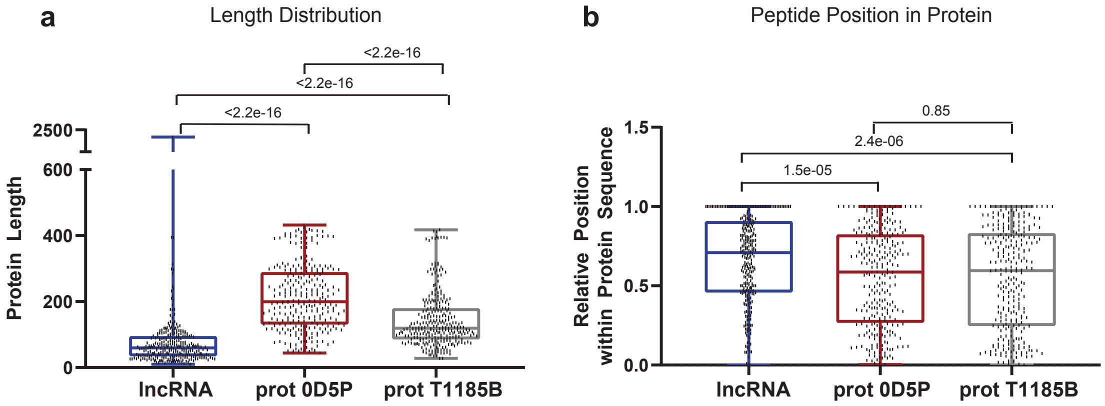
Chong et al.


a
1. Probability ratio for true and false PSMs for each cell
x = (XCorr, deltaCn, spScore)
Z = PSM charge
H = 1: true PSMs
H = 0: false PSMs
Target Comet PSMs
Decoy Comet PSMs
γ(x, Z) = p(x|Z, H = 1) / p(x|Z, H = 0)
2. Separate class probabilities for Nonc and Prot
prot PSMs π1/π0
nonc PSMs π1/π0
3. Local FDR calculation for Nonc and Prot and each cell
lFDR(x, Z) = (1 + π1/π0 γ(x, Z))^-1
4. Adjustment to global FDR ≤ 3%
max_θ(FDR(lFDR ≤ θ) ≤ 3%) protHLAp
max_θ(FDR(lFDR ≤ θ) ≤ 3%) noncHLAp
spScore
deltaCn
Xcorr
b
c
d
1D, 1 Group: Comet XCorr; global lFDR
3D, 1 Group: Comet XCorr, deltaCn, spScore; global lFDR
3D, 2 Groups: Comet XCorr, deltaCn, spScore; group-specific lFDR
Comet only
log10(Unique peptide count)
461
14212
522
17437
145
18821
1D, 1Group non-canonical
1D, 1Group proteome-derived
3D, 1Group non-canonical
3D, 1Group proteome-derived
3D, 2 Groups non-canonical
3D, 2 Groups proteome-derived
-log10(p-value)
55.5
94.8
55.2
94.5
84.8
92.9
abs (HI error)
4.556
2.265
4.644
2.214
3.438
2.455
e
f
g
NewAnce 3% FDR: NewAnce method used in this study
NewAnce, 3% FDR, 1D, 1Group: NewAnce method using only Comet XCorr and global lFDR before taking the intersection
NewAnce 1% FDR: NewAnce method used in this study at 1% FDR for MaxQuant and Comet
NewAnce
73
10443
110
9280
28
9196
NewAnce 3% FDR non-canonical
NewAnce 3% FDR proteome-derived
NewAnce, 1D, 1 Group non-canonical
NewAnce, 1D, 1 Group proteome-derived
NewAnce 1% FDR non-canonical
NewAnce 1% FDR proteome-derived
97.3
97.5
87.3
97.5
96.4
97.6
2.227
2.104
2.384
2.122
2.724
2.111
h
Uniprot
TEs
lncRNAs
% binders (p-value ≤0.05)
ns
<0.0001
0.0295
<0.0001
0.0204
MaxQuant
Comet
NewAnce
i
j
k
l
MaxQuant 1% FDR
NewAnce
protHLAlp
noncHLAlp
HLAlp Count
% binders (p-value ≤0.05)
FPKM>0
FPKM>2
FPKM>5
FPKM>10

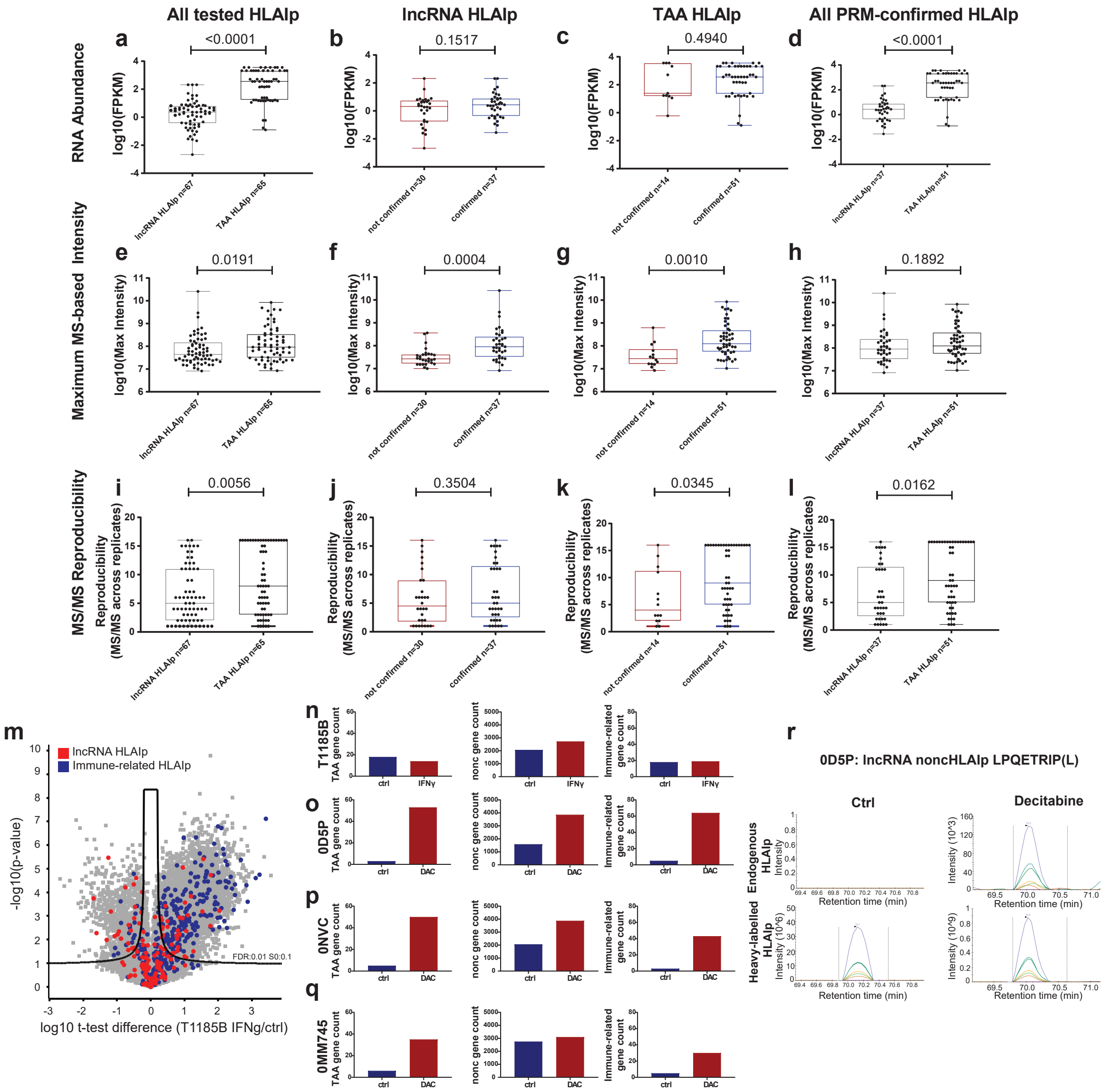
Supplementary Fig. 1 NewAnce for robust noncHLAIp identification. Group specific FDR calculation and the combination of two MS search tools enable the robust identification of noncHLAIp. **a** Schematic description of Comet FDR calculation workflow within NewAnce. 1) For PSMs of charge *Z*, the 3D Comet score space was divided into 40x40x40 cells. For every cell the probability ratios were calculated. 2) PSMs were split into non-canonical and proteome-derived groups and the class probability ratios were estimated for each group separately. 3) *lFDR* values were calculated for each cell and group. 4) Finally, the *lFDR* threshold corresponding to a global FDR of 3% was calculated and used to filter the PSMs. **b** The log10 of the number of unique peptides is shown for the comparison of 3 processing strategies tested for the identification of lncRNA- and proteome derived peptides at FDR of 3% in 0D5P sample. **c** The p-values for the MixMHCpred binding predictions are shown for the same comparisons as in b. The percentages of predicted binders (MixMHC p-values ≤ 0.05) are indicated as numbers above the boxplots. **d** The residual absolute errors of hydrophobicity index calculations by SSRCalc are shown for the same comparisons as in b. Standard errors are indicated as numbers above the boxplots. **e** The log10 of the number of unique peptides is shown for the comparison of 3 combiner options tested for the identification of lncRNA- and proteome derived peptides in the 0D5P sample. **f** The p-values for the MixMHCpred binding predictions are shown for the same comparisons as in e. The percentages of predicted binders (MixMHC p-values ≤ 0.05) are shown as numbers above the plots. **g** The residual absolute error of hydrophobicity index calculations by SSRCalc are shown for the same comparisons as in e. Standard errors are indicated as numbers above the boxplots. Please refer to the Methods section for boxplot parameters. **h** Systematic assessment of percentages of proteome-derived and predicted non-canonical HLA-I binders for each MS search tool (MaxQuant and Comet at FDR 3%) and NewAnce, for all n=11 samples, were performed. Ordinary one-way ANOVA, Sidak's multiple comparisons test was performed separately for Uniprot, TEs and lncRNAs. P-values between MaxQuant and NewAnce, and Comet and NewAnce are shown above the boxplots. ns: non-significant. **i-l** MaxQuant identified 0D5P prot- and noncHLAIp were compared to the NewAnce output, while reducing database sizes at FPKM thresholds. The total number of protHLAIp identified is plotted in **i**, and their corresponding percentage of predicted HLA binders in **j**. Similarly, the total number of noncHLAIp identified is plotted in **k** and the corresponding % of predicted HLA binders in **l**. Source data are provided as a Source Data file.


NewAnce protHLAIp vs noncHLAIp
lncRNA noncHLAIp per search tool
TE noncHLAIp per search tool
· NewAnce proteome-derived peptides
● NewAnce non-canonical peptides
▵ Comet non-canonical peptides
● MaxQuant non-canonical peptides
0D5P
0NVC
0MM745
T1015A
T1185B
Me275
Me290
C3N-02289
C3N-02671
HI (calculated)
Observed RT (min)
40
30
20
10
0
-10
0
50
100

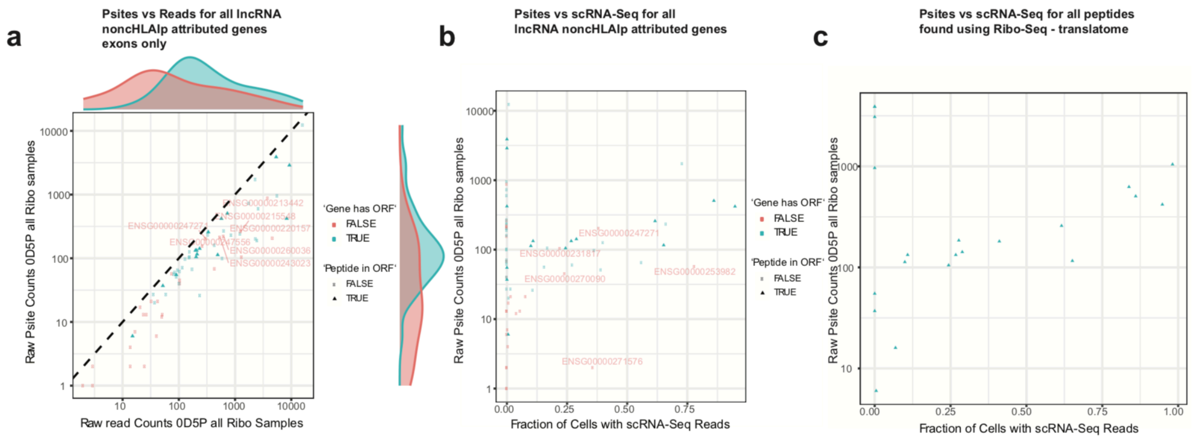
Supplementary Fig. 2 Hydrophobicity index calculation for all identified peptides. The data is shown for all assessed patient samples. In the leftmost panel, the observed mean retention time (RT) is plotted against the hydrophobicity indices (HI) for NewAnce-identified proteome-derived versus lncRNA-derived non-canonical peptides. All lncRNA-derived peptides (middle panel) or TE-derived peptides (rightmost panel) identified with each tool (MaxQuant, Comet, NewAnce) were analysed based on their hydrophobicity indices. Source data are provided as a Source Data file.



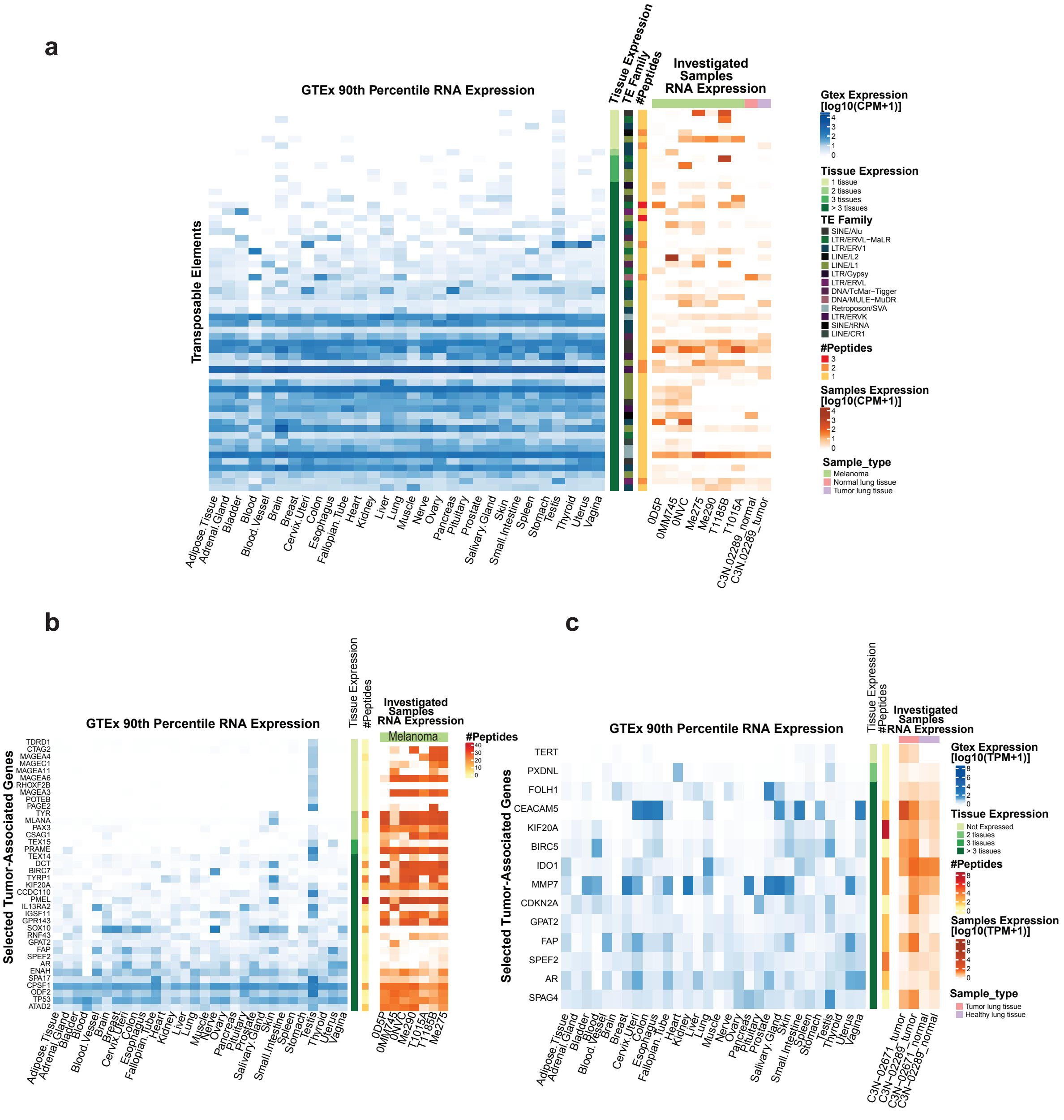
Supplementary Fig. 3 The origin of lncRNA-derived nonHLAps. LncRNA-derived noncHLAps are mainly derived from the C-terminus of the source translation products. **a** Protein length differences were assessed by sampling a matching-sized subset of both of the proteome-derived datasets fitting the length distribution of the lncRNA dataset (n=276) for a fair comparison. **b** Using the same dataset as that in (a), the corresponding HLA peptide's relative position (0 for N-terminus, 1 for C-terminus) was calculated for source lncRNA non-canonical and proteome-derived sequences. Statistical significance was performed with Wilcoxon testing. Please refer to the Methods section for boxplot parameters. Source data are provided as a Source Data file.



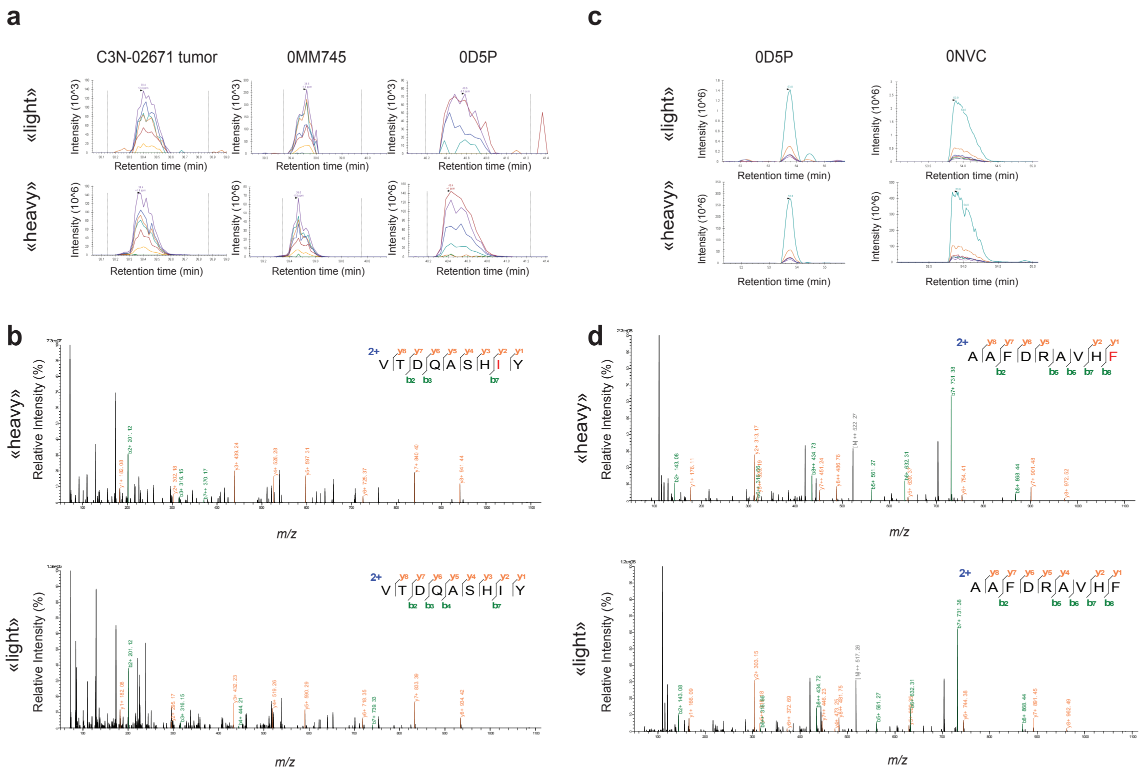
Supplementary Fig. 4 MS-based validation of noncHLAIp presentation and drug treatment effects. a-d Statistical analyses of MS-validated lncRNA-derived noncHLAIps and TAAs in the melanoma cell line 0D5P. The same comparisons were made first for all PRM-tested HLAIps regardless of the validation status (lncRNA HLAIp n=67, TAA HLAIp n=65); made second for lncRNA- (not confirmed n=30, confirmed n=37) and TAA HLAIps (not confirmed n=14, confirmed n=51) separately; and finally, made for only the PRM-confirmed HLAIps (lncRNA HLAIp n=37, TAA HLAIp n=51). RNA abundance in FPKM was extracted from the RNA-Seq data and compared within the corresponding groups. **e-h** MS-based intensity values were taken from the MaxQuant peptide output table and compared within the corresponding groups. **i-l** Last, MS/MS reproducibility, based on fragmentation by MS/MS per raw file (16 raw files for 0D5P in total), was analysed and compared within the corresponding groups. Unpaired two-sided t-test at 95% confidence interval. P-values are indicated above the plots. **m** Volcano plot depicting t-test analysis of HLAIps of IFNγ-treated versus untreated T1185B melanoma cells. Peptides located above the lines are statistically significantly up- or downregulated (FDR: 0.01, S0:0.1). All HLAIps derived from immunity-related genes are highlighted in dark blue, whereas all lncRNA-derived noncHLAIps are highlighted in red. **n** RNA expression analyses upon IFNγ treatment in T1185B. Induction of the total number of genes in per gene set was analysed for control and treated samples separately. The following gene sets were analysed: a selected set of TAA genes, all non-coding genes, and a subset of hypomethylating agent-induced immune-related genes (see main text). **o-q** Decitabine-treated melanoma cell lines were investigated at the RNA level. Only the total number of genes of interest that were exclusively expressed in each condition were taken into account. The same groups described above were analysed. For each gene category, decitabine induced the expression of more genes. **r** One example of a lncRNA-derived noncHLAIp that was induced by decitabine treatment in 0D5P, which was analysed by PRM. Co-elution of heavy and endogenous light transitions was found in only decitabine-treated samples. Source data are provided as a Source Data file.



Supplementary Fig. 5 Limits of detection by Ribo-Seq analysis. a Scatterplot showing processed P-sites vs Raw Ribo-Seq reads for the noncHLAps detected using RNA-Seq data. Triangles indicate peptides that were contained within an ORF with a periodic Ribo-Seq signal. The blue symbol indicates peptides originating from genes that contained at least one periodic Ribo-Seq signal. We had difficulty determining correct P-site offsets for some read lengths, as the mapping quality and other factors reduced the number of P-sites available for the detection of periodicity, with detection becoming difficult for genes with low expression; however, all noncHLAIps showed at least some raw Ribo-Seq signals. **b** P-sites vs scRNA-Seq for noncHLAIps detected using RNA-Seq data. With some exceptions due to imperfect mappability, genes with few P-sites tended to show low scRNA-Seq signals and were detected in few cells. Note that only one gene (ENSG00000247271 - labelled) showed more than 100 P-sites across all samples and was detected in more than 10% of the cells. **c** scRNA-Seq signals for HLAIps detected using a Ribo-Seq-derived translatome. NoncHLAIps (blue) detected using the Ribo-Seq translatome showed a higher rate of detection in scRNA-Seq experiments, again indicating that SaTAnn identifies ORFs that show reproducible evidence of translation. Source data are provided as a Source Data file.

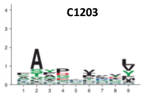
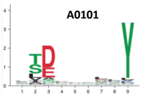
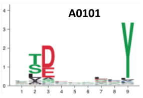
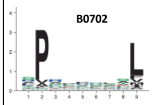
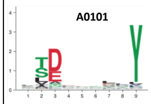
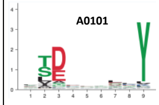
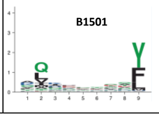
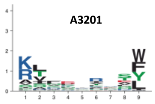


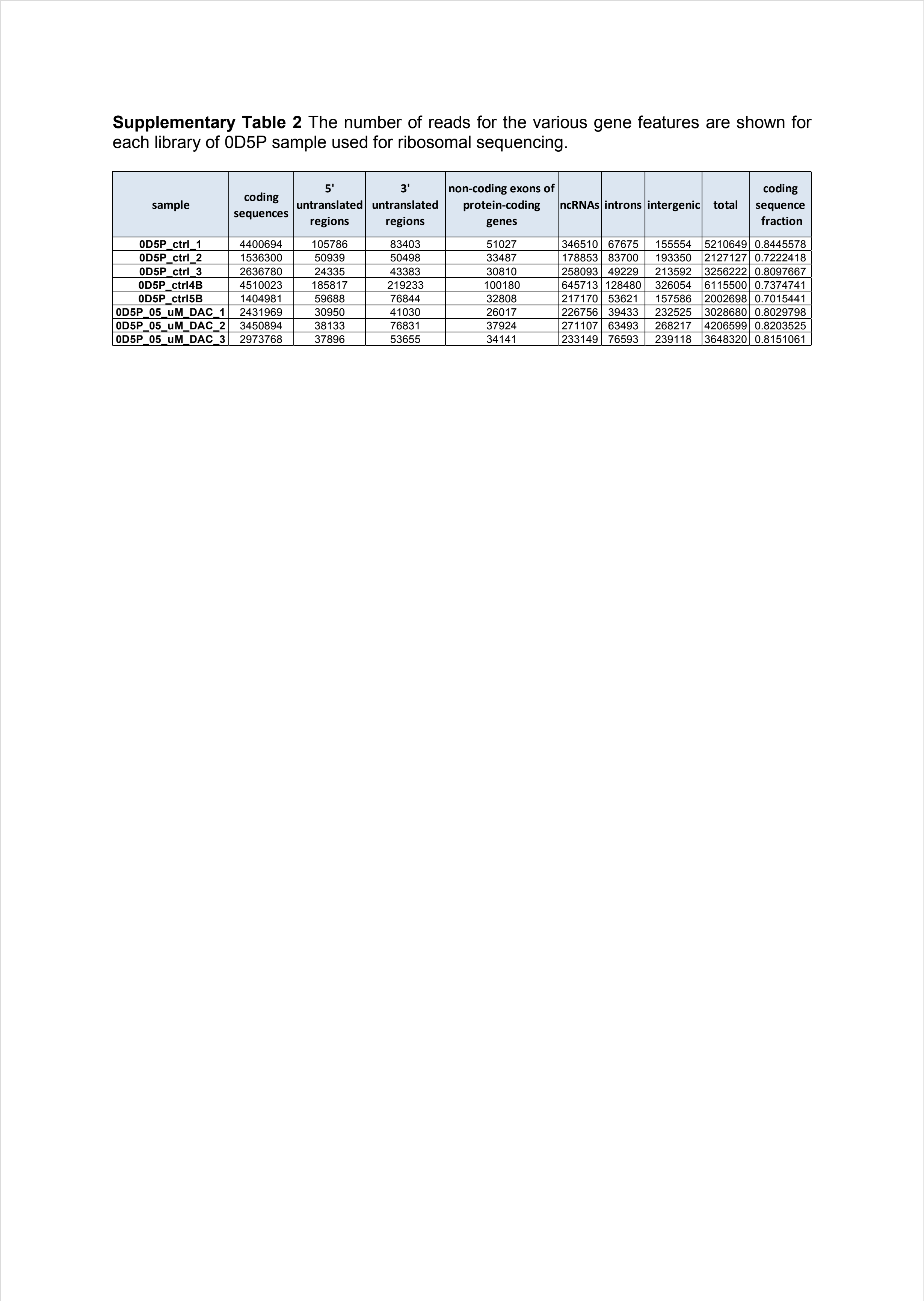
Supplementary Fig. 6 TE and tumor associated gene expression in healthy tissues. A comparison of presumed non-coding source gene expression in the investigated samples and healthy tissues (GTEx) for TE-derived HLAIps and TAAs. **a** Heatmap of TEs showing the 90th percentile gene expression levels in 30 healthy tissues on the left and the TE expression levels in our investigated samples on the right. Samples were classified as not expressed (90th percentile CPM ≤ 1) in any, 1-3, or more than 3 tissues other than testis to assess tumor specificity. The TEs were classified according to their TE family. Ten percent of the noncHLAIps derived from TEs were found to be expressed in only a single healthy tissue excluding the testis. **b** The same data described above are also plotted for selected source tumor-associated protein-coding genes in melanoma samples and for **c** lung tissue samples in TPM. Source data are provided as a Source Data file.



Supplementary Fig. 7 NoncHLAIp presentation can be shared across individuals. a Elution profiles of light and heavy labelled transitions and **b** representative MS/MS fragmentation pattern for the noncHLAIp VTDQASHIY. **c-d** The same representation is shown for TE-HLAIp AAFDRAVHF. Source data are provided as a Source Data file.

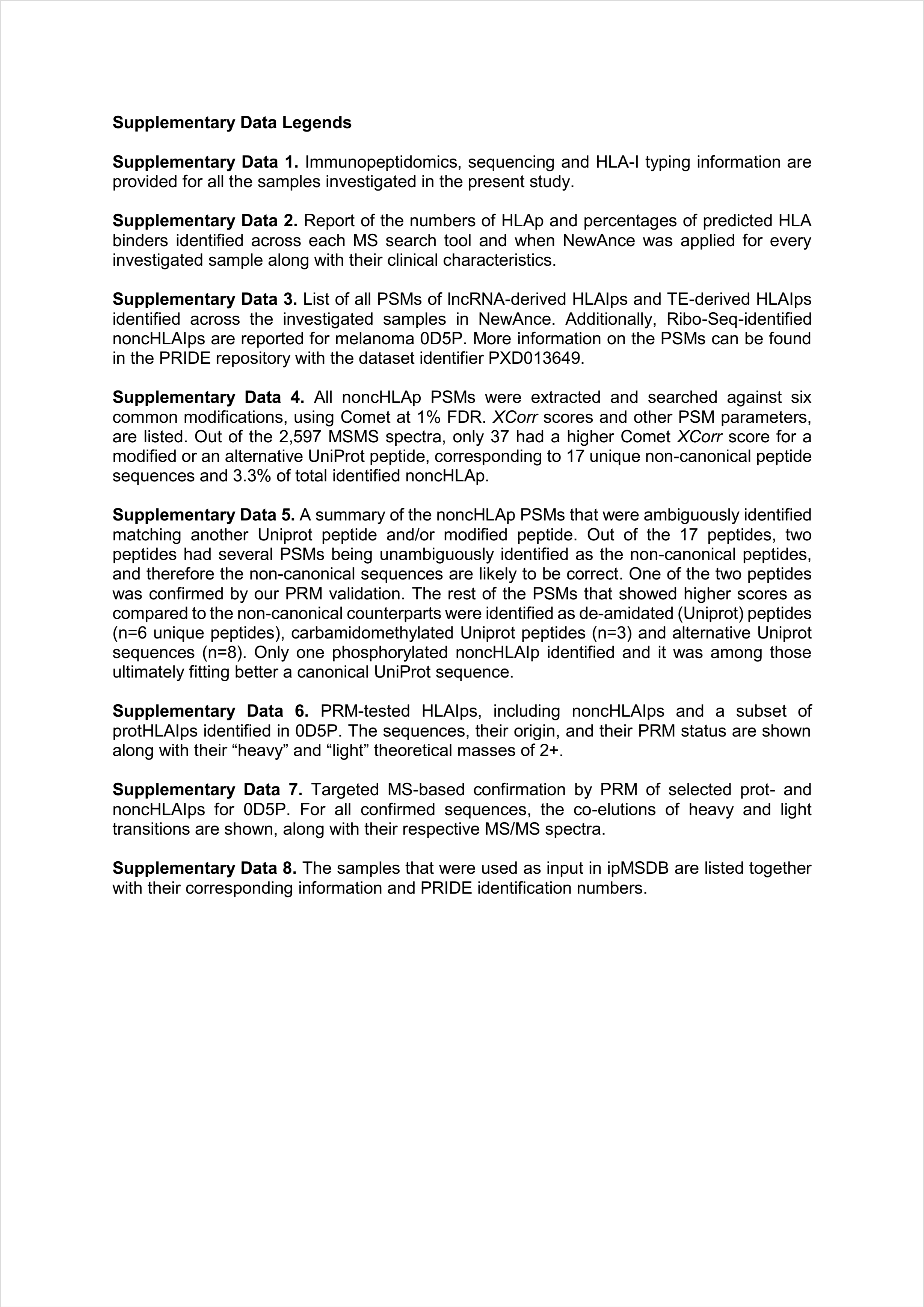
Supplementary Table 1 PRM-confirmed noncHLAIps that are shared across different samples. These patients express HLA allotypes that have identical or highly similar binding specificities.

Class	HLAIp	0D5P	0NVC	0MM745	C3N02671	Me275	T1015A	Motif	Motif
TE/lncRNA	**AAFDRAVHF**	C1203	C1203						
lncRNA	**VTDQASHIY**	A0101		A0101	A0101				
lncRNA	**KSDLSKPLSY**	A0101			A0101				
lncRNA	**APKSSSGFSL**	B0702				B0702			
lncRNA	**YLDPAQQNLY**	A0101			A0101				
lncRNA	**ETDIEMETRY**	A0101		A0101	A0101				
TE	**KVFKNGNAF**	B1501					A3201		



Supplementary Table 2 The number of reads for the various gene features are shown for each library of 0D5P sample used for ribosomal sequencing.

sample	coding sequences	5' untranslated regions	3' untranslated regions	non-coding exons of protein-coding genes	ncRNAs	introns	intergenic	total	coding sequence fraction
0D5P_ctrl_1	4400694	105786	83403	51027	346510	67675	155554	5210649	0.8445578
0D5P_ctrl_2	1536300	50939	50498	33487	178853	83700	193350	2127127	0.7222418
0D5P_ctrl_3	2636780	24335	43383	30810	258093	49229	213592	3256222	0.8097667
0D5P_ctrl4B	4510023	185817	219233	100180	645713	128480	326054	6115500	0.7374741
0D5P_ctrl5B	1404981	59688	76844	32808	217170	53621	157586	2002698	0.7015441
0D5P_05_uM_DAC_1	2431969	30950	41030	26017	226756	39433	232525	3028680	0.8029798
0D5P_05_uM_DAC_2	3450894	38133	76831	37924	271107	63493	268217	4206599	0.8203525
0D5P_05_uM_DAC_3	2973768	37896	53655	34141	233149	76593	239118	3648320	0.8151061



Supplementary Data Legends

Supplementary Data 1. Immunopeptidomics, sequencing and HLA-I typing information are provided for all the samples investigated in the present study.

Supplementary Data 2. Report of the numbers of HLAp and percentages of predicted HLA binders identified across each MS search tool and when NewAnce was applied for every investigated sample along with their clinical characteristics.

Supplementary Data 3. List of all PSMs of lncRNA-derived HLAIps and TE-derived HLAIps identified across the investigated samples in NewAnce. Additionally, Ribo-Seq-identified noncHLAIps are reported for melanoma 0D5P. More information on the PSMs can be found in the PRIDE repository with the dataset identifier PXD013649.

Supplementary Data 4. All noncHLAp PSMs were extracted and searched against six common modifications, using Comet at 1% FDR. *XCorr* scores and other PSM parameters, are listed. Out of the 2,597 MSMS spectra, only 37 had a higher Comet *XCorr* score for a modified or an alternative UniProt peptide, corresponding to 17 unique non-canonical peptide sequences and 3.3% of total identified noncHLAp.

Supplementary Data 5. A summary of the noncHLAp PSMs that were ambiguously identified matching another Uniprot peptide and/or modified peptide. Out of the 17 peptides, two peptides had several PSMs being unambiguously identified as the non-canonical peptides, and therefore the non-canonical sequences are likely to be correct. One of the two peptides was confirmed by our PRM validation. The rest of the PSMs that showed higher scores as compared to the non-canonical counterparts were identified as de-amidated (Uniprot) peptides (n=6 unique peptides), carbamidomethylated Uniprot peptides (n=3) and alternative Uniprot sequences (n=8). Only one phosphorylated noncHLAIp identified and it was among those ultimately fitting better a canonical UniProt sequence.

Supplementary Data 6. PRM-tested HLAIps, including noncHLAIps and a subset of protHLAIps identified in 0D5P. The sequences, their origin, and their PRM status are shown along with their "heavy" and "light" theoretical masses of 2+.

Supplementary Data 7. Targeted MS-based confirmation by PRM of selected prot- and noncHLAIps for 0D5P. For all confirmed sequences, the co-elutions of heavy and light transitions are shown, along with their respective MS/MS spectra.

Supplementary Data 8. The samples that were used as input in ipMSDB are listed together with their corresponding information and PRIDE identification numbers.

9.3 Curriculum Vitae

Chloe Chong

Avenue de Crousaz 11, 1010 Lausanne
+ 41 (0) 79 576 30 79 – chloechong91.cc@gmail.com

Research Experience

2016 - Present
Doctoral Researcher
Ludwig Institute for Cancer Research, Immunopeptidomics Lab, CHUV/UNI Lausanne, CH
Headed by: Prof. George Coukos, doctoral co-supervisor: Dr. Michal Bassani-Sternberg
Research focuses on developing a high-throughput immunopeptidomics platform for cancer immunotherapy applications, and advancing proteogenomics approaches for novel immunogenic targets

2015 - 2016
Graduate Research Assistant
Swiss Federal Institute of Technology (ETH), Pharmacogenomics Lab, Zurich, CH
Headed by: Prof. Michael Detmar
9-month research placement focused on extending Master Thesis results, and publishing work on the "*In vivo* visualization and quantification of collecting lymphatic vessel contractility using near-infrared imaging"

2013
Pharmaceutical Internship
F. Hoffmann-La Roche AG, Laboratory of Toxicology, Basel, CH
Headed by: Dr. Stefan Kustermann
2-month industrial research experience focused on establishing new experimental methods to measure oxidative stress *in vitro*

Education

2016 - Present
PhD Programme of Cancer and Immunology, CHUV / UNI, Lausanne, CH
Ludwig Institute for Cancer Research, Immunopeptidomics Lab

2013 - 2015
Master of Sciences Biology at ETH, Zurich, CH | Grade: 5.95
Major in Biochemistry with Distinction, Received Best in Class Award
Master Thesis | Pharmacogenomics Lab | Prof. Michael Detmar
“Characterization and Stimulation of Ocular Lymphatic Vessels”

2010 – 2013
Bachelor of Sciences in Biology at ETH, Zurich, CH | Grade: 5.74

ERASMUS Exchange Programme at Imperial College, London, UK
Third Bachelor Year | Major: Biochemistry

Bachelor Thesis | Department of Life Sciences, Glycobiology, Imperial | Dr. S. Haslam
“Glycomic Investigation by Mass Spectrometry Reveals Differences that Discriminate the Prognostic Subsets in Chronic Lymphocytic Leukemia”

Semester Thesis | Institute of Molecular Health Sciences, ETH | Prof. Sabine Werner
“Functional Characterization of Various FGF-Regulated Proteins in the Regenerating Liver”

Semester Thesis | Institute of Biochemistry, ETH | Prof. Matthias Peter
“Dissecting Cullin-RING E3 Ligase Regulation through Critical Interfaces of Nedd8”

2004 - 2010 **Swiss Matura at Realgymnasium Rämibühl, Zurich, CH**
Foreign language major in Spanish

Honors and Awards

2019 **Selected Speaker | Cancer Immunotherapy Conference (CICON), Paris, FR**
Category: Tumor Antigens, chosen to present PhD research at an international conference with a scientific audience of over 1,000 participants

2018 **Best Poster Award | Cancer Immunotherapy Conference (CIMT), Mainz, DE**
Category: New Targets and New Leads

2017 **Selected Speaker | Faculty and Staff Retreat of the Swiss Cancer Centre, Lausanne, CH**
Chosen to present PhD research at an internal conference with a scientific audience of over 250 participants

2016 **Willi Studer Prize | ETH, Zurich, CH**
Awarded for achieving highest grade among peers in Master’s Degree Programme Biology ETH

2013 **Best in Class, Glycobiology | Imperial College London, UK**
Awarded for achieving highest grade in Glycobiology course

Scientific Publications

2020 **Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes**
Chong C.*, Müller M., Pak H., Harnett D., Huber F., Grun D., Leleu M., Auger A., Arnaud M., Stevenson B., Michaux J., Bilic I., Hirsekorn A., Calviello L., Simó-Riudalbas L., Planet E., Lubiński J., Bryśkiewicz M., Wiznerowicz M., Xenarios I., Zhang L., Trono D., Harari A., Ohler U., Coukos G., Bassani-Sternberg, M. *Nature Communications, accepted on 12th February 2020*

2019 **High-throughput, fast, and sensitive immunopeptidomics sample processing for mass spectrometry**
Marino F*., **Chong C.*,** Michaux J., Bassani-Sternberg M. *In Immune Checkpoint Blockade (pp. 67-79). Humana Press, New York*

2019 **Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes**
Racle J*., Michaux J., Rockinger G., Arnaud M., Bobisse S., **Chong C.**, Guillaume P., Coukos G., Harari A., Jandus C., Bassani-Sternberg M., Gfeller D. *Nature biotechnology*

2019 **Tumour-reactive T cell subsets in the microenvironment of ovarian cancer**
Westergaard M.*, Andersen R., **Chong C.**, Kjeldsen J., Pedersen M., Friese C., Hasselager T., Lajer H., Coukos G., Bassani-Sternberg M., Donia M., Marie Svane I. *British journal of cancer*

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

2018 High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome
Chong C.*, Marino F.*, Pak H., Racle J., Daniel R., Müller M., Gfeller D., Coukos G., Bassani-Sternberg M. *Molecular & Cellular Proteomics*

2017 Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity
Bassani-Sternberg M.*, **Chong C.**, Guillaume P., Solleder M., Pak H., Gannon P., Kandalaft L., Coukos G., Gfeller D. *PLoS computational biology*

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

2016 *In vivo* visualization and quantification of collecting lymphatic vessel contractility using near-infrared imaging
Chong C.*, Scholkmann F.*, Bachmann S., Luciani P., Leroux J., Detmar M., Proulx S. *Scientific reports*

2016 DeepCAGE transcriptomics identify HOXD10 as a transcription factor regulating lymphatic endothelial responses to VEGF-C.
Klein S.*, Dieterich L.*, Mathelier A., **Chong C.**, Sliwa-Primorac A., Hong Y., Shin J., Lizio M., Itoh M., Kawaji H., Lassmann T., Daub C., Arner E., Fantom Consortium, Carninci P., Hayashizaki Y., Forrest A., Wasserman W., Detmar M. *Journal of cell science*

2016 Regulation of lymphangiogenesis in the diaphragm by macrophages and VEGFR-3 signaling.
Ochsenbein A.*, Karaman S., Proulx S., Goldmann R., Chittazhathu J., Dasargyri A., **Chong C.**, Leroux J., Stanley E., Detmar M. *Angiogenesis*

TBD Immunogenicity of BRCA1-deficient ovarian cancers is driven through DNA sensing and is augmented by PARP inhibition
Bruand M.*, Barras D., Mina M., Ghisoni E., Lanitis E., Zhang H, **Chong C.**, Chee S., Tsianou T., Dorier J., Stevenson B., Iseli C., Ronet C., Bobisse S., Genolet R., Walton J., Bassani-Sternberg M., Kandalaft L., Ren B., McNeish I., Swisher E., Harari A., Delorenzi M., Ciriello G., Irving M., Rusakiewicz S., Foukas P., Martinon F., Dangaj D., Coukos G. *Manuscript in revision*

TBD A pre-clinical humanized mouse model of antigenic epitope identification using mass spectrometry-based immunopeptidomics
Semilietof A.*, Stefanidis E., **Chong C.**, Pak H., Bobisse S., Marino F., Varrin M., Girotra M., Rotmistrovsky-Valcarcel N., Mathevet P., Barnier-Quer C., Sandaltzopoulos R., Collin N., Harari A., Coukos G., Bassani-Sternberg M., Vanhecke D. *Manuscript in preparation*

Relevant Scientific Presentations

2019	**Poster** at 8th SCCL Faculty & Staff Retreat, Lausanne, CH
	Poster and Talk at Cancer Immunotherapy Conference (CICON19), Paris, FR
	Poster at 6th joint Novartis-EPFL-UNIL meeting, Basel, CH
2018	**Poster** at CIMT, Mainz, DE
	Talk at Wolfsberg Meeting, Meeting of the Swiss Immunology PhD students, Thun, CH
2017	**Poster and Talk** at 6th SCCL Faculty & Staff Retreat, Lausanne, CH
	Poster at EMBO workshop on Antigen Processing and Presentation, Salamanca, ES
2016	**Poster** at 5th SCCL Faculty & Staff Retreat, Lausanne, CH
	Poster at MaxQuant Summer School, Oxford, UK

REFERENCES

1. Galon, J. and D. Bruni, *Tumor Immunology and Tumor Evolution: Intertwined Histories.* Immunity, 2020. **52**(1): p. 55-81.
2. Parkin, J. and B. Cohen, *An overview of the immune system.* Lancet, 2001. **357**(9270): p. 1777-89.
3. Vesely, M.D., et al., *Natural innate and adaptive immunity to cancer.* Annu Rev Immunol, 2011. **29**: p. 235-71.
4. Chen, D.S. and I. Mellman, *Oncology Meets Immunology: The Cancer-Immunity Cycle.* Immunity, 2013. **39**(1): p. 1-10.
5. O'Donnell, J.S., M.W.L. Teng, and M.J. Smyth, *Cancer immunoediting and resistance to T cell-based immunotherapy.* Nat Rev Clin Oncol, 2019. **16**(3): p. 151-167.
6. Binnewies, M., et al., *Understanding the tumor immune microenvironment (TIME) for effective therapy.* Nat Med, 2018. **24**(5): p. 541-550.
7. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
8. Egen, J.G., W. Ouyang, and L.C. Wu, *Human Anti-tumor Immunity: Insights from Immunotherapy Clinical Trials.* Immunity, 2020. **52**(1): p. 36-54.
9. Kruger, S., et al., *Advances in cancer immunotherapy 2019 - latest trends.* J Exp Clin Cancer Res, 2019. **38**(1): p. 268.
10. Christofi, T., et al., *Current Perspectives in Cancer Immunotherapy.* Cancers (Basel), 2019. **11**(10).
11. Couzin-Frankel, J., *Breakthrough of the year 2013. Cancer immunotherapy.* Science, 2013. **342**(6165): p. 1432-3.
12. Chamoto, K., R. Hatae, and T. Honjo, *Current issues and perspectives in PD-1 blockade cancer immunotherapy.* Int J Clin Oncol, 2020.
13. Brower, V., *Checkpoint blockade immunotherapy for cancer comes of age.* J Natl Cancer Inst, 2015. **107**(3).
14. Gubin, M.M., et al., *Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens.* Nature, 2014. **515**(7528): p. 577-81.
15. Snyder, A., et al., *Genetic basis for clinical response to CTLA-4 blockade in melanoma.* N Engl J Med, 2014. **371**(23): p. 2189-2199.
16. Robert, C., et al., *Pembrolizumab versus Ipilimumab in Advanced Melanoma.* N Engl J Med, 2015. **372**(26): p. 2521-32.
17. Guedan, S., M. Ruella, and C.H. June, *Emerging Cellular Therapies for Cancer.* Annu Rev Immunol, 2019. **37**: p. 145-171.
18. Magalhaes, I., et al., *Facing the future: challenges and opportunities in adoptive T cell therapy in cancer.* Expert opinion on biological therapy, 2019: p. 1-17.
19. June, C.H., S.R. Riddell, and T.N. Schumacher, *Adoptive cellular therapy: a race to the finish line.* Sci Transl Med, 2015. **7**(280): p. 280ps7.
20. Martinez, M. and E.K. Moon, *CAR T Cells for Solid Tumors: New Strategies for Finding, Infiltrating, and Surviving in the Tumor Microenvironment.* Front Immunol, 2019. **10**: p. 128.
21. Zhang, J. and L. Wang, *The Emerging World of TCR-T Cell Trials Against Cancer: A Systematic Review.* Technol Cancer Res Treat, 2019. **18**: p. 1533033819831068.
22. Garber, K., *Driving T-cell immunotherapy to solid tumors.* Nat Biotechnol, 2018. **36**(3): p. 215-219.
23. Bagley, S.J. and D.M. O'Rourke, *Clinical investigation of CAR T cells for solid tumors: Lessons learned and future directions.* Pharmacol Ther, 2020. **205**: p. 107419.
24. Giordano-Attianese, G., et al., *A computationally designed chimeric antigen receptor provides a small-molecule safety switch for T-cell therapy.* Nat Biotechnol, 2020.

25. Mata, M., et al., *Inducible Activation of MyD88 and CD40 in CAR T Cells Results in Controllable and Potent Antitumor Activity in Preclinical Solid Tumor Models.* Cancer Discov, 2017. **7**(11): p. 1306-1319.

26. Hu, B., et al., *Augmentation of Antitumor Immunity by Human and Mouse CAR T Cells Secreting IL-18.* Cell Rep, 2017. **20**(13): p. 3025-3033.

27. Zhang, L., et al., *Tumor-infiltrating lymphocytes genetically engineered with an inducible gene encoding interleukin-12 for the immunotherapy of metastatic melanoma.* Clin Cancer Res, 2015. **21**(10): p. 2278-88.

28. Krenciute, G., et al., *Transgenic Expression of IL15 Improves Antiglioma Activity of IL13Ralpha2-CAR T Cells but Results in Antigen Loss Variants.* Cancer Immunol Res, 2017. **5**(7): p. 571-581.

29. Spolski, R. and W.J. Leonard, *Interleukin-21: a double-edged sword with therapeutic potential.* Nat Rev Drug Discov, 2014. **13**(5): p. 379-95.

30. Ren, J., et al., *Multiplex Genome Editing to Generate Universal CAR T Cells Resistant to PD1 Inhibition.* Clin Cancer Res, 2017. **23**(9): p. 2255-2266.

31. Tardón, M.C., et al., *Peptides as cancer vaccines.* Current opinion in pharmacology, 2019. **47**: p. 20-26.

32. Ventola, C.L., *Cancer Immunotherapy, Part 1: Current Strategies and Agents.* P T, 2017. **42**(6): p. 375-383.

33. Fennemann, F.L., et al., *Attacking Tumors From All Sides: Personalized Multiplex Vaccines to Tackle Intratumor Heterogeneity.* Frontiers in Immunology, 2019. **10**: p. 824.

34. Chiang, C.L.L., G. Coukos, and L.E. Kandalaft, *Whole Tumor Antigen Vaccines: Where Are We?* Vaccines, 2015. **3**(2): p. 344-372.

35. Homicsko, K., et al., *Combine and Conquer: Double CTLA-4 and PD-1 Blockade Combined with Whole Tumor Antigen Vaccine Cooperate to Eradicate Tumors.* Cancer Research, 2016. **76**(23): p. 6765-6767.

36. Duraiswamy, J., G.J. Freeman, and G. Coukos, *Dual blockade of PD-1 and CTLA-4 combined with tumor vaccine effectively restores T-cell rejection function in tumors--response.* Cancer Res, 2014. **74**(2): p. 633-4; discussion 635.

37. Mougel, A., M. Terme, and C. Tanchot, *Therapeutic Cancer Vaccine and Combinations With Antiangiogenic Therapies and Immune Checkpoint Blockade.* Front Immunol, 2019. **10**: p. 467.

38. Seliger, B., *Combinatorial Approaches With Checkpoint Inhibitors to Enhance Anti-tumor Immunity.* Front Immunol, 2019. **10**: p. 999.

39. Marshall, H.T. and M.B.A. Djamgoz, *Immuno-Oncology: Emerging Targets and Combination Therapies.* Front Oncol, 2018. **8**: p. 315.

40. Jones, P.A., et al., *Epigenetic therapy in immune-oncology.* Nat Rev Cancer, 2019. **19**(3): p. 151-161.

41. Covre, A., et al., *Antitumor activity of epigenetic immunomodulation combined with CTLA-4 blockade in syngeneic mouse models.* Oncoimmunology, 2015. **4**(8): p. e1019978.

42. Dunn, J. and S. Rao, *Epigenetics and immunotherapy: The current state of play.* Mol Immunol, 2017. **87**: p. 227-239.

43. Li, X., et al., *Decitabine: a promising epi-immunotherapeutic agent in solid tumors.* Expert Review of Clinical Immunology, 2015. **11**(3): p. 363-375.

44. Zhang, Z., et al., *Decitabine treatment sensitizes tumor cells to T-cell-mediated cytotoxicity in patients with myelodysplastic syndromes.* American Journal of Translational Research, 2017. **9**(2): p. 454-465.

45. Brocks, D., et al., *DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats (vol 49, pg 1052, 2017).* Nature Genetics, 2017. **49**(11): p. 1661-1661.

46. Chiappinelli, K.B., et al., *Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses.* Cell, 2015. **162**(5): p. 974-86.

47. Roulois, D., et al., *DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts.* Cell, 2015. **162**(5): p. 961-973.

48. Ventola, C.L., *Cancer Immunotherapy, Part 3: Challenges and Future Trends.* P T, 2017. **42**(8): p. 514-521.

49. Haslam, A. and V. Prasad, *Estimation of the Percentage of US Patients With Cancer Who Are Eligible for and Respond to Checkpoint Inhibitor Immunotherapy Drugs.* JAMA Netw Open, 2019. **2**(5): p. e192535.

50. Duffy, M.J. and J. Crown, *Biomarkers for Predicting Response to Immunotherapy with Immune Checkpoint Inhibitors in Cancer Patients.* Clin Chem, 2019. **65**(10): p. 1228-1238.

51. Galuppini, F., et al., *Tumor mutation burden: from comprehensive mutational screening to the clinic.* Cancer Cell Int, 2019. **19**: p. 209.

52. Rizvi, N.A., et al., *Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer.* Science, 2015. **348**(6230): p. 124-8.

53. Schumacher, T.N., C. Kesmir, and M.M. van Buuren, *Biomarkers in cancer immunotherapy.* Cancer Cell, 2015. **27**(1): p. 12-4.

54. Hollingsworth, R.E. and K. Jansen, *Turning the corner on therapeutic cancer vaccines.* NPJ Vaccines, 2019. **4**: p. 7.

55. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade.* Science, 2016.

56. McGranahan, N. and C. Swanton, *Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.* Cell, 2017. **168**(4): p. 613-628.

57. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nat Rev Immunol, 2011. **11**(12): p. 823-36.

58. Rock, K.L., E. Reits, and J. Neefjes, *Present Yourself! By MHC Class I and MHC Class II Molecules.* Trends Immunol, 2016. **37**(11): p. 724-737.

59. van Deutekom, H.W.M. and C. Kesmir, *Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most?* Immunogenetics, 2015. **67**(8): p. 425-436.

60. Paul, S., et al., *HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity.* J Immunol, 2013. **191**(12): p. 5831-9.

61. Falk, K., et al., *Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. 1991.* J Immunol, 2006. **177**(5): p. 2741-7.

62. McClelland, E.E., D.J. Penn, and W.K. Potts, *Major histocompatibility complex heterozygote superiority during coinfection.* Infection and Immunity, 2003. **71**(4): p. 2079-2086.

63. Penn, D.J., K. Damjanovich, and W.K. Potts, *MHC heterozygosity confers a selective advantage against multiple-strain infections.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(17): p. 11260-11264.

64. Chowell, D., et al., *Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy.* Science, 2018. **359**(6375): p. 582-587.

65. McGranahan, N., et al., *Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution.* Cell, 2017. **171**(6): p. 1259-1271 e11.

66. Oppermann, U. and L. Moutsianas, *Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility (vol 43, pg 761, 2011).* Nature Genetics, 2011. **43**(9): p. 919-919.

67. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biology, 2017. **18**.

68. Lima-Junior Jda, C. and L.R. Pratt-Riccio, *Major Histocompatibility Complex and Malaria: Focus on Plasmodium vivax Infection.* Front Immunol, 2016. **7**: p. 13.

69. Kloetzel, P.M., *Antigen processing by the proteasome.* Nature Reviews Molecular Cell Biology, 2001. **2**(3): p. 179-187.

70. Kisselev, A.F., et al., *The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes - Implications for understanding the degradative mechanism and antigen presentation.* Journal of Biological Chemistry, 1999. **274**(6): p. 3363-3371.

71. Aki, M., et al., *Interferon-Gamma Induces Different Subunit Organizations and Functional Diversity of Proteasomes.* Journal of Biochemistry, 1994. **115**(2): p. 257-269.

72. Murata, S., et al., *The immunoproteasome and thymoproteasome: functions, evolution and human disease.* Nat Immunol, 2018. **19**(9): p. 923-931.

73. Mishto, M., et al., *Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation.* Eur J Immunol, 2014. **44**(12): p. 3508-21.

74. Vigneron, N. and B.J. Van den Eynde, *Proteasome subtypes and the processing of tumor antigens: increasing antigenic diversity.* Current Opinion in Immunology, 2012. **24**(1): p. 84-91.

75. de Verteuil, D., et al., *Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules.* Mol Cell Proteomics, 2010. **9**(9): p. 2034-47.

76. Kincaid, E.Z., et al., *Mice completely lacking immunoproteasomes show major changes in antigen presentation.* Nature Immunology, 2012. **13**(2): p. 129-135.

77. Yewdell, J.W., E. Reits, and J. Neefjes, *Making sense of mass destruction: quantitating MHC class I antigen presentation.* Nat Rev Immunol, 2003. **3**(12): p. 952-61.

78. Croft, N.P., et al., *Kinetics of antigen expression and epitope presentation during virus infection.* PLoS Pathog, 2013. **9**(1): p. e1003129.

79. Schubert, U., et al., *Rapid degradation of a large fraction of newly synthesized proteins by proteasomes.* Nature, 2000. **404**(6779): p. 770-4.

80. Granados, D.P., et al., *MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements.* Blood, 2012. **119**(26): p. e181-91.

81. Anton, L.C. and J.W. Yewdell, *Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors.* J Leukoc Biol, 2014. **95**(4): p. 551-62.

82. Apcher, S., et al., *Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(28): p. 11572-11577.

83. Qian, S.B., et al., *Characterization of rapidly degraded polypeptides in mammalian cells reveals a novel layer of nascent protein quality control.* J Biol Chem, 2006. **281**(1): p. 392-400.

84. Roche, P.A. and K. Furuta, *The ins and outs of MHC class II-mediated antigen processing and presentation.* Nat Rev Immunol, 2015. **15**(4): p. 203-16.

85. Embgenbroich, M. and S. Burgdorf, *Current Concepts of Antigen Cross-Presentation.* Frontiers in Immunology, 2018. **9**.

86. Cruz, F.M., et al., *The Biology and Underlying Mechanisms of Cross-Presentation of Exogenous Antigens on MHC-I Molecules.* Annu Rev Immunol, 2017.

87. Ackerman, A.L. and P. Cresswell, *Cellular mechanisms governing cross-presentation of exogenous antigens.* Nat Immunol, 2004. **5**(7): p. 678-84.

88. Schuette, V. and S. Burgdorf, *The ins-and-outs of endosomal antigens for cross-presentation.* Curr Opin Immunol, 2014. **26**: p. 63-8.

89. Guermonprez, P., et al., *ER-phagosome fusion defines an MHC class I cross-presentation compartment in dendritic cells.* Nature, 2003. **425**(6956): p. 397-402.

90. Gros, M. and S. Amigorena, *Regulation of Antigen Export to the Cytosol During Cross-Presentation.* Front Immunol, 2019. **10**: p. 41.
91. Reeves, E. and E. James, *Antigen processing and immune regulation in the response to tumours.* Immunology, 2017. **150**(1): p. 16-24.
92. Leone, P., et al., *MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells.* J Natl Cancer Inst, 2013. **105**(16): p. 1172-87.
93. Meissner, M., et al., *Defects in the human leukocyte antigen class I antigen processing machinery in head and neck squamous cell carcinoma: association with clinical outcome.* Clin Cancer Res, 2005. **11**(7): p. 2552-60.
94. Ogino, T., et al., *HLA class I antigen down-regulation in primary laryngeal squamous cell carcinoma lesions as a poor prognostic marker.* Cancer Research, 2006. **66**(18): p. 9281-9289.
95. Rolland, P., et al., *Human leukocyte antigen class I antigen expression is an independent prognostic factor in ovarian cancer.* Clinical Cancer Research, 2007. **13**(12): p. 3591-3596.
96. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.
97. Coulie, P.G., et al., *Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy.* Nature Reviews Cancer, 2014. **14**(2): p. 135-146.
98. Ilyas, S. and J.C. Yang, *Landscape of Tumor Antigens in T Cell Immunotherapy.* Journal of Immunology, 2015. **195**(11): p. 5117-5122.
99. Schmidt, M. and J.R. Lill, *MHC class I presented antigens from malignancies: A perspective on analytical characterization & immunogenicity.* J Proteomics, 2019. **191**: p. 48-57.
100. Smith, C.C., et al., *Alternative tumour-specific antigens.* Nat Rev Cancer, 2019. **19**(8): p. 465-478.
101. Schumacher, T.N., W. Scheper, and P. Kvistborg, *Cancer Neoantigens.* Annual Review of Immunology, Vol 37, 2019, 2019. **37**: p. 173-200.
102. Skipper, J.C., et al., *Mass-spectrometric evaluation of HLA-A*0201-associated peptides identifies dominant naturally processed forms of CTL epitopes from MART-1 and gp100.* Int J Cancer, 1999. **82**(5): p. 669-77.
103. Hogquist, K.A., T.A. Baldwin, and S.C. Jameson, *Central tolerance: Learning self-control in the thymus.* Nature Reviews Immunology, 2005. **5**(10): p. 772-782.
104. Cobbold, M., et al., *MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia.* Sci Transl Med, 2013. **5**(203): p. 203ra125.
105. Lamers, C.H., et al., *Treatment of metastatic renal cell carcinoma with autologous T-lymphocytes genetically retargeted against carbonic anhydrase IX: first clinical experience.* J Clin Oncol, 2006. **24**(13): p. e20-2.
106. Schietinger, A., M. Philip, and H. Schreiber, *Specificity in cancer immunotherapy.* Seminars in Immunology, 2008. **20**(5): p. 276-285.
107. Vigneron, N., *Human Tumor Antigens and Cancer Immunotherapy.* Biomed Res Int, 2015. **2015**: p. 948501.
108. Tio, D., et al., *Expression of cancer/testis antigens in cutaneous melanoma: a systematic review.* Melanoma Res, 2019. **29**(4): p. 349-357.
109. Caballero, O.L. and Y.T. Chen, *Cancer/testis (CT) antigens: potential targets for immunotherapy.* Cancer Sci, 2009. **100**(11): p. 2014-21.
110. van der Bruggen, P., et al., *A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma.* Science, 1991. **254**(5038): p. 1643-7.
111. Fratta, E., et al., *The biology of cancer testis antigens: putative function, regulation and therapeutic potential.* Mol Oncol, 2011. **5**(2): p. 164-82.
112. Connerotte, T., et al., *Functions of Anti-MAGE T-cells induced in melanoma patients under different vaccination modalities.* Cancer Res, 2008. **68**(10): p. 3931-40.

113. Krishnadas, D.K., et al., *A phase I trial combining decitabine/dendritic cell vaccine targeting MAGE-A1, MAGE-A3 and NY-ESO-1 for children with relapsed or therapy-refractory neuroblastoma and sarcoma.* Cancer Immunol Immunother, 2015. **64**(10): p. 1251-60.
114. Zajac, P., et al., *MAGE-A Antigens and Cancer Immunotherapy.* Front Med (Lausanne), 2017. **4**: p. 18.
115. Engelhard, V.H., et al., *Post-translational modifications of naturally processed MHC-binding epitopes.* Curr Opin Immunol, 2006. **18**(1): p. 92-7.
116. Thygesen, C., et al., *Characterizing disease-associated changes in post-translational modifications by mass spectrometry.* Expert Rev Proteomics, 2018: p. 1-14.
117. Marino, F., et al., *Extended O-GlcNAc on HLA Class-I-Bound Peptides.* J Am Chem Soc, 2015. **137**(34): p. 10922-5.
118. Zarling, A.L., et al., *Phosphorylated peptides are naturally processed and presented by major histocompatibility complex class I molecules in vivo.* Journal of Experimental Medicine, 2000. **192**(12): p. 1755-1762.
119. Zarling, A.L., et al., *Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy.* Proc Natl Acad Sci U S A, 2006. **103**(40): p. 14889-94.
120. Malaker, S.A., et al., *Identification of glycopeptides as posttranslationally modified neoantigens in leukemia.* Cancer immunology research, 2017. **5**(5): p. 376-384.
121. Yang, A., et al., *Current state in the development of candidate therapeutic HPV vaccines.* Expert Rev Vaccines, 2016. **15**(8): p. 989-1007.
122. Draper, L.M., et al., *Targeting of HPV-16+ Epithelial Cancer Cells by TCR Gene Engineered T Cells Directed against E6.* Clin Cancer Res, 2015. **21**(19): p. 4431-9.
123. Mesri, E.A., M.A. Feitelson, and K. Munger, *Human viral oncogenesis: a cancer hallmarks analysis.* Cell Host Microbe, 2014. **15**(3): p. 266-82.
124. Coulie, P.G., et al., *A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma.* Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7976-80.
125. Robbins, P.F., et al., *A mutated beta-catenin gene encodes a melanoma-specific antigen recognized by tumor infiltrating lymphocytes.* J Exp Med, 1996. **183**(3): p. 1185-92.
126. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
127. Richters, M.M., et al., *Best practices for bioinformatic characterization of neoantigens for clinical utility.* Genome Med, 2019. **11**(1): p. 56.
128. Matsushita, H., et al., *Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting.* Nature, 2012. **482**(7385): p. 400-4.
129. Castle, J.C., et al., *Exploiting the mutanome for tumor vaccination.* Cancer Res, 2012. **72**(5): p. 1081-91.
130. Robbins, P.F., et al., *Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells.* Nat Med, 2013. **19**(6): p. 747-52.
131. van Rooij, N., et al., *Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma.* J Clin Oncol, 2013. **31**(32): p. e439-42.
132. Stevanovic, S., et al., *Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer.* Science, 2017. **356**(6334): p. 200-205.
133. Snyder, A., J.D. Wolchok, and T.A. Chan, *Genetic basis for clinical response to CTLA-4 blockade.* N Engl J Med, 2015. **372**(8): p. 783.
134. Lauss, M., et al., *Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma.* Nature Communications, 2017. **8**.

135. Yarchoan, M., A. Hopkins, and E.M. Jaffee, *Tumor Mutational Burden and Response Rate to PD-1 Inhibition.* New England Journal of Medicine, 2017. **377**(25): p. 2500-2501.

136. Blank, C.U., et al., *The "cancer immunogram".* Science, 2016. **352**(6286): p. 658-660.

137. Linnebacher, M., et al., *Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens.* Int J Cancer, 2001. **93**(1): p. 6-11.

138. Saulquin, X., et al., *+1 frameshifting as a novel mechanism to generate a cryptic cytotoxic T lymphocyte epitope derived from human interleukin 10.* Journal of Experimental Medicine, 2002. **195**(3): p. 353-358.

139. Zook, M.B., et al., *Epitopes derived by incidental translational frameshifting give rise to a protective CTL response.* Journal of Immunology, 2006. **176**(11): p. 6928-6934.

140. Turajlic, S., et al., *Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis.* Lancet Oncology, 2017. **18**(8): p. 1009-1021.

141. Yang, W., et al., *Immunogenic neoantigens derived from gene fusions stimulate T cell responses.* Nature medicine, 2019. **25**(5): p. 767.

142. Saeterdal, I., et al., *A TGF beta RII frameshift-mutation-derived CTL epitope recognised by HLA-A2-restricted CD8(+) T cells.* Cancer Immunology Immunotherapy, 2001. **50**(9): p. 469-476.

143. Pinilla-Ibarz, J., et al., *Vaccination of patients with chronic myelogenous leukemia with bcr-abl oncogene breakpoint fusion peptides generates specific immune responses.* Blood, 2000. **95**(5): p. 1781-7.

144. Castle, J.C., et al., *Mutation-Derived Neoantigens for Cancer Immunotherapy.* Front Immunol, 2019. **10**: p. 1856.

145. Laumont, C.M. and C. Perreault, *Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy.* Cellular and Molecular Life Sciences, 2018. **75**(4): p. 607-621.

146. Boon, T. and A. Vanpel, *T-Cell-Recognized Antigenic Peptides Derived from the Cellular Genome Are Not Protein-Degradation Products but Can Be Generated Directly by Transcription and Translation of Short Subgenic Regions - a Hypothesis.* Immunogenetics, 1989. **29**(2): p. 75-79.

147. Malarkannan, S., M. Afkarian, and N. Shastri, *A rare cryptic translation product is presented by Kb major histocompatibility complex class I molecule to alloreactive T cells.* J Exp Med, 1995. **182**(6): p. 1739-50.

148. Guilloux, Y., et al., *A peptide recognized by human cytolytic T lymphocytes on HLA-A2 melanomas is encoded by an intron sequence of the N-acetylglucosaminyltransferase V gene.* J Exp Med, 1996. **183**(3): p. 1173-83.

149. Wang, R.F., et al., *Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen.* J Exp Med, 1996. **183**(3): p. 1131-40.

150. Robbins, P.F., et al., *The intronic region of an incompletely spliced gp100 gene transcript encodes an epitope recognized by melanoma-reactive tumor-infiltrating lymphocytes.* J Immunol, 1997. **159**(1): p. 303-8.

151. Wang, R.F., et al., *A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames.* J Immunol, 1998. **161**(7): p. 3598-606.

152. Van Den Eynde, B.J., et al., *A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription.* J Exp Med, 1999. **190**(12): p. 1793-800.

153. Weinzierl, A.O., et al., *A cryptic vascular endothelial growth factor T-cell epitope: Identification and characterization by mass spectrometry and T-cell assays.* Cancer Research, 2008. **68**(7): p. 2447-2454.

154. Vigneron, N., et al., *An antigenic peptide produced by peptide splicing in the proteasome.* Science, 2004. **304**(5670): p. 587-90.

155. Hanada, K., J.W. Yewdell, and J.C. Yang, *Immune recognition of a human renal cancer antigen through post-translational protein splicing.* Nature, 2004. **427**(6971): p. 252-6.

156. Charpentier, M., et al., *IRES-dependent translation of the long non coding RNA meloe in melanoma cells produces the most immunogenic MELOE antigens.* Oncotarget, 2016. **7**(37): p. 59704-59713.

157. Lupetti, R., et al., *Translation of a retained intron in tyrosinase-related protein (TRP) 2 mRNA generates a new cytotoxic T lymphocyte (CTL)-defined and shared human melanoma antigen not expressed in normal cells of the melanocytic lineage.* J Exp Med, 1998. **188**(6): p. 1005-16.

158. Moreau-Aubry, A., et al., *A processed pseudogene codes for a new antigen recognized by a CD8(+) T cell clone on melanoma.* Journal of Experimental Medicine, 2000. **191**(9): p. 1617-1623.

159. Siehl, J.M., et al., *Expression of Wilms' tumor gene 1 at different stages of acute myeloid leukemia and analysis of its major splice variants.* Ann Hematol, 2004. **83**(12): p. 745-50.

160. Kahles, A., et al., *Comprehensive analysis of alternative splicing across tumors from 8,705 patients.* Cancer cell, 2018. **34**(2): p. 211-224. e6.

161. Li, L.J., et al., *Translation of noncoding RNAs: Focus on lncRNAs, pri-miRNAs, and circRNAs.* Exp Cell Res, 2017.

162. Godet, Y., et al., *Frequent occurrence of high affinity T cells against MELOE-1 makes this antigen an attractive target for melanoma immunotherapy.* European Journal of Immunology, 2010. **40**(6): p. 1786-1794.

163. Carbonnelle, D., et al., *The Melanoma Antigens MELOE-1 and MELOE-2 Are Translated from a Bona Fide Polycistronic mRNA Containing Functional IRES Sequences.* Plos One, 2013. **8**(9).

164. Friedli, M. and D. Trono, *The developmental control of transposable elements and the evolution of higher species.* Annu Rev Cell Dev Biol, 2015. **31**: p. 429-51.

165. Wang-Johanning, F., et al., *Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer.* Int J Cancer, 2007. **120**(1): p. 81-90.

166. Buscher, K., et al., *Expression of human endogenous retrovirus K in melanomas and melanoma cell lines.* Cancer Res, 2005. **65**(10): p. 4172-80.

167. Wang-Johanning, F., et al., *Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma.* Cancer, 2003. **98**(1): p. 187-197.

168. Cherkasova, E., et al., *Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer.* Cancer Research, 2016. **76**(8): p. 2177-2185.

169. Smith, C.C., et al., *Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma.* The Journal of clinical investigation, 2019. **128**(11): p. 4804-4820.

170. Bobisse, S., et al., *Neoantigen-based cancer immunotherapy.* Ann Transl Med, 2016. **4**(14): p. 262.

171. Vita, R., et al., *The immune epitope database (IEDB) 3.0.* Nucleic Acids Res, 2015. **43**(Database issue): p. D405-12.

172. Nielsen, M., et al., *The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage.* Immunogenetics, 2005. **57**(1-2): p. 33-41.

173. Peters, B., et al., *Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors.* The Journal of Immunology, 2003. **171**(4): p. 1741-1749.

174. Bassani-Sternberg, M. and D. Gfeller, *Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions.* J Immunol, 2016. **197**(6): p. 2492-9.

175. Bassani-Sternberg, M., et al., *Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity.* PLoS Comput Biol, 2017. **13**(8): p. e1005725.

176. Jurtz, V., et al., *NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data.* The Journal of Immunology, 2017. **199**(9): p. 3360-3368.

177. Abelin, J.G., et al., *Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction.* Immunity, 2017. **46**(2): p. 315-326.

178. Racle, J., et al., *Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes.* Nature biotechnology, 2019: p. 1-4.

179. Abelin, J.G., et al., *Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction.* Immunity, 2019. **51**(4): p. 766-779. e17.

180. Chen, B.B., et al., *Predicting HLA class II antigen presentation through integrated deep learning.* Nature Biotechnology, 2019. **37**(11): p. 1332-+.

181. Hintzsche, J.D., W.A. Robinson, and A.C. Tan, *A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data.* Int J Genomics, 2016. **2016**: p. 7983236.

182. Gfeller, D., et al., *Current tools for predicting cancer-specific T cell immunity.* Oncoimmunology, 2016. **5**(7): p. e1177691.

183. Mei, S., et al., *A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction.* Briefings in bioinformatics, 2019.

184. Schmidt, J., et al., *In silico and cell-based analyses reveal strong divergence between prediction and observation of T-cell-recognized tumor antigen T-cell epitopes.* Journal of Biological Chemistry, 2017. **292**(28): p. 11840-11849.

185. Garcia-Garijo, A., C.A. Fajardo, and A. Gros, *Determinants for Neoantigen Identification.* Front Immunol, 2019. **10**: p. 1392.

186. Miyahira, Y., et al., *Quantification of Antigen-Specific Cd8(+) T-Cells Using an Elispot Assay.* Journal of Immunological Methods, 1995. **181**(1): p. 45-54.

187. Siota, M., et al., *ELISpot for measuring human immune responses to vaccines.* Expert Review of Vaccines, 2011. **10**(3): p. 299-306.

188. Chevalier, M.F., et al., *High-throughput monitoring of human tumor-specific T-cell responses with large peptide pools.* Oncoimmunology, 2015. **4**(10).

189. Kalaora, S., et al., *Combined Analysis of Antigen Presentation and T-cell Recognition Reveals Restricted Immune Responses in Melanoma.* Cancer Discov, 2018. **8**(11): p. 1366-1375.

190. Kalaora, S., et al., *Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens.* Oncotarget, 2016.

191. Wick, D.A., et al., *Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer.* Clin Cancer Res, 2014. **20**(5): p. 1125-34.

192. Danilova, L., et al., *The Mutation-Associated Neoantigen Functional Expansion of Specific T Cells (MANAFEST) Assay: A Sensitive Platform for Monitoring Antitumor Immunity.* Cancer Immunol Res, 2018. **6**(8): p. 888-899.

193. Schmidt, J., et al., *Analysis, Isolation, and activation of antigen-specific CD4(+) and CD8(+) T cells by soluble MHC-peptide complexes.* Frontiers in Immunology, 2013. **4**.

194. Bentzen, A.K. and S.R. Hadrup, *Evolution of MHC-based technologies used for detection of antigen-responsive T cells.* Cancer Immunol Immunother, 2017. **66**(5): p. 657-666.

195. Cohen, C.J., et al., *Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes.* J Clin Invest, 2015. **125**(10): p. 3981-91.

196. Bentzen, A.K., et al., *Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes.* Nat Biotechnol, 2016. **34**(10): p. 1037-1045.

197. Lu, Y.C., et al., *Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions.* Clin Cancer Res, 2014. **20**(13): p. 3401-10.

198. Tran, E., et al., *Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer.* Science, 2014. **344**(6184): p. 641-5.

199. Arnaud, M., et al., *Biotechnologies to tackle the challenge of neoantigen identification.* Curr Opin Biotechnol, 2020. **65**: p. 52-59.

200. Vizcaino, J.A., et al., *The Human Immunopeptidome Project: a roadmap to predict and treat immune diseases.* Mol Cell Proteomics, 2019.

201. Hunt, D.F., et al., *Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.* Science, 1992. **255**(5049): p. 1261-3.

202. Dubey, P., et al., *The immunodominant antigen of an ultraviolet-induced regressor tumor is generated by a somatic point mutation in the DEAD box helicase p68.* J Exp Med, 1997. **185**(4): p. 695-705.

203. Singh-Jasuja, H., N.P. Emmerich, and H.G. Rammensee, *The Tubingen approach: identification, selection, and validation of tumor-associated HLA peptides for cancer therapy.* Cancer Immunol Immunother, 2004. **53**(3): p. 187-95.

204. Weinschenk, T., et al., *Integrated functional genomics approach for the design of patient-individual antitumor vaccines.* Cancer Res, 2002. **62**(20): p. 5818-27.

205. Caron, E., et al., *Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry.* Mol Cell Proteomics, 2015. **14**(12): p. 3105-17.

206. Pritchard, A.L., et al., *Exploration of peptides bound to MHC class I molecules in melanoma.* Pigment Cell Melanoma Res, 2015. **28**(3): p. 281-94.

207. Jarmalavicius, S., Y. Welte, and P. Walden, *High immunogenicity of the human leukocyte antigen peptidomes of melanoma tumor cells.* J Biol Chem, 2012. **287**(40): p. 33401-11.

208. Dutoit, V., et al., *Exploiting the glioblastoma peptidome to discover novel tumour-associated antigens for immunotherapy.* Brain, 2012. **135**(Pt 4): p. 1042-54.

209. Bassani-Sternberg, M., et al., *Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.* Nat Commun, 2016. **7**: p. 13404.

210. Hilf, N., et al., *Actively personalized vaccination trial for newly diagnosed glioblastoma.* Nature, 2019. **565**(7738): p. 240-245.

211. Alpizar, A., et al., *A Molecular Basis for the Presentation of Phosphorylated Peptides by HLA-B Antigens.* Molecular & Cellular Proteomics, 2017. **16**(2): p. 181-193.

212. Solleder, M., et al., *Mass spectrometry based immunopeptidomics leads to robust predictions of phosphorylated HLA class I ligands.* Molecular & Cellular Proteomics, 2019.

213. Bassani-Sternberg, M. and G. Coukos, *Mass spectrometry-based antigen discovery for cancer immunotherapy.* Curr Opin Immunol, 2016. **41**: p. 9-17.

214. Zhang, X., et al., *Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy.* Biomed Pharmacother, 2019. **120**: p. 109542.

215. Caron, E., et al., *A Case for a Human Immuno-Peptidome Project Consortium.* Immunity, 2017. **47**(2): p. 203-208.

216. Purcell, A.W., S.H. Ramarathinam, and N. Ternette, *Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics.* Nat Protoc, 2019. **14**(6): p. 1687-1707.

217. Hassan, C., et al., *Accurate quantitation of MHC-bound peptides by application of isotopically labeled peptide MHC complexes.* J Proteomics, 2014. **109**: p. 240-4.

218. Demmers, L.C., A.J.R. Heck, and W. Wu, *Pre-fractionation Extends but also Creates a Bias in the Detectable HLA Class I Ligandome.* Journal of Proteome Research, 2019. **18**(4): p. 1634-1643.

219. Edman, P., *Method for Determination of the Amino Acid Sequence in Peptides.* Acta Chemica Scandinavica, 1950. **4**(2): p. 283-293.

220. Glish, G.L. and R.W. Vachet, *The basics of mass spectrometry in the twenty-first century.* Nature Reviews Drug Discovery, 2003. **2**(2): p. 140-150.

221. De Hoffmann, E. and V. Stroobant, *Mass Spectrometry Principles and Applications.* 2013.

222. Michalski, A., et al., *Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer.* Mol Cell Proteomics, 2011. **10**(9): p. M111 011015.

223. Smoluch, M., et al., *Mass spectrometry an applied approach.* 2019.

224. Gedela, S. and N.R. Medicherla, *Chromatographic techniques for the separation of peptides: Application to proteomics.* Chromatographia, 2007. **65**(9-10): p. 511-518.

225. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.

226. Paizs, B. and S. Suhai, *Fragmentation pathways of protonated peptides.* Mass Spectrometry Reviews, 2005. **24**(4): p. 508-548.

227. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.

228. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.

229. Eng, J.K., et al., *A face in the crowd: recognizing peptides through database search.* Mol Cell Proteomics, 2011. **10**(11): p. R111 009522.

230. UniProt Consortium, T., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2018. **46**(5): p. 2699.

231. Kapp, E.A., et al., *An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis.* Proteomics, 2005. **5**(13): p. 3475-3490.

232. Kapp, E. and F. Schutz, *Overview of tandem mass spectrometry (MS/MS) database search algorithms.* Curr Protoc Protein Sci, 2007. **Chapter 25**: p. Unit25 2.

233. Eng, J.K., T.A. Jahan, and M.R. Hoopmann, *Comet: an open-source MS/MS sequence database search tool.* Proteomics, 2013. **13**(1): p. 22-4.

234. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.

235. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

236. Zhang, J., et al., *PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.* Mol Cell Proteomics, 2012. **11**(4): p. M111 010587.

237. Shan, P. and H. Tran, *Integrating Database Search and De Novo Sequencing for Immunopeptidomics with DIA Approach.* J Biomol Tech, 2019. **30**(Suppl): p. S23.

238. Faridi, P., A.W. Purcell, and N.P. Croft, *In Immunopeptidomics We Need a Sniper Instead of a Shotgun.* Proteomics, 2018. **18**(12).

239. Granados, D.P., et al., *The nature of self for T cells—a systems-level perspective.* Current opinion in immunology, 2015. **34**: p. 1-8.

240. Mester, G., V. Hoffmann, and S. Stevanovic, *Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands.* Cell Mol Life Sci, 2011. **68**(9): p. 1521-32.

241. Ritz, D., et al., *Data-Independent Acquisition of HLA Class I Peptidomes on the Q Exactive Mass Spectrometer Platform.* Proteomics, 2017. **17**(19).

242. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Mol Cell Proteomics, 2012. **11**(6): p. O111 016717.

243. Rosenberger, G., et al., *A repository of assays to quantify 10,000 human proteins by SWATH-MS.* Sci Data, 2014. **1**: p. 140031.

244. Schuster, H., et al., *A tissue-based draft map of the murine MHC class I immunopeptidome.* Sci Data, 2018. **5**: p. 180157.

245. Ronsein, G.E., et al., *Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics.* Journal of Proteomics, 2015. **113**: p. 388-399.
246. Bourmaud, A., S. Gallien, and B. Domon, *Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications.* Proteomics, 2016. **16**(15-16): p. 2146-59.
247. Peterson, A.C., et al., *Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics.* Mol Cell Proteomics, 2012. **11**(11): p. 1475-88.
248. Croft, N.P., A.W. Purcell, and D.C. Tscharke, *Quantifying epitope presentation using mass spectrometry.* Molecular immunology, 2015. **68**(2): p. 77-80.
249. Tan, C.T., et al., *Direct quantitation of MHC-bound peptide epitopes by selected reaction monitoring.* Proteomics, 2011. **11**(11): p. 2336-40.
250. Laumont, C.M., et al., *Noncoding regions are the main source of targetable tumor-specific antigens.* Science translational medicine, 2018. **10**(470): p. eaau5516.
251. Ebrahimi-Nik, H., et al., *Mass spectrometry driven exploration reveals nuances of neoepitope-driven tumor rejection.* JCI Insight, 2019. **5**.
252. Mann, M. and N.L. Kelleher, *Precision proteomics: The case for high resolution and high mass accuracy.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(47): p. 18132-18138.
253. Cox, J., et al., *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.
254. Yates, J.R., 3rd, *Recent technical advances in proteomics.* F1000Res, 2019. **8**.
255. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.
256. Patterson, S.D. and R.H. Aebersold, *Proteomics: the first decade and beyond.* Nat Genet, 2003. **33 Suppl**: p. 311-23.
257. Yan, W., R. Aebersold, and E.W. Raines, *Evolution of organelle-associated protein profiling.* J Proteomics, 2009. **72**(1): p. 4-11.
258. Konermann, L., et al., *Mass spectrometry combined with oxidative labeling for exploring protein structure and folding.* Mass Spectrom Rev, 2010. **29**(4): p. 651-67.
259. Mayya, V. and D.K. Han, *Phosphoproteomics by mass spectrometry: insights, implications, applications and limitations.* Expert Rev Proteomics, 2009. **6**(6): p. 605-18.
260. Bassani-Sternberg, M., et al., *Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.* Mol Cell Proteomics, 2015. **14**(3): p. 658-73.
261. Hoof, I., et al., *Proteome sampling by the HLA class I antigen processing pathway.* PLoS Comput Biol, 2012. **8**(5): p. e1002517.
262. Javitt, A., et al., *Pro-inflammatory Cytokines Alter the Immunopeptidome Landscape by Modulation of HLA-B Expression.* Front Immunol, 2019. **10**: p. 141.
263. Loffler, M.W., et al., *Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma.* Genome Med, 2019. **11**(1): p. 28.
264. Milner, E., et al., *The turnover kinetics of major histocompatibility complex peptides of human cancer cells.* Mol Cell Proteomics, 2006. **5**(2): p. 357-65.
265. Mommen, G.P., et al., *Sampling from the proteome to the HLA-DR ligandome proceeds via high specificity.* Mol Cell Proteomics, 2016.
266. Pearson, H., et al., *MHC class I-associated peptides derive from selective regions of the human genome.* J Clin Invest, 2016. **126**(12): p. 4690-4701.

267. Frese, C.K., et al., *Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry.* Anal Chem, 2012. **84**(22): p. 9668-73.

268. Mommen, G.P., et al., *Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD).* Proc Natl Acad Sci U S A, 2014. **111**(12): p. 4507-12.

269. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for mass spectrometry-based proteomics.* Methods Mol Biol, 2010. **604**: p. 55-71.

270. Gupta, N., et al., *Target-decoy approach and false discovery rate: when things may go wrong.* J Am Soc Mass Spectrom, 2011. **22**(7): p. 1111-20.

271. Mylonas, R., et al., *Estimating the contribution of proteasomal spliced peptides to the HLA-I ligandome.* Molecular & Cellular Proteomics, 2018. **17**(12): p. 2347-2357.

272. Rolfs, Z., et al., *Comment on "A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands".* Science Immunology, 2019. **4**(38): p. eaaw1622.

273. Faridi, P., et al., *Response to Comment on "A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands".* Science Immunology, 2019. **4**(38): p. eaaw8457.

274. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.

275. Jaffe, J.D., H.C. Berg, and G.M. Church, *Proteogenomic mapping as a complementary method to perform genome annotation.* Proteomics, 2004. **4**(1): p. 59-77.

276. Liepe, J., et al., *A large fraction of HLA class I ligands are proteasome-generated spliced peptides.* Science, 2016. **354**(6310): p. 354-358.

277. Zhao, S.R., et al., *Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap.* Bmc Genomics, 2015. **16**.

278. Levin, J.Z., et al., *Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nature methods, 2010. **7**(9): p. 709.

279. Blakeley, P., I.M. Overton, and S.J. Hubbard, *Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies.* J Proteome Res, 2012. **11**(11): p. 5221-34.

280. Ji, Z., et al., *Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins.* elife, 2015. **4**: p. e08890.

281. Pauli, A., E. Valen, and A.F. Schier, *Identifying (non-) coding RNAs and small peptides: Challenges and opportunities.* Bioessays, 2015. **37**(1): p. 103-112.

282. Chugunova, A., et al., *Mining for small translated ORFs.* Journal of proteome research, 2017. **17**(1): p. 1-11.

283. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.

284. Faridi, P., et al., *A subset of HLA-I peptides are not genomically templated: Evidence for cis-and trans-spliced peptide ligands.* Science immunology, 2018. **3**(28): p. eaar3947.

285. Laumont, C.M., et al., *Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames.* Nat Commun, 2016. **7**: p. 10238.

286. Calviello, L., et al., *Detecting actively translated open reading frames in ribosome profiling data.* Nat Methods, 2016. **13**(2): p. 165-70.

287. Erhard, F., et al., *Improved Ribo-seq enables identification of cryptic translation events.* Nat Methods, 2018. **15**(5): p. 363-366.

288. Ingolia, N.T., et al., *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.

289. Klebanoff, C.A. and J.D. Wolchok, *Shared cancer neoantigens: Making private matters public.* J Exp Med, 2018. **215**(1): p. 5-7.

290. Smart, A.C., et al., *Intron retention is a source of neoepitopes in cancer.* Nature biotechnology, 2018.

291. Zhang, M., et al., *RNA editing derived epitopes function as cancer antigens to elicit immune responses.* Nature communications, 2018. **9**(1): p. 3919.

292. Kong, Y., et al., *Transposable element expression in tumors is associated with immune infiltration and increased antigenicity.* Nature Communications, 2019. **10**.

293. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101.

294. Li, H., et al., *Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification.* BMC genomics, 2016. **17**(13): p. 1031.

295. Probst-Kepper, M., et al., *An alternative open reading frame of the human macrophage colony-stimulating factor gene is independently translated and codes for an antigenic peptide of 14 amino acids recognized by tumor-infiltrating CD8 T lymphocytes.* Journal of Experimental Medicine, 2001. **193**(10): p. 1189-1198.

296. Schiavetti, F., et al., *A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes.* Cancer Res, 2002. **62**(19): p. 5510-6.

297. Ho, O. and W.R. Green, *Cytolytic CD8+ T cells directed against a cryptic epitope derived from a retroviral alternative reading frame confer disease protection.* J Immunol, 2006. **176**(4): p. 2470-5.

298. Mullins, C.S. and M. Linnebacher, *Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes.* Cancer Immunology Immunotherapy, 2012. **61**(7): p. 1093-1100.

299. Chong, C., et al., *High-throughput and sensitive immunopeptidomics platform reveals profound IFNgamma-mediated remodeling of the HLA ligandome.* Mol Cell Proteomics, 2017.

300. Krokhin, O., *Peptide retention prediction in reversed-phase chromatography: proteomic applications.* Expert Review of Proteomics, 2012. **9**(1): p. 1-4.

301. Lang, D., J.B. Mascarenhas, and C.R. Shea, *Melanocytes, melanocyte stem cells, and melanoma stem cells.* Clinics in Dermatology, 2013. **31**(2): p. 166-178.

302. Tachibana, M., et al., *Ectopic expression of MITF, a gene for Waardenburg syndrome type 2, converts fibroblasts to cells with melanocyte characteristics.* Nat Genet, 1996. **14**(1): p. 50-4.

303. Widlund, H.R., et al., *Beta-catenin-induced melanoma growth requires the downstream target Microphthalmia-associated transcription factor.* J Cell Biol, 2002. **158**(6): p. 1079-87.

304. Carithers, L.J. and H.M. Moore, *The Genotype-Tissue Expression (GTEx) Project.* Biopreserv Biobank, 2015. **13**(5): p. 307-8.

305. Müller, M., et al., *'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization.* Frontiers in Immunology, 2017. **8**(1367).

306. Schuster, H., et al., *The immunopeptidomic landscape of ovarian carcinomas.* Proc Natl Acad Sci U S A, 2017. **114**(46): p. E9942-E9951.

307. Klatt, M.G., et al., *Carcinogenesis of renal cell carcinoma reflected in HLA ligands: A novel approach for synergistic peptide vaccination design.* Oncoimmunology, 2016. **5**(8): p. e1204504.

308. Berlin, C., et al., *Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy.* Leukemia, 2016. **30**(4): p. 1003-4.

309. Khodadoust, M.S., et al., *Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens.* Nature, 2017. **543**(7647): p. 723-727.

310. Kearney, P., et al., *The building blocks of successful translation of proteomics to the clinic.* Curr Opin Biotechnol, 2018. **51**: p. 123-129.

311. Heather, J.M., et al., *Murine xenograft bioreactors for human immunopeptidome discovery.* Sci Rep, 2019. **9**(1): p. 18558.

312. Faridi, P., R. Aebersold, and E. Caron, *A first dataset toward a standardized community-driven global mapping of the human immunopeptidome.* Data Brief, 2016. **7**: p. 201-5.

313. Caron, E., et al., *An open-source computational and data resource to analyze digital maps of immunopeptidomes.* Elife, 2015. **4**.

314. Ghosh, M., et al., *Validation of a high-performance liquid chromatography-tandem mass spectrometry immunopeptidomics assay for the identification of HLA class I ligands suitable for pharmaceutical therapies.* BioRxiv, 2019: p. 821249.

315. Castro, F., et al., *Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion.* Frontiers in Immunology, 2018. **9**.

316. Karachaliou, N., et al., *Interferon gamma, an important marker of response to immune checkpoint blockade in non-small cell lung cancer and melanoma patients.* Therapeutic Advances in Medical Oncology, 2018. **10**.

317. Ramakrishna, V., et al., *Naturally occurring peptides associated with HLA-A2 in ovarian cancer cell lines identified by mass spectrometry are targets of HLA-A2-restricted cytotoxic T cells.* Int Immunol, 2003. **15**(6): p. 751-63.

318. Viborg, N., et al., *T cell recognition of novel shared breast cancer antigens is frequently observed in peripheral blood of breast cancer patients.* Oncoimmunology, 2019. **8**(12).

319. Westergaard, M.C.W., et al., *Correction: Tumour-reactive T cell subsets in the microenvironment of ovarian cancer.* Br J Cancer, 2019. **120**(8): p. 870.

320. Shraibman, B., et al., *HLA peptides derived from tumor antigens induced by inhibition of DNA methylation for development of drug-facilitated immunotherapy.* Mol Cell Proteomics, 2016.

321. Bauer, J., et al., *Mass Spectrometry-Based Immunopeptidome Analysis of Acute Myeloid Leukemia Cells Under Decitabine Treatment Delineates Induced Presentation of Cancer/Testis Antigens on HLA Class I Molecules.* Blood, 2018. **132**.

322. Molognoni, F., et al., *Epigenetic reprogramming as a key contributor to melanocyte malignant transformation.* Epigenetics, 2011. **6**(4): p. 451-465.

323. Micevic, G., et al., *Attenuation of genome-wide 5-methylcytosine level is an epigenetic feature of cutaneous malignant melanomas.* Melanoma research, 2017. **27**(2): p. 85.

324. Li, J.-L., et al., *Genome-wide methylated CpG island profiles of melanoma cells reveal a melanoma coregulation network.* Scientific reports, 2013. **3**: p. 2962.

325. Tellez, C.S., et al., *CpG island methylation profiling in human melanoma cell lines.* Melanoma research, 2009. **19**(3): p. 146-155.

326. Creech, A.L., et al., *The role of mass spectrometry and proteogenomics in the advancement of HLA epitope prediction.* Proteomics, 2018. **18**(12): p. 1700259.

327. Bassani-Sternberg, M., et al., *Soluble plasma HLA peptidome as a potential source for cancer biomarkers.* Proc Natl Acad Sci U S A, 2010. **107**(44): p. 18769-76.

328. Ritz, D., et al., *Purification of soluble HLA class I complexes from human serum or plasma deliver high quality immuno peptidomes required for biomarker discovery.* Proteomics, 2017. **17**(1-2).

329. Seale, B., et al., *Digital Microfluidics for Immunoprecipitation.* Anal Chem, 2016. **88**(20): p. 10223-10230.

330. Chiu, D.T., et al., *Small but perfectly formed? Successes, challenges, and opportunities for microfluidics in the chemical and biological sciences.* Chem, 2017. **2**(2): p. 201-223.

331. Michalski, A., J. Cox, and M. Mann, *More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS.* J Proteome Res, 2011. **10**(4): p. 1785-93.

332. Pfammatter, S., et al., *A Novel Differential Ion Mobility Device Expands the Depth of Proteome Coverage and the Sensitivity of Multiplex Proteomic Measurements.* Molecular & Cellular Proteomics, 2018. **17**(10): p. 2051-2067.

333. d'Atri, V., et al., *Adding a new separation dimension to MS and LC–MS: What is the utility of ion mobility spectrometry?* Journal of separation science, 2018. **41**(1): p. 20-67.

334. Pfammatter, S., E. Bonneil, and P. Thibault, *Improvement of Quantitative Measurements in Multiplex Proteomics Using High-Field Asymmetric Waveform Spectrometry.* Journal of Proteome Research, 2016. **15**(12): p. 4653-4665.

335. Sandow, J.J., et al., *Simplified high-throughput methods for deep proteome analysis on the timsTOF Pro.* BioRxiv, 2019: p. 657908.

336. Meier, F., et al., *Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer.* Mol Cell Proteomics, 2018. **17**(12): p. 2534-2545.

337. Marino, F., et al., *High-Throughput, Fast, and Sensitive Immunopeptidomics Sample Processing for Mass Spectrometry.* Methods Mol Biol, 2019. **1913**: p. 67-79.

338. Bassani-Sternberg, M., et al., *A Phase Ib Study of the Combination of Personalized Autologous Dendritic Cell Vaccine, Aspirin, and Standard of Care Adjuvant Chemotherapy Followed by Nivolumab for Resected Pancreatic Adenocarcinoma-A Proof of Antigen Discovery Feasibility in Three Patients.* Frontiers in Immunology, 2019. **10**.

339. Aldous, A.R. and J.Z. Dong, *Personalized neoantigen vaccines: A new approach to cancer immunotherapy.* Bioorg Med Chem, 2018. **26**(10): p. 2842-2849.

340. Katsnelson, A., *Mutations as munitions: Neoantigen vaccines get a closer look as cancer treatment.* Nature Medicine, 2016. **22**(2): p. 122-124.

341. Bouchie, A. and L. DeFrancesco, *Nature Biotechnology's academic spinouts of 2015.* Nature Biotechnology, 2016. **34**(5): p. 484-492.

342. Fritsche, J., et al., *Translating Immunopeptidomics to Immunotherapy-Decision-Making for Patient and Personalized Target Selection.* Proteomics, 2018. **18**(12): p. 1700284.

343. Freudenmann, L.K., A. Marcu, and S. Stevanovic, *Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry.* Immunology, 2018. **154**(3): p. 331-345.

344. Wolf, Y., et al., *UVB-Induced Tumor Heterogeneity Diminishes Immune Response in Melanoma.* Cell, 2019. **179**(1): p. 219-+.

345. Zhang, B., et al., *Clinical potential of mass spectrometry-based proteogenomics.* Nature Reviews Clinical Oncology, 2019. **16**(4): p. 256-268.

346. Samstein, R.M., et al., *Tumor mutational load predicts survival after immunotherapy across multiple cancer types.* Nature Genetics, 2019. **51**(2): p. 202-+.

347. Sexton, C.E. and M.V. Han, *Paired-end mappability of transposable elements in the human genome.* Mobile DNA, 2019. **10**.

348. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Res, 2012. **22**(9): p. 1775-89.

349. Zhao, S.R., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA plus selection versus rRNA depletion.* Scientific Reports, 2018. **8**.

350. Weinzierl, A.O., et al., *Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface.* Mol Cell Proteomics, 2007. **6**(1): p. 102-13.

351. Shteynberg, D., et al., *Combining Results of Multiple Search Engines in Proteomics.* Molecular & Cellular Proteomics, 2013. **12**(9): p. 2383-2393.

352. Quandt, A., et al., *Using synthetic peptides to benchmark peptide identification software and search parameters for MS/MS data analysis.* EuPA Open Proteomics, 2014. **5**: p. 21-31.

353. Rech, A.J., et al., *Tumor immunity and survival as a function of alternative neopeptides in human cancer.* Cancer immunology research, 2018. **6**(3): p. 276-287.

354. Ghorani, E., et al., *Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma.* Annals of Oncology, 2018. **29**(1): p. 271-279.

355. Duan, F., et al., *Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity.* J Exp Med, 2014. **211**(11): p. 2231-48.

356. Gross, D.-A., et al., *High vaccination efficiency of low-affinity epitopes in antitumor immunotherapy.* The Journal of clinical investigation, 2004. **113**(3): p. 425-433.

357. Prabakaran, S., et al., *Quantitative profiling of peptides from RNAs classified as noncoding.* Nature Communications, 2014. **5**.

358. Rauniyar, N., *Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry.* International Journal of Molecular Sciences, 2015. **16**(12): p. 28566-28581.

359. Calviello, L. and U. Ohler, *Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome.* Trends Genet, 2017. **33**(10): p. 728-744.

360. Zhao, J., et al., *Translatomics: The Global View of Translation.* Int J Mol Sci, 2019. **20**(1).

361. Gilboa, E., *The risk of autoimmunity associated with tumor immunotherapy.* Nat Immunol, 2001. **2**(9): p. 789-92.

362. Marshall, M.J., R.J. Stopforth, and M.S. Cragg, *Therapeutic antibodies: what have we learnt from targeting CD20 and where are we going?* Frontiers in immunology, 2017. **8**: p. 1245.

363. Shah, N.N., et al., *Multi Targeted CAR-T Cell Therapies for B-Cell Malignancies.* Front Oncol, 2019. **9**: p. 146.

364. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & development, 2011. **25**(18): p. 1915-1927.

365. Aarnoudse, C.A., et al., *Interleukin-2-induced, melanoma-specific T cells recognize camel, an unexpected translation product of LAGE-1.* International Journal of Cancer, 1999. **82**(3): p. 442-448.

366. Rosenberg, S.A., et al., *Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy.* J Immunol, 2002. **168**(5): p. 2402-7.

367. Mayrand, S.M., D.A. Schwarz, and W.R. Green, *An alternative translational reading frame encodes an immunodominant retroviral CTL determinant expressed by an immunodeficiency-causing retrovirus.* J Immunol, 1998. **160**(1): p. 39-50.

368. Cardinaud, S., et al., *Identification of cryptic MHC I-restricted epitopes encoded by HIV-1 alternative reading frames.* J Exp Med, 2004. **199**(8): p. 1053-63.

369. Schatton, T., et al., *Identification of cells initiating human melanomas.* Journal of Investigative Dermatology, 2008. **128**: p. S213-S213.

370. Sana, G., et al., *Exome Sequencing of ABCB5 Identifies Recurrent Melanoma Mutations that Result in Increased Proliferative and Invasive Capacities.* Journal of Investigative Dermatology, 2019. **139**(9): p. 1985-+.

371. Wilson, B.J., et al., *ABCB5 Maintains Melanoma-Initiating Cells through a Proinflammatory Cytokine Signaling Circuit.* Cancer Research, 2014. **74**(15): p. 4196-4207.

372. Andersen, R.S., et al., *Dissection of T-cell antigen specificity in human melanoma.* Cancer Res, 2012. **72**(7): p. 1642-50.

373. Gallimore, A., et al., *Protective immunity does not correlate with the hierarchy of virus-specific cytotoxic T cell responses to naturally processed peptides.* Journal of Experimental Medicine, 1998. **187**(10): p. 1647-1657.

374. La Gruta, N.L., et al., *A virus-specific CD8+ T cell immunodominance hierarchy determined by antigen dose and precursor frequencies.* Proc Natl Acad Sci U S A, 2006. **103**(4): p. 994-9.

375. Marchand, M., et al., *Tumor regressions observed in patients with metastatic melanoma treated with an antigenic peptide encoded by gene MAGE-3 and presented by HLA-A1.* International journal of cancer, 1999. **80**(2): p. 219-230.

376. Anichini, A., et al., *An expanded peripheral T cell population to a cytotoxic T lymphocyte (CTL)-defined, melanocyte-specific antigen in metastatic melanoma patients impacts on generation of peptide-specific CTLs but does not overcome tumor escape from immune surveillance in metastatic lesions.* The Journal of experimental medicine, 1999. **190**(5): p. 651-668.

377. Rosenberg, S.A., et al., *Immunologic and therapeutic evaluation of a synthetic peptide vaccine for the treatment of patients with metastatic melanoma.* Nature medicine, 1998. **4**(3): p. 321-327.

378. Kahles, A., et al., *SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data.* Bioinformatics, 2016. **32**(12): p. 1840-7.

379. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nat Rev Genet, 2016. **17**(2): p. 93-108.

380. Krokhin, O.V. and V. Spicer, *Generation of accurate peptide retention data for targeted and data independent quantitative LC-MS analysis: Chromatographic lessons in proteomics.* Proteomics, 2016. **16**(23): p. 2931-2936.

381. Hsu, P.Y., et al., *Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis.* Proc Natl Acad Sci U S A, 2016. **113**(45): p. E7126-E7135.