Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering, and Natural Sciences

of the

Ruprecht-Karls-University

Heidelberg

Presented by

Arne Knudsen, M.Sc.

born in Bremen, Germany

Oral examination: 28.07.2022

Alternative pre-mRNA splicing as a source of cancer neoepitopes

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. Dr. Andreas Kulozik

I. Abstract

Robust identification of neoepitopes is crucial for the efficacy and safety of immunotherapy, the most promising treatment strategy for several cancer types. Current approaches have provided limited numbers of immunogenic and tumor-specific targets, thus preventing the broad application of targeted immunotherapy. Here, the focus on somatic mutation-derived neoantigens often overlooks possible neoepitopes originating from mRNA processing events. A potential new source of tumor-specific peptides is alternative pre-mRNA splicing, a widely dysregulated process in several cancer subtypes. However, there is limited insight regarding the potential of alternative splicing to generate peptides that are also presented on the cell surface. Thus, in this thesis, I aimed to investigate how perturbation of the splicing machinery contributes to the neoepitope repertoire in tumor cells.

To explore alternative splicing-derived neoantigens, I performed immunopeptidomics to determine the HLA-I ligandome of wild-type RPE-1 cells and RPE-1 cell lines carrying common cancer mutations. To facilitate the presentation of alternative splicing-derived neoepitopes, I treated these cell lines with the splicing inhibitor GEX1A. I then performed HLA-I immunopurification to recover HLA-I-bound peptides of these cells, followed by peptide identification through mass spectrometry. To be able to identify non-canonical peptides from mass spectra, I generated sample-specific custom reference databases based on matching RNA-seq data. This strategy allowed me to identify more than 8,000 unique HLA-I-presented peptides per cell line.
In parallel, I specifically identified neoepitopes originating from aberrant alternative splice events. By performing differential splicing analysis between the various conditions, I obtained thousands of differentially regulated splice junction events. Particularly in cells treated with the splicing inhibitor GEX1A, alternative splicing analysis revealed numerous novel, non-annotated splice events. To examine whether these dysregulated events were translated into novel peptides, I subsequently mapped the candidate peptides to the differential splice events. With this strategy, I was able to identify and validate several alternative splicing-derived neoepitope candidates that exhibited a high immunogenic potential in *in vivo* immunization assays.

In conclusion, my work demonstrates that pharmacological modulation of the splicing machinery has the potential to promote the presentation of neoepitopes derived from alternative splice variants. These findings have potential implications for immunotherapy of cancer types with low tumor mutational burden. Exploring the splicing-derived neopeptidome could reveal novel therapeutic targets and serve as a predictive biomarker for response to immune checkpoint blockade therapy.

Note: A previous version of this abstract was published at CIMT 2021 Conference

II. Zusammenfassung

Die Krebsimmuntherapie ist die zurzeit vielversprechendste Behandlungsmethode für eine Vielzahl von Tumorerkrankungen. Für die Wirksamkeit und Sicherheit einer zielgerichteten Therapie ist die Identifizierung von Neoepitopen essenziell. Allerdings waren bisherige Strategien, diese tumorspezifischen Peptide zu identifizieren, nur begrenzt erfolgreich. Da hier der Fokus vor allem auf Neoepitopen liegt, die aus Veränderungen der DNA-Sequenz resultieren, werden andere Mechanismen, die Neoepitope hervorbringen können, oftmals vernachlässig. Eine vielversprechende weitere Quelle für Neoepitope ist der Prozess des alternativen prä-mRNA Spleißens, welcher in vielen Krebsarten beeinträchtigt ist. Allerdings gibt es bisher nur wenige Einblicke, inwiefern diese alternativen Spleißevents Peptide produzieren können, die dann auf der Zelloberfläche als Neoepitope präsentiert werden können. Daher hatte meine Doktorarbeit das Ziel, herauszufinden, ob die Beeinträchtigungen des Spleißprozesses in Tumorzellen deren Neoepitopanzahl erweitern und dadurch die Anwendung von Krebsimmuntherapien verbessern kann.

Mithilfe von Immunopeptidomics habe ich das HLA-I-Ligandom in RPE-1-Zelllinien erforscht, wobei ich sowohl Wildtyp-Zelllinien untersucht habe als auch Zellen mit Defekten im DNA-Reparaturmechanismus, welche häufig in Tumorzellen auftreten. Diese Zellen habe ich mit dem Spleiß-Inhibitor GEX1A behandelt, um den deregulierten Spleiß-Phänotypen in Krebszellen nachzustellen. Dies sollte zu vermehrter Präsentation von Peptiden führen, die durch alternative Spleißevents entstehen. Mithilfe einer HLA-I-Immunopurifikation konnte ich die HLA-I gebundenen Peptide aufreinigen und deren Peptidsequenzen mittels Massenspektrometrie analysieren. Dafür habe ich basierend auf RNA-Sequenzierungsdaten spezifische Datenbanken generiert, welche mir ermöglicht haben, auch neuartige, nicht annotierte Peptide zu identifizieren. Insgesamt konnte ich so mehr als 8000 verschiedene HLA-I-Peptide für jede Zelllinie detektieren.

Zusätzlich habe ich das Transkriptom dieser Zellen analysiert, um alternative Speißevents zu identifizieren, indem ich eine differentiale Spleißanalyse zwischen den einzelnen Konditionen durchgeführt habe. Dadurch konnte ich tausende deregulierte Spleißevents entdecken, besonders in den Zellen, welche mit Spleiß-Inhibitor behandelt wurden. Diese Spleißevents habe ich in Peptidsequenzen translatiert, um anschließend die HLA-I-Peptide damit abzugleichen. Dadurch konnte ich mehrere Neoepitope feststellen, die von alternativen Spleißevents stammen. Schließlich habe ich die Immunogenität dieser alternativen Spleiß-Neoepitope mithilfe von *in vivo* Experimenten untersucht, wodurch ich das hohe immunogene Potential der Kandidaten nachweisen konnte.

In meiner Doktorarbeit konnte ich zeigen, dass die pharmakologische Modulation des RNA-Spleißprozesses das Potential aufweist, Neoepitope zu generieren, die durch alternative Spleißevents entstehen. Die Ergebnisse meiner Arbeit eröffnen neue therapeutische Möglichkeiten für die Krebsimmuntherapie auf. Insbesondere in Krebsarten, die kaum somatische Mutationen aufweisen, können Neoepitope, die von alternativen Spleißevents stammen, neue Behandlungenstrategien ermöglichen. Außerdem könnte die durch fehlerhaftes Spleißen verursachte Tumorlast ein neuer prädiktiver Biomarker für die Therapie mit Immuncheckpoint-Inhibitoren werden.

III. Acknowledgements

During the past four years, I could count on many people who helped, supported, and guided me on my journey towards my PhD. In particular, I would like to thank:

Jan Mauer, for giving me the opportunity to conduct my research at your group and supervise me throughout the years. Thank you, Jan, for your guidance in research and life, for challenging and motivating me, and for keeping my PhD journey on track.

Lina Gallego-Paez, for your supervision and being a great source of knowledge whenever I had a question on my mind. Thank you, Lina, for being my Sudo-san, solving problems in my bioinformatic code, and for your inspiration.

Merck KGaA, as well as **Christian Tidona and the whole BioMed X support team** for funding my PhD project and enabling me to conduct my research at BioMed X.

Christina Esdar and **Juliane Braun,** for your support and scientific input during our monthly meetings.

Andreas Kulozik, for being the initial spark of this project and your input and support in numerous meetings and as a member of my Thesis Advisory Committee.

Jonas Becker, for being a great help to establish the methodological part of the immunopeptidomics workflow by sharing your resources and knowledge. Thank you for answering all my questions.

Alejandro Hernandez-Sanchez, **Matthias Kloor**, and **Magnus von Knebel Doeberitz**, for your expertise and experimental support on the immunological part of my PhD project.

My dear colleagues and friends at BioMed X, for being there whenever I needed your support and making BioMed X a place where I enjoyed going to work (almost) every day.

Benedikt Brors, **Christiane Opitz**, and **Daniel Hübschmann** for evaluating my thesis and being my Thesis Defense Committee members.

The Lunch Crew members, for all the good times and conversations about all sorts of things in life, which gave me a break from my daily routine.

My friends, for freeing my mind from science and being fellow sufferers on the PhD journey.

My family, especially my parents **Kirsten** and **Karl**, for your support and motivation throughout the whole time.

Carolin, for your love and support, for inspiring me, and for being my partner in all different aspects of life. Thank you for being there for me since the early days of my studies. Without you, I would not have made it!

IV. Table of Contents

V. Abbreviations

Abbreviation	Meaning
aa	Amino acid(s)
ACN	Acetonitrile
AML	Acute myeloid leukemia
ASO	Antisense oligonucleotide
BCA	Bicinchoninic acid
BRD9	Bromodomain-Containing Protein 9
BSA	Bovine serum albumin
CAR	Chimeric antigen receptor
CML	Chronic myeloid leukemia
ConA	Concanavalin A
CTA	Cancer/testis antigen
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DKMS	Deutsche Knochenmarkspenderdatei
DMP	Dimethyl pimelimidate
DTT	Dithiothreitol
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
ER	Endoplasmic reticulum
FBS	Fetal bovine serum
FDR	False discovery rate
GR	Growth rate
GSEA	Gene Set Enrichment Analysis
HCD	Higher-energy collisional dissociation
HLA-I	Human leukocyte antigen I
HLA-I IP	HLA- I immunoaffinity purification
hnRNP	Heterogeneous nuclear ribonucleoproteins
HNSC	Head and neck cancer
HPV	Human papilloma virus
HRP	Horseradish peroxidase
ID	Identifier
IDH2	Isocitrate Dehydrogenase (NADP(+)) 2
kDA	Kilodalton
LDS	Lithiumdodecyl sulfate
logFC	log fold change
MES	2-(N-morpholino)ethanesulfonic acid
MMR	Mismatch repair
MMR-D	Mismatch repair deficiency/deficient
MS	Mass spectrometry
NMD	Nonsense-mediated decay

nt	Nucleotide(s)
ORF	Open reading frame
pAdj	Adjusted p-value
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PD1	Programmed cell death protein 1
PD-1	Programmed death protein 1
PD-L1	Programmed death ligand 1
poly(Y)	Polypyrimidine
ppm	Parts per million
PRMT5	Protein arginine N-methyltransferase 5
PROTACS	Proteolysis targeting chimeras
psi	Pound per square inch
PSM	Peptide-spectrum match
PTC	Premature termination codon
RBM protein	RNA binding motif protein
RBM39	RNA-binding protein 39
RBP	RNA binding protein
RPKM	Reads per kilobase of transcript per Million mapped reads
rpm	Revolutions per minute
RT	Room temperature
RT-PCR	Reverse transcription polymerase chain reaction
SDS-PAGE	Sodium dodecyl sulphate–polyacrylamide gel electrophoresis
SF3B1	Splicing factor 3B subunit 1
SJDB	Splice Junction-derived Database
snRNP	Small ribonucleoprotein complex
SNV	Somatic nucleotide variant
SR protein	Serine/arginine-rich protein
SRSF2	Serine/arginine-rich splicing factor 2
TAA	Tumor-associated antigen
TAP	Transporter associated with antigen processing
TBS-T	Tris-buffered saline supplemented with Tween 20
TCGA	The Cancer Genome Atlas
TCR	T cell receptor
TFA	Trifluoracetic acid
TIL	Tumor infiltrating lymphocytes
T_m	Melting temperature
TP53	Tumor protein P53
TSA	Tumor-specific antigen
U2AF1	U2 Small Nuclear RNA Auxiliary Factor 1
x*g*	Times gravity
ZRSR2	Zinc finger CCCH-type, RNA binding motif and serine/arginine-rich protein 2

VI. List of figures

VII. List of tables

1. Introduction

1.1 Alternative mRNA splicing

1.1.1 Conventional mRNA splicing

Almost all human premature mRNA transcripts undergo RNA splicing as a crucial part of their maturation process to remove non-coding intronic regions from the exonic coding regions within the transcript before it is translated into protein (Gallego-Paez et al., 2017). In 99% of human introns, this task is carried out by the major spliceosome, a tightly regulated machinery consisting of five small ribonucleoprotein complexes (snRNPs) (Dvinge et al., 2016). The remaining 1% introns are targeted by the minor spliceosome consisting of a different snRNP composition with similar functional properties. The assembly of spliceosomes is promoted by the recognition of splice sites through conserved nucleotide elements in the pre-mRNA (Figure 1). These short sequence motifs mainly occur at exon-intron boundaries. The donor or 5' splice site is characterized by a GU sequence element, the acceptor or 3' splice site by an AG sequence element. In addition, a polypyrimidine tract and a branch point facilitate the recruitment of the spliceosome (Figure 1).

The spliceosomal assembly is further coordinated by *cis*-acting sequence elements that are bound by *trans*-acting serine/arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs), which can both either promote or suppress splicing depending on the vicinity at the mRNA binding site (Zhang and Manley, 2013). Apart from SR protein and hnRNPs, the recruitment of the spliceosome is governed by various other RNA-binding proteins (RBPs) (Fu and Ares, 2014) as well as by alterations in the expression levels of spliceosomal components (Saltzman et al., 2011). Studies that aimed to reveal the complexity of the regulatory system of the splicing machinery identified more than 200 proteins associated with the spliceosome (Wahl et al., 2009) and detected hundreds of regulatory sequence motifs (Barash et al., 2010). To date, the exact number of splicing factors and other regulators of the splicing machinery remains unknown. Indeed, splicing is mediated by the complex interplay between *cis*-regulatory elements and *trans*-acting proteins, especially at alternative splice sites (Wang and Aifantis, 2020).

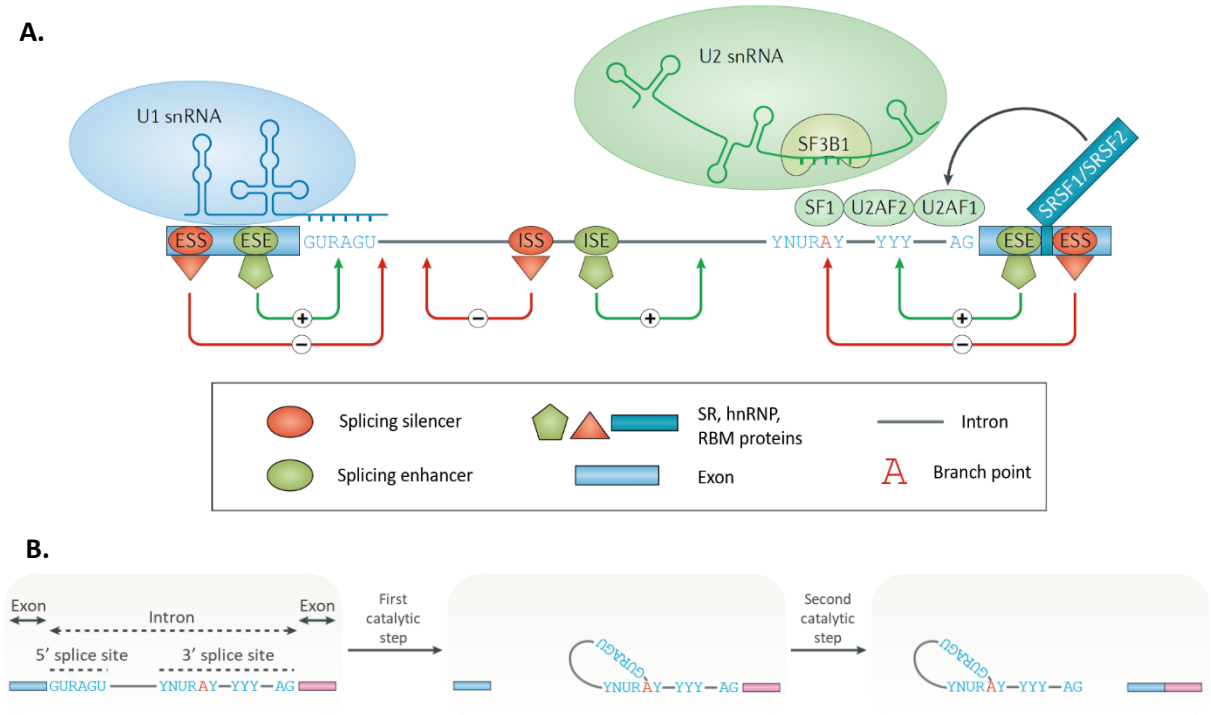


Figure 1 | Mechanism and regulation of mRNA splicing. *Assembly of the spliceosome is facilitated by specific sequence elements and governed by exonic or intronic regulatory elements. U1 small nuclear ribonucleoprotein (U1 snRNP) recognizes a GU sequence element at the 5′ end of the intron, the splice donor site. Recognition of the splice acceptor site at the 3′ end of the intron by U2 snRNP is facilitated by various splicing factors that bind to the AG sequence element and the further upstream polypyrimidine tract and branch point. The spliceosomal assembly is promoted or prevented by the binding of serine/arginine-rich (SR) proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs), and RNA binding motif (RBM) proteins to exonic and intronic regulatory elements which affect splice site choices and thus alternative splicing.* ***(B)*** *Two concatenated transesterification reactions remove the intronic sequence during canonical RNA splicing. The first step generates a lariat-shaped structure with a phosphodiester bond between the splice donor site and the branch point 15-30 nucleotides upstream of the 3′ end of the intron. In the second step, the intron is released as lariat and the 5′ exon and the 3′ exon are spliced together. ESE/ESS = exonic splicing enhancer/silencer; ISE/ISS = intronic splicing enhancer/silencer; SF1 = splicing factor 1; SF3B1 = splicing factor 3B subunit 1 (SF3B1); SRSF1/2 = serine/arginine-rich splicing factor 1/2; U2AF1/2 = U2 Small Nuclear RNA Auxiliary Factor 1. Figures reprinted and adapted from (Bonnal et al., 2020) with permission from Springer Nature under license number 5241381017664.*

1.1.2 The biological significance of alternative splicing

In addition to canonical intron splicing, exonic elements can be assembled differently so that several distinct transcripts can be derived from a similar mRNA precursor sequence. This process of alternative splicing was first discovered in the primary transcript of human adenovirus type 2, which comprises almost the whole viral genome but is spliced into numerous smaller mRNA transcripts (Chow et al., 1977). While the first alternatively spliced transcripts were identified through an electron microscope, the advances in genomic sequencing allow a far more comprehensive assessment of the alternative splicing landscape. In 2021, more than 900 different mRNA transcripts of human adenovirus 2 were identified using long-read sequencing (Westergren Jakobsson et al., 2021). This highlights how alternative splicing can enhance the coding capacity of an organism, a principle that was found to

increase with organismal complexity (Lee and Rio, 2015). In humans, nearly 95% of genes produce at least one alternative transcript isoform (Pan et al., 2008; Wang et al., 2008).

The alternative mRNA transcripts not only differ in stability, localization, or translation, but alternative splicing can also introduce changes in the reading frame, thereby producing distinct protein isoforms with different cellular functionalities (Baralle and Giudice, 2017). Thus, alternative splicing is helpful for cells to obtain proteins with different functions from the same genetic information at different developmental stages, for example. Indeed, it has been demonstrated that alternative splicing is necessary for cell differentiation, lineage determination, tissue identity acquisition and maintenance, as well as organ development (Gallego-Paez et al., 2017; Wang et al., 2008). Similar to conventional RNA splicing, the process of alternative splicing is tightly regulated by *cis*- and *trans*-acting regulatory elements. Hence, the expression of a particular splicing factor during cell development can change splice patterns by influencing splice site selection. Dependent on the selected splice site, different types of splice events can occur (Figure 2). Thereby, alternative splicing networks mediate the transition from embryonic to adult functions (Baralle and Giudice, 2017). Investigating these splicing transitions is crucial for understanding pathological mechanisms implicated in human diseases, such as cancer, where the highly orchestrated splicing machinery is largely affected by mutations and dysregulations.

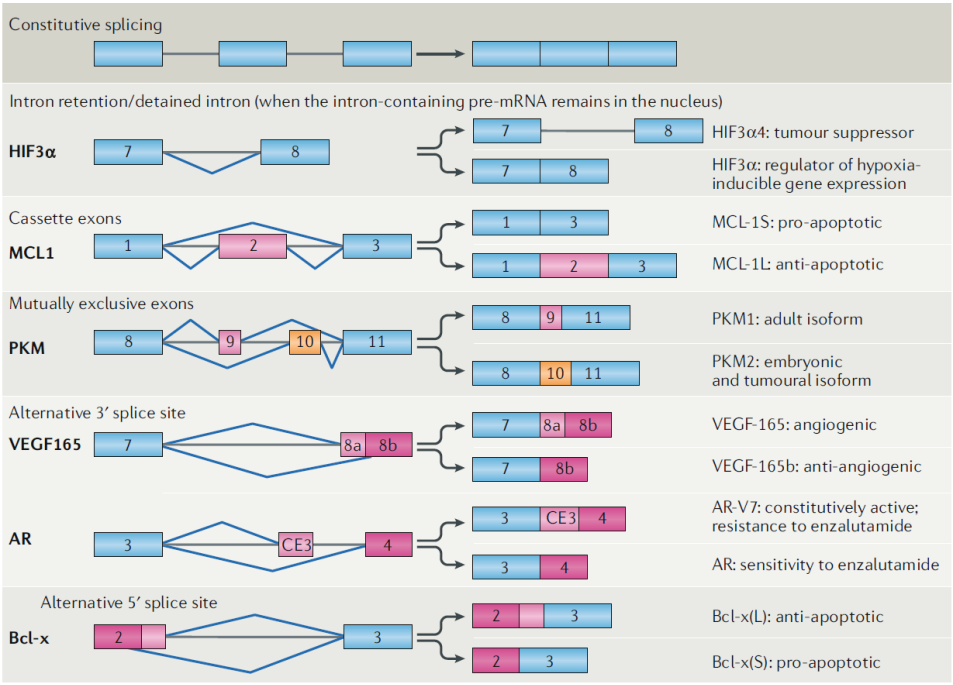


Figure 2 | Different classes of alternative mRNA processing events and their implication in cancer. *The expression of alternative isoforms is modulated by different splice site choices. Tumor cells exploit perturbations in their splicing machinery to promote the expression of isoforms that facilitate tumor growth and survival. Figure reprinted from (Bonnal et al., 2020) with permission from Springer Nature under license number 5241381017664.*

1.2 Aberrant alternative splicing – the 15th hallmark of cancer?

Considering the crucial process of RNA splicing in gene regulation and the essential role of alternative splicing in cellular processes such as proliferation, survival, and differentiation, it is unsurprising that altered splice patterns are associated with many human malignancies (Wang and Aifantis, 2020). These alterations result from mutations and dysregulated expression of spliceosomal components and regulators (Baeza-Centurion et al., 2019; Jayasinghe et al., 2018), causing an increase in splicing burden leading to alternative junction usage in patients (Kahles et al., 2018). Given its significant contribution to cancer pathogenesis, aberrant alternative splicing was described as an underappreciated hallmark of cancer (Bonnal et al., 2020; Farina et al., 2020; Ladomery, 2013; Oltean and Bates, 2014; Seiler et al., 2018a).

1.2.1 Dysregulation of oncogenes and tumor suppressors

At post-transcriptional level, tumorigenesis is frequently promoted by activation of oncogenes or loss of tumor suppressor function. Here, MYC is a prominent oncogenic transcription factor whose function is frequently altered in cancer cells (Wang and Aifantis, 2020). MYC can induce the expression of spliceosomal components, such as snRNPs, leading to splicing changes associated with cancer cell survival (Koh et al., 2015). In addition, splicing of important tumor suppressor genes was found to be impaired in cancer patients due to somatic nucleotide variants (SNVs), which affect splice site recognition, causing intron retention and thus degradation of tumor protein P53 (*TP53*) transcripts (Jung et al., 2015). In general, several *TP53*/p53 isoforms are known to have implications for cancer biology and were associated with cancer therapy and patient prognosis (Vieler and Sanyal, 2018).

A pan-cancer splicing analysis of cancer patient data from The Cancer Genome Atlas (TCGA) reported that aberrant alternative transcripts are frequently found across tumors compared to normal tissue (Kahles et al., 2018). Indeed, oncogenes in cancer cells often undergo isoform switches to enhance tumorigenesis. For example, cancer cells modulate isoform usage of the *BCL2L1* gene, favoring the expression of the longer protein isoform BCL-XL, which blocks apoptosis over the shorter *BCL-XS* isoform with tumor suppressor function (Danial, 2007). Isoform switches also facilitate tumors to develop resistance against cancer therapies, as observed in melanoma tumors which adapt to B-Raf inhibitor treatment by expressing splice variants of the oncogene $BRAF^{V600E}$ missing the RAS-binding domain (Poulikakos et al., 2010).

Cancer cells generate these aberrant splicing patterns by overburdening the splicing machinery, resulting in the generation of cancer-specific isoforms (Kahles et al., 2018). Hence, several SR proteins, hnRBPs, and other RBPs are known to function as proto-oncogenes or tumor suppressors and their activity is frequently exploited by cancer cells to promote tumor growth (Zhang and Manley, 2013). A genomic and transcriptomic analysis revealed broad alterations in the expression of RBPs important for the regulation of (alternative) splicing (Sebestyén et al., 2016). For example, downregulation of serine/arginine-rich splicing factor 3 (*SRSF3*), which governs splicing of *TP53*, produces an alternative spliced p53 isoform that promotes p53-mediated cellular senescence (Tang et al., 2013).

In addition to dysregulated expression of splicing factors, it was reported that 119 splicing factor genes are recurrently mutated across various cancer types with a significant impact on splice patterns of patients (Seiler et al., 2018a). Since these mutations result in loss or gain of gene function, they play a tumor-suppressive or oncogene-like role in cancer development and undermine the role of aberrant alternative splicing for tumorigenesis.

1.2.2 Mutations in core spliceosomal proteins

Mutations in the core components of the spliceosome are recurrently found across cancer types, especially in hematopoietic malignancies (Yoshida et al., 2011). More recently, it has been summarized that spliceosomal genes are also recurrently mutated in solid tumor types (Dvinge et al., 2016). The most frequently occurring spliceosomal mutations are found in four genes: splicing factor 3B subunit 1 (*SF3B1*), U2 Small Nuclear RNA Auxiliary Factor 1 (*U2AF1*), serine/arginine-rich splicing factor 2 (*SRSF2*), and zinc finger CCCH-type, RNA-binding motif and, serine/arginine-rich 2 (*ZRSR2*). While ZRSR2 mutations result in loss of function, mutations in the other spliceosomal components cause an alteration or gain of gene function (Yoshida et al., 2011). Notably, mutations in spliceosomal genes are thought to be mutually exclusive, potentially due to functional redundancy or synthetic lethality (Dvinge et al., 2016).

One of the most frequently mutated spliceosomal components is **SF3B1**, which is part of the SF3b complex and responsible for recruiting U2 snRNP by recognizing the intronic branchpoint sequence. Here, the mutations cause alterations on SF3B1 residue K700, resulting in mis-splicing of genes through selection of cryptic 3' splice sites (Wang et al., 2016). By analyzing the transcriptome of *SF3B1*-mutated tumors of chronic lymphocytic leukemia, this study revealed that cancer cells can modulate DNA damage response and Notch signaling pathways through mis-spliced variants resulting from *SF3B1* mutations. In addition, *SF3B1*-mutated patients express mis-spliced variants of the non-canonical BAF

complex subunit Bromodomain-containing protein 9 (*BRD9*) (Inoue et al., 2019). Mis-splicing results in the inclusion of a poison exon within *BRD9* transcripts, leading to degradation by nonsense-mediated decay. Depletion of BRD9 transcripts was shown to promote tumor growth (Inoue et al., 2019). Thus, mutations in spliceosomal components could not only have a bystander role in disease progression but also could be drivers of tumorigenesis.

Another example of spliceosome alterations that drive tumor growth is the occurrence of mutations at residues S34 and Q157 of **U2AF1**. These mutations are located within two zinc finger domains and affect the 3' splice site recognition of U2AF1 (Yoshida et al., 2011). Instead of interacting with the AG dinucleotide element at 3'splice sites, mutated U2AF1 recognizes alternative splice site motifs resulting in differentially spliced exons (Ilagan et al., 2015). Hence, *U2AF1* mutations can promote tumorigenesis by affecting splicing of various cellular pathways such as DNA damage and apoptosis (Ilagan et al., 2015).

Apart from *SF3B1* and *U2FA1*, ***SRSF2*** is another core spliceosomal proto-oncogene found frequently mutated in cancers, especially in hematopoietic malignancies such as chronic myelomonocytic leukemia (Yoshida et al., 2011). SRSF2 promotes splicing by binding exonic elements to recruit additional splicing factors such as U2AF1 for spliceosomal assembly (Wang and Aifantis, 2020). Canonically, SRSF2 binds the CCNG or GGNG sequence motif through its RNA recognition motif domain at similar affinities. However, recurrent mutations in P95 cause a shift in preference of SRSF2 binding towards the CCNG motif leading to specific exon inclusion or exclusion (Kim et al., 2015). This causes widespread splicing aberrancies, including mis-splicing of Enhancer of zeste homolog 2 (*EZH2*), leading to aberrant transcripts targeted for degradation. The loss of EZH2 was associated with impaired hematopoiesis, highlighting the clinical importance of *SRSF2* mutations (Kim et al., 2015). The *SRSF2* mutation-dependent changes are well-known to cause myelodysplastic syndromes; however, mutated SRSF2 was also connected to more aggressive types of blood cancer such as acute myeloid leukemia (AML). A recent transcriptome study of AML patients identified an overlap of mutations in *SRSF2* and isocitrate dehydrogenase (NADP(+)) 2 (*IDH2*), which together contribute to AML pathogenesis by affecting epigenetic processes as well as RNA splicing (Yoshimi et al., 2019). This highlights how the interplay between spliceosomal mutations and other genetic alterations can drive tumorigenesis.

In contrast to the proto-oncogenic function of the other spliceosomal genes described before, ***ZRSR2*** is considered to have a tumor-suppressive role. ZRSR2 is responsible for accurate 3' splice site recognition to promote the excision of major and minor classes of introns (Shen et al., 2010). Mutations in *ZRSR2* are found on the entire length of the gene, resulting in in-frame stop codons and thus loss-of-function of the *ZRSR2* gene (Yoshida et al., 2011). Notably, these splicing defects only affect processing of U12-type introns but do not involve U2-type introns (Madan et al., 2015). Since dysregulated U12-type intron splicing has been associated with contributing to various hematopoietic malignancies (Wang and Aifantis, 2020), *ZRSR2* mutations are another example of how alterations in the core spliceosomal machinery facilitate tumorigenesis.

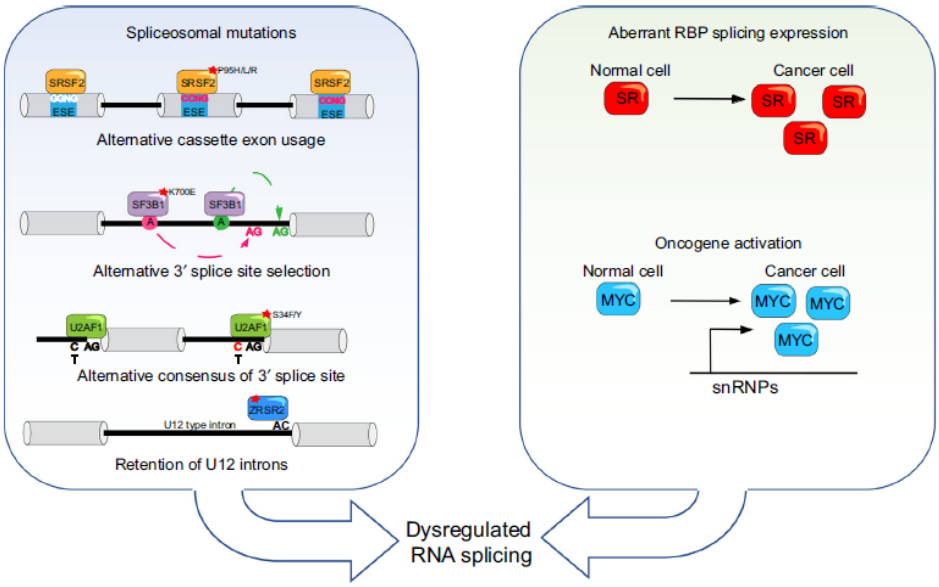


Figure 3 | Dysregulation of RNA splicing by spliceosomal mutations and aberrant expression of splicing factors and oncogenes. *Left: Core components of the spliceosome are frequently mutated in cancer and thus affect splice site recognition leading to the expression of tumorigenic isoforms. Right: Aberrant expression of RNA-binding proteins (RBPs) which modulate RNA splicing or the activation of oncogenes can introduce cancer-associated changes in RNA splicing patterns. Figure reprinted and adapted from (Wang and Aifantis, 2020) with permission from Elsevier under license number 5241391049967.*

1.2.3 Therapeutic targeting of the spliceosomal machinery

Given the critical contribution of dysregulated splicing for tumorigenesis, many therapeutic approaches have been undertaken to target spliceosomal perturbations in cancer over the past years (Figure 4). Since the spliceosomal process is essential for cell survival, but its efficiency or fidelity is reduced in various cancer types, tumor cells are more susceptible to further perturbations of the splicing machinery (Bonnal et al., 2020). This susceptibility provides a therapeutic window targeted by small-molecule spliceosomal inhibitors, which broadly prevent the assembly of the spliceosome. The first compounds identified to modulate splicing were discovered from bacterial strains in the early 1990s, but the anti-tumor activity of the so-called herboxidienes, spliceostatines, and pladienolides was described a decade later (Schneider-Poetsch et al., 2021). Although numerous compounds have been submitted to clinical trials in the past, no splicing inhibitors are in clinical use yet due to severe side effects. Initial clinical trials of inhibitors such as E7107 targeting the SF3b complex unexpectedly exhibited toxicity of bilateral optic neuritis (Hong et al., 2013). However, a more recent study in which the SF3b complex inhibitor H3B-8800 is evaluated across various leukemias reported no such toxicities (Steensma et al., 2021).

Apart from targeting the core components of the spliceosome, small molecules aim at interfering with post-translational modifications of splicing factors. For example, kinase inhibitors developed to broadly inhibit phosphorylation of SR proteins by SR protein kinases (SRPK) and CDC2-like kinase (CLK) families exhibit anti-angiogenic potential (Ohe and Hagiwara, 2015).

In addition to directly targeting spliceosomal components, pharmacological inhibitors of splicing-associated factors are tested. For example, since MYC-overexpressing tumor cells directly modulate the expression of *PRMT5* to increase the expression of spliceosomal genes, they are more susceptible to PRMT5 inhibition (Koh et al., 2015). Hence, a PRMT5 inhibitor is evaluated for its anti-tumor effect across various cancer types in clinical trials (Fedoriw et al., 2019).

Recent therapeutic approaches include targeted protein degradation of spliceosomal components by proteolysis targeting chimeras (PROTACS). For instance, sulfonamides have demonstrated an anti-tumor effect in clinical trials before, but the mechanism of action was unraveled recently. Here, a study showed that the sulfonamide compound indisulam interacts with DDB1- and CUL4-associated factor 15 (DCAF15) to promote degradation of RNA-binding protein 39 (RBM39) (Han et al., 2017). This splicing factor was described to play a central role in AML, and pharmacological degradation of RBM39

has been shown to have an antileukemic effect (Wang et al., 2019). More recently, indisulam was shown to increase anti-tumor immune response by generating cancer neoepitopes originating from aberrant splice patterns (Lu et al., 2021). While indisulam is currently undergoing clinical trials, the question remains whether PROTACs can be engineered to target other RBP splicing factors.

With the recent success of modulating splice events with antisense oligonucleotide (ASO) therapy, also approaches in cancer therapy focus on targeting specific mRNA molecules. ASOs are short single-stranded DNA sequences designed to match the target RNA sequence and can have three different mechanisms of action (Schoch and Miller, 2017). First, ASOs were designed to pair with target transcripts to trigger RNAse H-mediated mRNA degradation. This is the method of choice for diseases associated with accumulation of specific RNA/protein, such as neurodegenerative disorders. In cancer, targeting the Mushashi RNA-binding protein 1 (MSI1) with ASOs has been reported to reduce tumor growth in a patient-derived xenograft model of pancreas (Fox et al., 2016).

Vice versa, in the second strategy ASOs are used to restore protein function, which is impaired by dysfunctional splicing. Here, ASOs directly modulate specific splice events and change alternative splice patterns by binding exon-intron junctions which was shown to affect recruitment and binding of splicing factor or splice site detection (Havens and Hastings, 2016). A proof-of-concept example of this type of therapy is the FDA-approved ASO Nursinersen to treat spinal muscular atrophy. Nursinersen promotes the inclusion of exon 7 in transcripts of survival of motor neuron 2, preventing its degradation and restoring protein levels which results in a milder disease outcome (Rigo et al., 2012). Also, in cancer, ASOs have been designed to target splice events associated with anti-tumor effects. As described earlier, tumor suppressor *BRD9* is mis-spliced in cancer, leading to nonsense-mediated decay of aberrant mRNA transcripts and thus reduced protein functionality. By preventing the inclusion of the poison exon, ASOs stabilize *BRD9* transcripts, which was shown to decrease tumor growth (Inoue et al., 2019).

The third option to use ASOs includes targeting microRNAs to inhibit their function in regulating gene expression through binding mRNA targets (Schoch and Miller, 2017). ASOs are designed to bind to microRNA targets and promote their degradation to prevent their interaction with mRNA transcripts. Since microRNAs can act as oncogenes or tumor suppressors, miRNA-based therapeutic strategies are currently being evaluated for their use in cancer treatment (Shah et al., 2016).

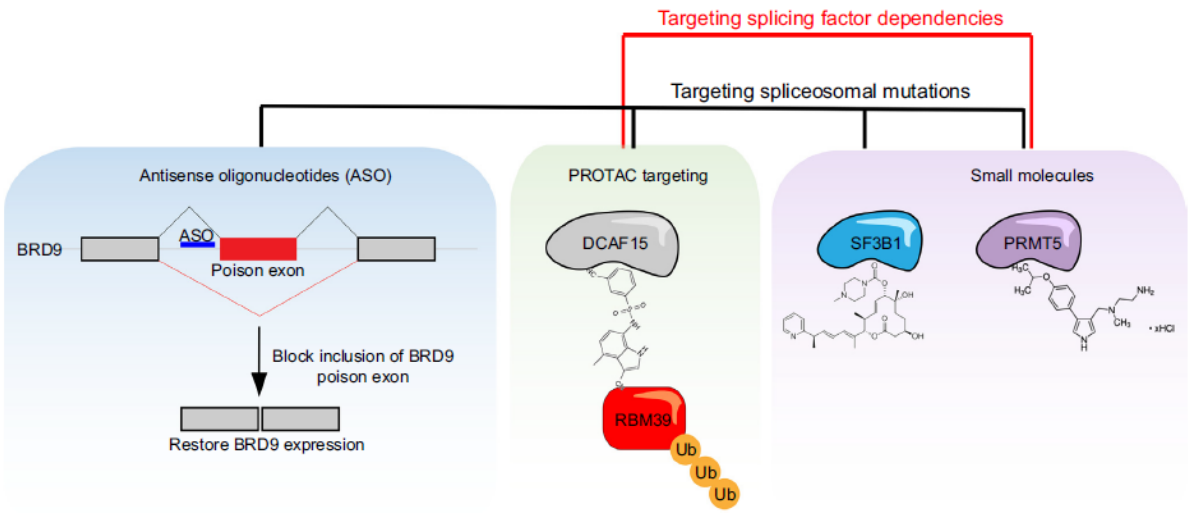


Figure 4 | Therapeutic strategies to target RNA splicing in cancer. *The antitumor effect of RNA splicing modulation has been described for several cancer types. Current clinical approaches focus on pharmacological modulation of splicing. Small molecules either directly inhibit the assembly of the core spliceosome by binding to core components such as SF3B1 or inhibit splicing-associated factors like PRMT5. In addition, proteolysis targeting chimeras (PROTACS) are developed to promote degradation of specific splicing factors. For example, degradation of RNA-binding protein 39 (RBM39) leads to antileukemic effects. In a pre-clinical approach, antisense oligonucleotides (ASOs) were able to restore the expression of the tumor suppressor Bromodomain-containing protein 9 (BRD9), which is frequently mis-spliced in cancer. Ub = Ubiquitin. Figure reprinted and adapted from (Wang and Aifantis, 2020) with permission from Elsevier under license number 5241391049967.*

Overall, several therapeutic strategies have been developed to target dysregulated splicing in cancer. Although most therapies have shown limited success in the clinic so far, recent advances in understanding the splicing mechanisms provided new exciting therapeutic approaches, such as PROTACS targeting RBP splicing factors. Importantly, tumor cells can develop resistance against splicing-based therapies and can also modulate RNA splicing to escape immune-based therapies (Wang and Aifantis, 2020). Thus, a better understanding of how tumor cells promote these resistance mechanisms will further contribute to the efficacy of cancer therapies.

1.3 The role of cancer neoantigens for immunotherapy

The recent advances in the field of immunotherapy make it the most promising treatment strategy for several different cancer types. In 2013, the *Science* magazine recognized cancer immunotherapy as 'breakthrough of the year', highlighting its potential in the fight against human malignancies (Couzin-Frankel, 2013). Furthermore, James P. Allison and Tasuku Honjo received the Nobel Prize in Physiology and Medicine in 2018 for their discovery of cancer therapy by inhibition of negative immune regulation (The Nobel Assembly at Karolinska Institutet, 2018). To date, several dozen immunotherapies have been approved for several cancer types. The success of immunotherapeutic approaches is based on their immune-modulatory functions, aiming to enhance anti-tumor immune response. Here, tumor-specific neoepitopes play a crucial role in the efficacy and safety of cancer immunotherapies. These short peptides are exclusively presented by cancer cells and thus have the ability to activate T cell-mediated cytotoxicity, thereby promoting the elimination of neoepitope-presenting tumor cells.

1.3.1 Surface presentation of neoantigens

With the advent of next-generation sequencing, researchers have discovered that the genomic instability within tumors leads to thousands of different somatic mutations. While most of them are passenger mutations and do not contribute to tumorigenesis, more than 140 genes have been identified that directly promote tumor growth and survival (Yarchoan et al., 2017a). Beyond that, nonsynonymous mutations alter the amino acid coding sequence, resulting in mutant proteins exclusively expressed in tumor cells (Figure 5). Like all other endogenously expressed proteins, mutant proteins are degraded by the proteasome and cleaved into short peptide sequences. These short peptides have a length of 8-11 amino acids and can be presented on specific human leukocyte antigen class I (HLA-I) receptors on the cell surface as snapshots of the total proteome of a cell (Gfeller et al., 2018a). Here, the peptides are recognized as antigens by $CD8^+$ cytotoxic T cells, which can discriminate between self-antigens derived from canonical, endogenous proteins, and non-self-antigens derived from mutant protein sequences (Murphy and Weaver, 2016). During their maturation, T cells are negatively selected against self-antigens, which are shared between all normal cells, to avoid autoreactive T cell response. In contrast, T cells have not encountered non-self-peptides before and can thus exhibit a strong binding affinity towards these so-called neoepitopes (Murphy and Weaver, 2016). Binding the T cell receptor (TCR) to the neoepitope activates T cell expansion and recruits other immune cells, eventually eliminating the neoepitope-presenting cell. Hence, the immune system can selectively target neoepitope-presenting cells, which makes the identification of neoepitopes of high therapeutic interest to promote immune-mediated tumor elimination (Waldman et al., 2020).

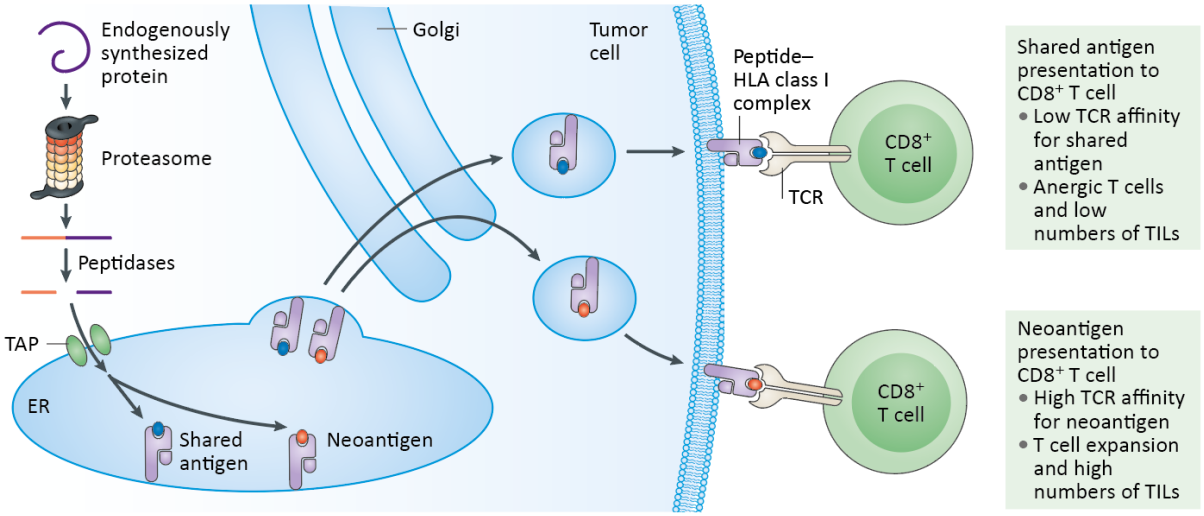


Figure 5 | Processing and presentation of antigens on HLA class I. *Endogenous peptides are labeled for proteasomal degradation by ubiquitination. The proteasome cleaves proteins into short peptide sequences, which can be further processed by peptidases resulting in short peptides of 8-11 amino acids. These peptides enter the endoplasmic reticulum (ER) through the transporter associated with antigen processing (TAP) complex. In the ER, peptides bind to human leukocyte antigen class I (HLA-I) if their sequence matches the specific binding pattern of the HLA-I subtype. The HLA-I:peptide complexes are delivered through the Golgi complex to the cell surface, where CD8⁺ cytotoxic T cells recognize presented peptides as antigens. T cell receptors (TCRs) bind with a higher affinity to neoantigens, inducing an immune response against the neoantigen-presenting cell. TIL = tumor-infiltrating lymphocytes. Reprinted and adapted from* (Yarchoan et al., 2017a) *with permission from Springer Nature under license number 5237000258875.*

1.3.2 Types of tumor antigens and their role in immunotherapy

Since neoepitopes are selectively expressed in tumors and completely absent in normal cells, they are also called **tumor-specific antigens** (TSAs). This selective expression makes TSAs an attractive target for immunotherapy. This has been already demonstrated by the infusion of CD8⁺ cells targeting mutant KRAS, which mediated effective antitumor response in patients (Tran et al., 2016). Notably, TSAs can not only arise from genetic alterations but also from oncogenic viral proteins. Viral neoantigens, such as the oncoproteins E6 and E7 of human papilloma virus (HPV), are attractive targets for immunotherapeutic approaches and were shown to facilitate immune activation in HPV-associated cancers (Aggarwal et al., 2019). However, most TSAs are derived from unique mutations specific to individual tumors. This prevents the broad application of immunotherapy because it does not only make the identification of neoantigens challenging, but the highly personalized therapy goes along with high treatment cost (Yarchoan et al., 2017a).

For these reasons, **tumor-associated antigens** (TAAs) were initially thought to be favorable targets for immunotherapy since they are shared between patients and are found in several tumor types (Frankiw et al., 2019). However, most therapeutic approaches to target TAAs remained unsuccessful. Although

TAAs are overexpressed on tumor cells, they can also be presented on normal cells at lower levels (Ward et al., 2016). Hence, immune cells are unresponsive to many TAAs due to clonal deletion of autoreactive B and T cells as a mechanism of central and peripheral tolerance (Coulie et al., 2014). Still, some TAA-specific T cells can avoid negative selection, but their TCRs usually have no high affinity for the presented TAA (Stone et al., 2015). Since TCR binding affinity is correlated with T cell cytotoxicity and activation, the immune response against TAAs might be less pronounced compared to TSAs (Tian et al., 2007).

Moreover, TAA-targeting immunotherapies can cause severe collateral damage to non-target tissues because they are also expressed on normal cells. There are several cases of autoimmune toxicities reported, for example, a therapeutic approach to target ERRB2 using chimeric antigen receptor (CAR) T cells recognizing overexpressed erb-b2 receptor tyrosine kinase 2 (ERBB2/HER2) caused respiratory failure and death in a metastatic colon patient since CAR T cells also induced an immune response against lung epithelial cells which probably also expressed ERBB2 at low levels (Morgan et al., 2010).
Similar to TAAs, **cancer/testis antigens** (CTAs) were targeted by immunotherapies. CTAs are also shared between patients, but their expression is limited to specific tumor types (Yarchoan et al., 2017a). It was thought that CTAs have a high specificity because they are selectively expressed on tumor cells but absent on normal cells, except in reproductive tissues. However, studies reported severe side effects in clinical cases using adoptive cell therapies directed against CTAs, such as melanoma-associated antigen 3 (MAGE-A3) (Cameron et al., 2013; Linette et al., 2013).

These reports highlight the importance of tumor antigens being exclusively expressed in tumor cells to avoid cross-reactivity, suggesting that TSAs are the preferred target for new immunotherapeutic approaches to avoid both immune tolerance and autoimmune toxicity.

1.3.3 Neoantigens and the immune response

With the mechanisms of antigen presentation and clonal selection of immune cells, the immune system has developed a powerful ability to identify and target non-self-antigens on cancer cells. Here, antitumor T cell response is mainly driven by the recognition of neoantigens. The central role of neoantigens was supported when researchers discovered that T cells in melanoma patients predominately respond to TSAs resulting from nonsynonymous mutations rather than to TAAs resulting from expression changes (Lennerz et al., 2005). In addition, adoptive cell therapies where antitumor lymphocytes were expanded *ex vivo* and then infused into the patient identified neoantigens as the leading cause of antitumor response (Zhou et al., 2005). Since then, several

immunotherapies have been developed that specifically target neoantigens. For example, the TCRs of CAR T cells are genetically engineered to bind a specific neoantigen of interest. However, these adoptive cell therapies are highly personalized and cost-intensive because most neoantigens arise from unique mutations and are not shared between patients (Yarchoan et al., 2017a).

Although antitumor immune response occurs naturally, tumor cells escape immune surveillance to evade immune destruction, a concept described as a hallmark of cancer (Hanahan and Weinberg, 2011). Thus, more current immunotherapeutic strategies aim to enhance the natural antitumor immune response rather than targeting a specific neoantigen. Here, immune checkpoint inhibitors demonstrated clinical efficacy and have proven that the immune system can effectively eliminate cancer cells (Figure 6A). Although all tumor antigens theoretically could contribute to antitumor T cell response, TSAs have been identified as main T cell targets in melanoma patients treated with immune checkpoint inhibitor (Gros et al., 2016).

The relevance of neoantigens for immunotherapies is further supported by the correlation between clinical response rates to immune checkpoint inhibitors and the frequency of somatic mutations (Figure 6B). Cancers formed due to certain mutagens like UV light or tobacco exhibit a high tumor mutational burden (TMB). Consequently, non-small-cell lung cancer and melanoma patients respond well to immune checkpoint blockade due to the high number of potential neoantigen targets arising from numerous somatic mutations (Yarchoan et al., 2017a). Similarly, a small subset (> 5%) of colorectal cancer patients with mismatch repair deficiency (MMR-D) shows a decent response. In contrast, the majority of colorectal cancer patients do not respond to immune checkpoint inhibition (Yarchoan et al., 2017a). Although the TMB seems to be an important clinical marker, the patient's response rate cannot be entirely explained by the somatic mutation frequency. This suggests that other factors contribute to the success of immune checkpoint therapy. In addition, the TMB varies across different tumor types and patients. Many solid tumor types, such as melanoma, carry numerous somatic mutations and are thus considered immunologically 'hot' tumors. In contrast, especially hematological malignancies exhibit a low TMB and are thus immunologically 'cold'. The different response rates of immune checkpoint inhibitors for different tumor types highlight the need for additional immunotherapeutic targets that can contribute to the antitumor immune response and/or can be specifically targeted by immunotherapeutic approaches.

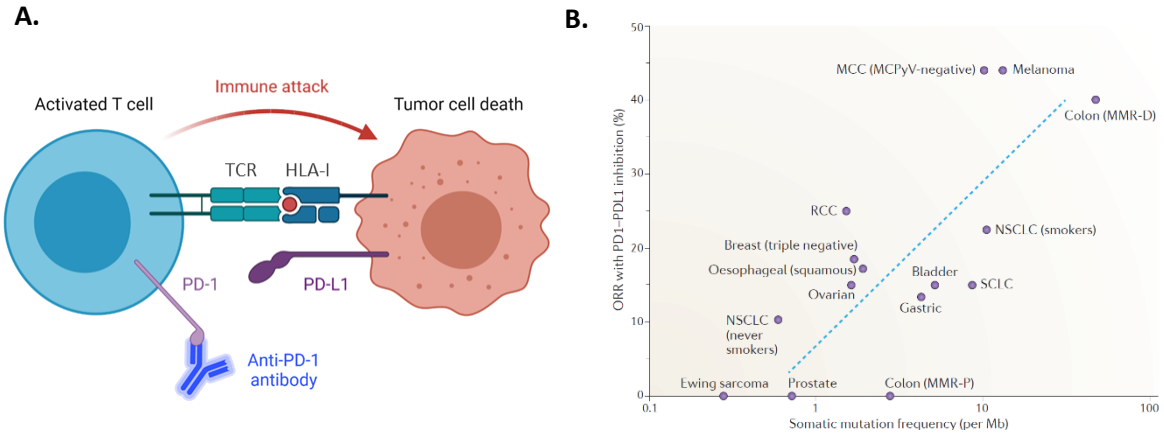


Figure 6 | Immune checkpoint inhibition enhances antitumor response in highly mutated cancers. (A) *T cells induce an antitumor immune response when their T cell receptors (TCRs) bind to non-self-antigens (neoantigens) presented by human leukocyte antigen class I (HLA-I) on the tumor cell surface. Tumors aim at escaping immune surveillance by upregulating programmed death ligand 1 (PD-L1) to inhibit T cell activation. Immune checkpoint inhibitors target this negative feedback pathway by binding to PD-L1 or programmed death protein 1 (PD1), thereby enhancing the antitumor T cell response.* ***(B)*** *Correlation of somatic mutation frequency and objective response rate (ORR) to PD1-PDL1 inhibition across various solid tumor types. The available clinical data suggest a clinical benefit of highly mutated tumors for immune checkpoint inhibition therapy, probably due to a high number of potential neoepitopes arising from nonsynonymous mutations. The line of best fit is represented as a blue dashed line on the graph. MCC = Merkel cell carcinoma, RCC = renal cell carcinoma, NSCLC = non-small-cell lung cancer, SCLC = small-cell lung cancer, MMR-D/P = mismatch repair-deficient/proficient. Figure 6B reprinted from* (Yarchoan et al., 2017a) *with permission from Springer Nature under license number 5237000258875.*

1.4 Alternative splicing-derived neoantigens

Conventionally, it is thought that neoantigens arise from aberrant protein expression resulting from nonsynonymous mutations. However, the focus on somatic mutation-derived targets often overlooks possible neoepitopes originating from mRNA processing events. An unexplored source for neoantigens is alternative pre-mRNA splicing, a widely dysregulated process in several cancer subtypes. Changes in splicing patterns promote the expression of tumor-specific splice isoforms, giving rise to potential neoepitopes with promising features (Figure 7). The recent advances in next-generation sequencing methods and bioinformatic tools enabled exploration of the transcriptome for neoepitope discovery and highlighted that investigation of the alternative splicing landscape in cancer can expand the target space for immunotherapeutic approaches.

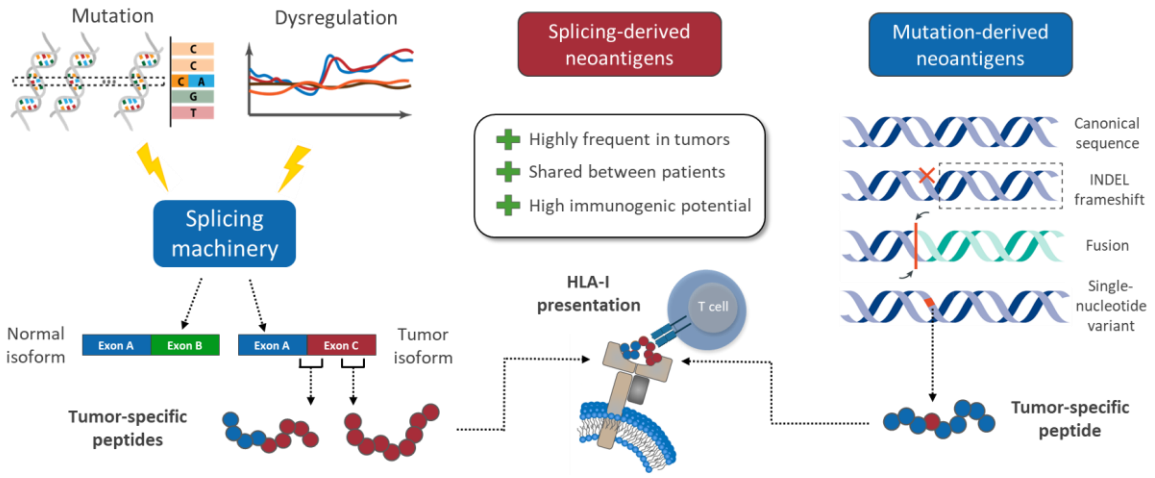


Figure 7 | Aberrant alternative splicing as a source of cancer neoepitopes. *The splicing machinery of tumors is widely affected by mutations of core spliceosomal components and dysregulated expression of splicing factors. Changes in splice junction usage generate tumor-specific transcript variants which can be translated into proteins. Proteasomal degradation of aberrant proteins results in short peptides of 8-11 amino acids that can function as neoepitopes if presented by human leukocyte antigen class I (HLA-I). These splicing-derived neoepitopes are thought to have superior features over conventional neoepitopes derived from somatic mutations and other genetic alterations. INDEL = insertion or deletion (of bases).*

1.4.1 The advantages of splicing-derived neoepitopes over conventional neoantigens

Neoepitopes resulting from dysregulated alternative splicing have been described to have distinct advantages over mutation-derived neoantigens (Smith et al., 2019). Splicing-derived neoepitopes are predicted to occur frequently across different cancers due to the wide dysregulation of alternative splicing in tumors driven by loss-of-function mutations in tumor suppressor gene-like factors and hotspot mutations in oncogene-like splicing factors (Seiler et al., 2018a). A recent study reported that at least one splicing-derived neoepitope was found in 68% of breast and ovarian cancer patients, whereas only 30% of the tumors contained a target-derived from SNVs (Kahles et al., 2018). Somatic mutations can only give rise to a limited number of novel peptide sequences, while one single splice event can give rise to a multitude of different novel peptide sequences. Cancer mutations creating tumor-specific splice junctions generate approximately twice as many neoepitopes per event as nonsynonymous mutations (Jayasinghe et al., 2018). This dramatically increases the chances that a peptide will match the binding pattern of an HLA-I allotype and can be presented on the cell surface. Moreover, peptide sequences originating from aberrant splice events can be very dissimilar from the canonical sequence. In the case of SNV, the novel peptide sequences usually only differ in one single amino acid. In contrast, alternative splicing can generate neoepitopes that are entirely different from any canonical peptide sequence. The greater dissimilarity towards self-peptides can be advantageous for the immunogenic properties of splicing-derived neoepitopes (Smith et al., 2019).

Furthermore, splicing-derived neoepitopes are predicted to be shared between patients because transcripts of splice variants are found across multiple tumors (Pellagatti et al., 2018). In contrast, conventional neoantigens result from unique mutations specific to individual tumors and thus can only be targeted by highly personalized, patient-specific therapeutic approaches, which prevents the broad application of immunotherapy (Yarchoan et al., 2017a). Hence, expanding the neoepitope repertoire by identifying splicing-derived peptides would allow a more universal immunotherapeutic approach.
In conclusion, neoepitopes originating from alternative splice events in tumors are particularly interesting targets due to their high immunogenic potential and their occurrence across multiple patients and thus have the potential to increase the target space for cancer immunotherapies.

1.4.2 Computational approaches and their limitations to identify splicing-derived neoepitopes

In 2018, the evidence was growing that peptides originating from dysregulated splice events in cancer significantly contribute to the tumor neoepitope repertoire. The first attempts to identify splicing-derived neoepitopes focused on computational strategies (Figure 8). A computational approach to identify splice-site-creating mutations across several thousand TCGA tumors predicted numerous potential neoepitopes which could arise from these events (Jayasinghe et al., 2018). Another study also aimed to identify neoepitopes from tumor-specific splice events across TCGA patients and further matched obtained neoepitope candidates to mass spectrometry data from the Clinical Proteomic Tumor Analysis Consortium (Kahles et al., 2018). Although this study already demonstrated that tumor-specific splice events can significantly increase the target space for immunotherapies, it might still underestimate the full potential of splicing-derived neoepitopes. For mass spectrometry sample preparation, trypsin was used to cleave full-length protein into small peptide fragments at arginine or lysine residues. However, these two amino acids are encoded by the conserved nucleotides at exon-exon boundaries. Thus, trypsin has a cleavage bias towards exon-exon junctions and the junction-spanning peptides are likely lost in Trypsin-digested datasets (Wang et al., 2018). Accordingly, it was reported that splice junction-spanning peptides are underrepresented in mass spectrometry datasets (Caron et al., 2015). Hence, conventional mass spectrometry datasets are not suitable for the identification of neoepitopes from tumor-specific splice junctions.

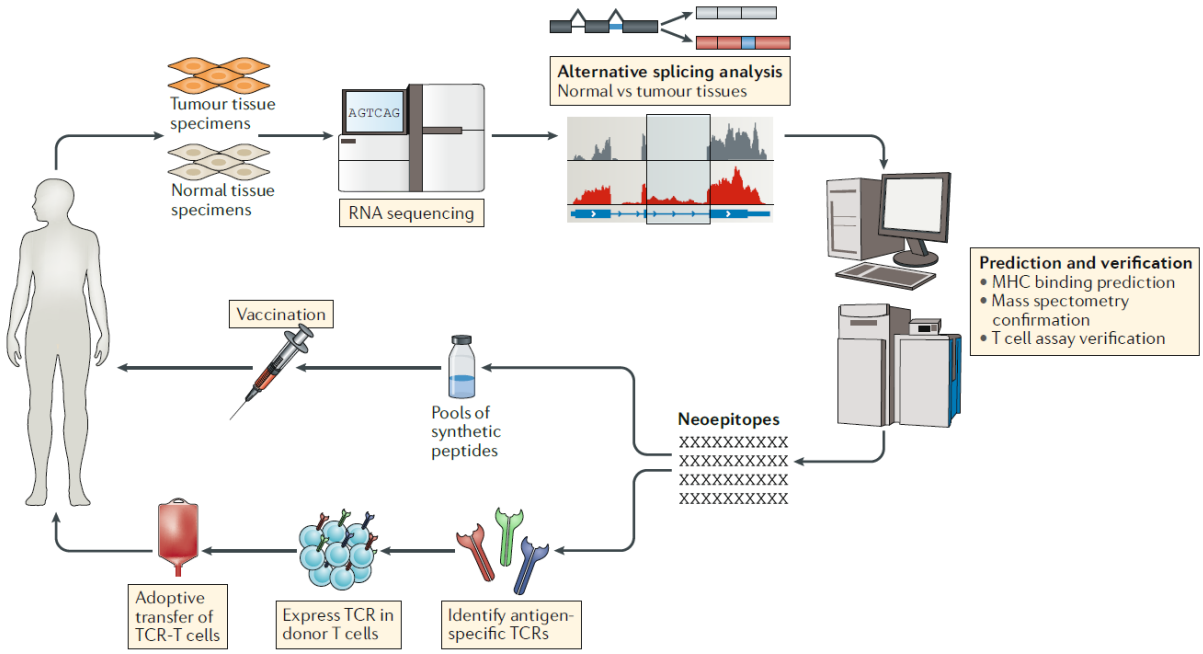


Figure 8 | Computational-based strategy for the identification of neoantigens derived from dysregulated splice events in cancer patients. *Current strategies to identify neoepitopes derived from differential splicing changes rely on computational prediction methods. These approaches start with the identification of aberrant splice events from RNA-seq data of patient tumor tissues. Candidate splice events are then translated into peptide sequences and matched to publicly available mass spectrometry datasets. TCR = T cell receptor. Reprinted from (Frankiw et al., 2019) with permission from Springer Nature under license number 5241391326960.*

1.4.3 Immunopeptidomics for discovery of splicing-derived neoepitopes

In general, solely computational approaches have only yielded a limited number of neoepitope candidates with the potential to induce an *in vivo* immune response because most of the candidates are not presented by HLA-I receptors *in vivo* (Schmidt et al., 2017). Although HLA binding prediction tools have significantly evolved over the past years, the quality of results varies depending on the HLA subtype due to limited availability of training datasets (Jurtz et al., 2017). To date, mass spectrometry is the only analytical methodology that allows the direct identification of HLA-I peptides naturally presented *in vivo* (Bassani-Sternberg and Coukos, 2016; Chong et al., 2020). Mass spectrometry-driven exploration of the immunopeptidome has a significant advantage: it avoids the caveats of error-prone computational approaches by providing direct evidence of peptide binding independent of HLA binding prediction tools. In addition, this approach addresses the limited sensitivity of mass spectrometry since it only analyzes the immunopeptidome instead the complete proteome of the cell. This also circumvents the above-described enzymatic sample digest, thereby preserving splice junction-spanning peptide sequences.

For a mass spectrometry-driven exploration of the immunopeptidome, HLA- I immunoaffinity purification (HLA-I IP) is the most suitable experimental approach to recover HLA-I peptides for mass spectrometry analysis (Bassani-Sternberg and Coukos, 2016). Here, HLA-I complexes are extracted from cell lysates and the bound peptides are subsequently purified, leading to highly enriched input samples for mass spectrometry analysis. Still, it is debated whether the sensitivity of an explorative mass spectrometry approach is sufficient to detect splicing-derived neoepitopes since most alternative splicing events are found in low abundant transcripts (Pickrell et al., 2010). However, a recent study focusing on neoepitopes derived from cancer-specific intron retention events in melanoma patients could predict several candidates that were later identified in melanoma immunopeptidomics datasets (Smart et al., 2018). More recently, another study used immunopeptidomics to identify hundreds of tumor-specific, non-canonical peptides by generating sample-specific reference databases, which facilitated the identification of non-canonical peptides from mass spectrometry datasets (Chong et al., 2020). These reports suggest that mass spectrometry-based immunopeptidomics is the most promising experimental strategy to identify neoepitopes from tumor-specific splice events.

1.5 Aim of the thesis

Cancer immunotherapy is the most promising treatment strategy for treating several different cancer types. For the efficacy and safety of immunotherapeutic approaches, the identification of tumor-specific neoepitopes plays a crucial role. However, conventional approaches have only yielded a limited amount of highly immunogenic targets. Here, the focus on somatic mutation-derived neoantigens often overlooks possible neoepitopes originating from mRNA processing events. Alternative mRNA splicing is widely dysregulated across several cancer types, generating tumor-specific splice events which can give rise to highly immunogenic peptides. These splicing-derived neoepitopes have several advantages over mutation-derived neoepitopes and can significantly increase the target space for immunotherapy.

However, the prediction of peptides originating from tumor-specific splice events is more challenging and complex than neoepitope prediction from non-synonymous mutations. Computational strategies for the discovery of splicing-derived neoepitopes have been described as error-prone and incomplete. Instead, the recent advantages in mass spectrometry-based immunopeptidomics allow direct identification of HLA-I peptides naturally presented *in vivo*. Combined with sample-specific reference databases, it enables the detection of tumor-specific, non-canonical peptides as promising targets for cancer immunotherapy.

In this thesis, I describe my efforts to establish a workflow for which I combined high throughput HLA-I IP, high-resolution mass spectrometry, and custom computational pipelines for an explorative approach to identify neoepitopes derived from aberrant alternative splice events (Figure 9).

My first aim was to establish an immunopeptidomics pipeline allowing me to explore the immunopeptidome of cell lines with different splicing burdens. This also included the generation of sample-specific reference libraries based on matching RNA-seq data. These custom databases enabled the identification of non-canonical, not-annotated peptides from raw mass spectrometry data.

My second aim was to analyze the transcriptome and identify all relevant, significantly differentially spliced events that can give rise to potential neoepitopes. I then generated sample-specific peptide libraries by translating the genetic information of identified splice events into peptide sequences.

Eventually, I wanted to combine these two efforts and match potential neoepitopes obtained from the immunopeptidomics workflow with the peptide reference library to specifically identify neoepitope candidates originating from differential splice events.

Ultimately, I aimed to explore whether I could obtain neoepitopes with highly immunogenic properties from aberrant splice events. In addition, I wanted to evaluate how many potential new targets are generated by an increase in the splicing burden and whether this could be a predictive biomarker for response to immune checkpoint inhibitor therapy.

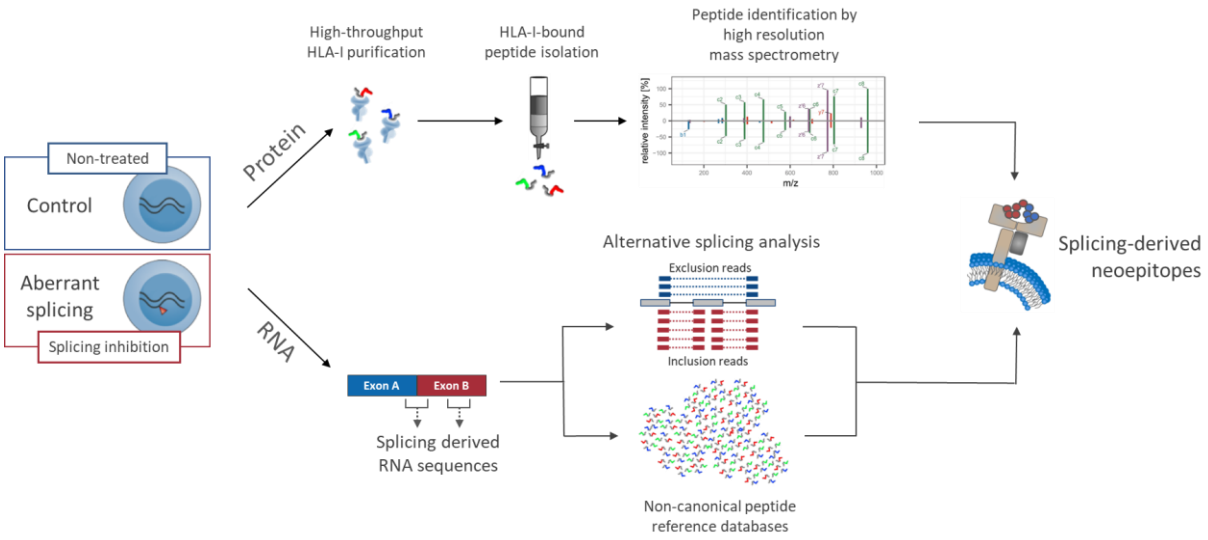


Figure 9 | Mass spectrometry-based exploration of the immunopeptidome of cell lines with different splicing burdens. *Cells were treated with a splicing inhibitor to recapitulate the spliceosomal stress that tumor cells experience. HLA-I immunoaffinity purification (HLA-I IP) recovered HLA-I-bound peptides from cell lysates. Subsequently, peptides were analyzed by mass spectrometry and their sequence detection was facilitated by sample-specific reference databases based on matching RNA-seq data. In parallel, an alternative splicing analysis was performed to detect dysregulated splice events between the conditions. By combining transcriptomic and proteomic data, splicing-derived neoepitopes were identified.*

2. Materials & Methods

2.1 Cell culture

2.1.1 Cell lines

Cell culture experiments were performed under a sterile Herasafe laminar flow hood (Thermo Fisher Scientific, Waltham, USA). K562 WT cells as well as K562 cells harboring the K700E mutation in the splicing factor gene *SF3B1* were retrieved from Queens University Belfast, United Kingdom. The K562 *U2AF1*^{S34F} cell line was generated by Team RSC at BioMed X Institute, Heidelberg. All K562 cell lines were cultivated in RPMI 1640 (ATCC modification) supplemented by 10% fetal bovine serum and 1% Gibco Penicillin-Streptomycin (Thermo Fisher Scientific, Waltham, USA).

RPE-1 WT and *TP53* knockout cell lines were obtained from Team DDC (BioMed X Institute, Heidelberg, Germany). RPE-1 *TP53*$^{-/-}$ + *MSH2*$^{-/-}$ knockout cells were generated by Team DDC. RPE-1 *TP53*$^{-/-}$ + *POLE*^{P286R} cells were generated by Nadja Ballin from Team DDC (BioMed X Institute, Heidelberg, Germany). All RPE-1 cell lines were cultivated in DMEM/F-12 GlutaMAX supplement containing 10% fetal bovine serum and 1% Gibco Penicillin-Streptomycin (Thermo Fisher Scientific, Waltham, USA).

All cell lines were cultured at 37 °C and 5% CO_2 in the Heracell VIOS 160i CO_2 incubator (Thermo Fisher Scientific, Waltham, USA). Cell growth was monitored under an inverted light microscope and cells were passaged 2-3 times a week. K562 suspension cell lines were passaged when reaching an approximate density of 1 million cells per ml. K562 cells were transferred to a 50 ml falcon, resuspended thoroughly, and counted using the Z1 Cell Counter (Beckman Coulter, Brea, USA). Approximately 100,000 cells per ml were transferred to a new culture flask containing fresh media. For passaging or harvesting adherent RPE-1 cell lines, cells were washed with pre-warmed phosphate-buffered saline (PBS) before they were detached by incubation in TrypLE Express Enzyme (1X, no phenol red; Thermo Fisher Scientific, Waltham, USA) for 5-10 min. Detached cells were resuspended in media and split at 1:10 or 1:16.

Table 1 | Human cell lines used for HLA-I IP experiments

Cell line	Source	Comment
K562 WT	Queens University Belfast	MTA #CCRCB1704
K562 *SF3B1*^{K700E}	Queens University Belfast	MTA #CCRCB1704
K562 *U2AF1*^{S34F}	Team RSC	Derived from K562 WT
RPE-1 WT	EMBL - EMBLEM	MTA Team DDC
RPE-1 *TP53*$^{-/-}$	EMBL - EMBLEM	MTA Team DDC
RPE-1 *TP53*$^{-/-}$ *MSH2*$^{-/-}$	Team DDC	Two clones: T2, T6
RPE-1 *TP53*$^{-/-}$ *POLE*^{P286R}	Team DDC	Two clones: T16, T17

2.1.2 Cell seeding concentrations

To determine cell densities suitable for HLA-I IP experiments, adherent RPE-1 cell lines were seeded in different densities (40-100,000 cells/ml) on white 96-well microplates (BRANDplates, transparent bottom, Sigma Aldrich, St. Louis, USA). 4,000-10,000 cells were seeded in 100 µl seeding volume in each well. Cell growth was monitored for 48 h using the IncuCyte S3 Live Cell Imaging and Analysis System (Sartorius, Göttingen, Germany). Cell images were analyzed using the IncuCyte software version 2019B (Sartorius, Göttingen, Germany). Following parameters were applied to the confluency mask for RPE-1 cell detection:

- Segmentation adjustment = 1.1
- Hole fill = 300 µm²
- Adjust size = -1 pixel
- Area filter = minimum 500 µm²

Growth curves were generated based on confluency values derived from image analysis. For future experiments, cell densities were chosen that guaranteed in 90-100% confluency after 24 h growth. Notably, RPE-1 cell lines could continue growing into dense clusters as there was no contact growth inhibition observed.

2.1.3 Reagents for *in vitro* cell treatment

Herboxidiene spliceosome inhibitor GEX1A was obtained from Focus Biomolecules, Plymouth Meeting, USA. The lyophilized drug was reconstituted in dimethyl sulfoxide (DMSO) (Carl Roth, Karlsruhe, Germany) to a stock concentration of 10 mM. Working dilutions were prepared freshly by diluting the stock in media.

Recombinant human IFN-γ was obtained from PeproTech, Cranbury, USA. The lyophilized IFN-γ stock was reconstituted in dH_2O to 1 mg/ml. The stock was further diluted 1:10 in PBS containing 1 % FBS to 100 µg/ml.

2.1.4 Optimization of IFN-γ-induced HLA-I expression in K562 cells

Induction of HLA-I expression of K562 cells by IFN-γ was evaluated by determining protein levels via western blot. First, K562 cells were seeded on 6-well plates. For K562 WT and *U2AF1*^{S34F} cell lines, 500,000 cells (500,000 cells/ml) were seeded per well. For K562 *SF3B1*^{K700E} cells, 600,000 cells (600,000 cells/ml) were seeded per well to compensate the slower proliferation rate of this cell line. Cells were stimulated with different amount of IFN-γ (0, 10, 50, 100 ng/ml). After 48 h, cells were harvested, and samples were prepared for western blot analysis.

For time course experiments, K562 WT cells 500,000 cells (500,000 cells/ml) were seeded in 6-well plate format and stimulated with 50 ng/ml IFN-γ along with matching non-treated control samples. Samples were harvested after 24 h and after 48 h and HLA-I protein levels were evaluated by Western blot. Moreover, the media of additional samples was replaced after 24 h. Samples were cultured in IFN-γ-free media for additional 24 and 48 h before they were harvested, and HLA-I levels were evaluated by western blot.

2.1.5 Sample preparation for flow cytometry measurement

Cell surface HLA-I expression was evaluated by flow cytometry measurements. For K562 WT and *U2AF1*^{S34F} cell lines, 500,000 cells (500,000 cells/ml) were seeded per well in a 6-well plate format. For K562 *SF3B1*^{K700E} cells, 600,000 cells (600,000 cells/ml) were seeded per well to compensate the slower proliferation rate of this cell line. K562 cell lines were stimulated with 50 ng/ml IFN-γ to induce HLA-I expression. After 24 h, K562 cells were harvested and counted. RPE-1 cell lines were detached and counted. For each sample, 250,000 cells were used as input for cell staining. One additional sample of K562 WT and RPE-1 WT cells serving as unstained controls. First, all samples were washed with 500 µl PBS and centrifuged at for 5 min at 300 x *g* at 4 °C. Supernatants were discarded and Fc Receptor Blocking Solution was added. Each sample was resuspended in 2 µl Human TruStain FcX (BioLegend, San Diego, USA) reagent diluted in 100 µl HBSS+++ (w/o calcium, magnesium, no phenol red) and incubated for 20 min at 4 °C. Next, 1.25 µl of APC anti-human HLA-A,B,C antibody was added per sample (for 250,000 cells). Again, samples were incubated for 20 min at 4°C. Subsequently, samples were centrifuged at for 5 min at 300 x *g* at 4 °C and washed with PBS. After pelleting cells again, PBS was discarded and cells were resuspended in 200 µl HBS+++ containing 1 µl SYTOX Blue Dead Cell Stain (1:2000, 0.5 µM).

Flow cytometry measurements were performed on a FACS Aria II Cell Sorter (BD Biosciences, Franklin Lakes, USA). After successful fluidics startup, a new experiment was setup. Fluorochrome compensation was not necessary since there was no overlap between APC and BV-421 spectra. Following gating strategy was applied: FSC-A/SSC-A -> FSC-W/-H -> SSC-W/-H -> fluorochromes
First, unstained controls were measured, and the gating and the voltage for individual parameters were adjusted accordingly. Next, the stained positive controls were acquired. Then, stained samples were measured, and 50,000 events were recorded per sample. Result data was exported and analyzed using the FlowJo software (BD Biosciences, Franklin Lakes, USA).

2.1.6 GEX1A dose response curves

Dose response curves were determined by viability assays in a 96-well plate format. For K562 WT and *U2AF1*^{S34F} cell lines, 5,000 cells (50,000 cells/ml) were seeded per well. For K562 *SF3B1*^{K700E} cells, 6,000 cells (60,000 cells/ml) were seeded per well to compensate the slower proliferation rate of this cell line. Different morphologies between RPE-1 cell lines resulted in specific seeding concentrations: WT & *TP53*$^{-/-}$ = 17.500 cells/ml; *TP53*$^{-/-}$*+MSH2*$^{-/-}$ & *TP53*$^{-/-}$*+POLE*^{P286R} clone #17 = 15.000 cells/ml; *TP53*$^{-/-}$*+POLE*^{P286R} clone #16 = 10.000 cells/ml.

To achieve spliceosomal inhibition, 50 µl GEX1A dilution was added to the cell seeding volume of 100 µl. GEX1A stock was diluted to 750 nM (1.5x) and six 1:1 dilutions were prepared in a serial dilution. 50 µl of each dilution were added to the plates in quadruplicates, resulting in a total volume of 150 µl per well. Non-treated control cells received additional 50 µl of fresh media. After 48 h, viability assays were performed following the instructions of the CellTiter-Glo 2.0 Assay protocol (Promega, Madison, USA). First, a 1:1 mix of media and CellTiter-Glo 2.0 reagent was prepared. Cell supernatant was aspirated and 100 µl of the mix were added per well. Plates were incubated for cell lyses shaking for 5 min at RT. Subsequently, luminescence was recorded at 485 nm_{Ex}/520 nm_{Em} using the GloMax Microplate Reader (Promega, Madison, USA). After subtraction of the background signal, the average value of non-treated cells was determined. Next, individual values were divided by the average control value to obtain the ratio of cells alive in treated conditions compared to the non-treated control. The results were further analyzed using GraphPad Prism (v8.43, Graphpad Software Inc., San Diego, USA). Values were normalized and a nonlinear fit curve was applied to the data to determine GR_{50} values. Final values were plotted in a dose response curve from which GEX1A doses for HLA-I IP experiments were determined.

2.1.7 Sample preparation for HLA-I immunoaffinity purification (HLA-I IP)

For HLA-I IP experiments, cells were cultured in bulk to obtain around 100 million cells per sample. K562 suspension cells were expanded in T175 flasks (Figure 10). For K562 WT and *U2AF1*^{S34F} cell lines, 50 million cells (250,000 cells/ml) were seeded per flask. For K562 *SF3B1*^{K700E} cells, 60 million cells (300,000 cells/ml) were seeded per flask to compensate the slower proliferation rate of this cell line. Cells were added to a total culture volume of 200 ml and stimulated with 50 ng/ml IFN-γ to induce HLA-I expression. After 24 h, half of the samples were treated with 10 nM GEX1A. The other half served as non-treated controls. After 24 h of GEX1A treatment and a total culture time of 48 h, cells were harvested. K562 cells were resuspended thoroughly to prevent cell clumping. Cells were distributed to 4x 50 ml falcons chilled on ice. Aliquots of 1 ml of each sample were taken for Western blot analysis.

The aliquots were further processed as described in Section 2.4.1. For replicate #3 additional aliquots of 7.5 ml were taken as input for RNA-seq analysis which were processed as described in Section 2.4.1. The remaining cell suspension samples (50 ml falcons) were centrifuged for 5 min at 300 x *g* at 4 °C. Supernatants were aspirated and the four cell pellets of each sample were pooled by resuspending them in a total of 15 ml ice-cold PBS. Each cell suspension was transferred into a new 15 ml falcon and centrifuged for 5 min at 300 x *g* at 4 °C. Supernatants were aspirated, cell pellets were snap-frozen in liquid nitrogen and stored at -80 °C.

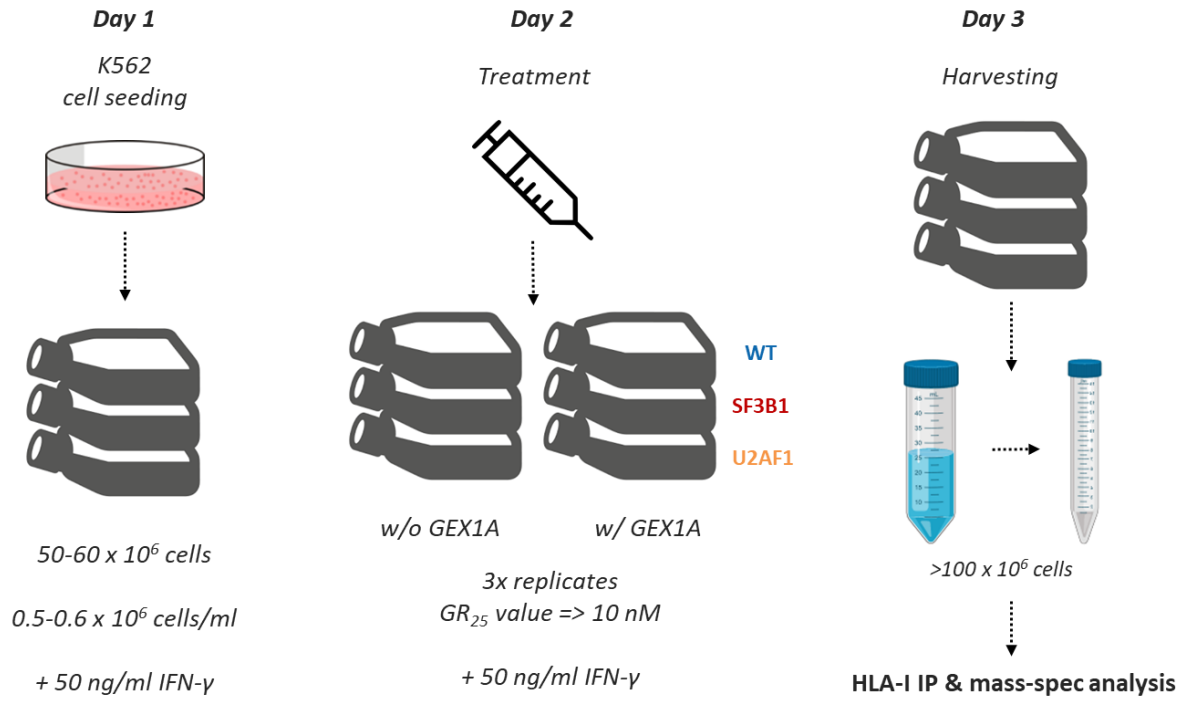


Figure 10 | Cell culture workflow for generating K562 samples as input for HLA-I IP.

Adherent RPE-1 cells were expanded in 5-layer Rectangular Straight Neck Cell Culture Multi-Flasks (Thermo Fisher Scientific, Waltham, USA). For RPE-1 WT and *TP53*$^{-/-}$ cell lines 5 million cells were seeded per flask and for *TP53*$^{-/-}$*+MSH2*$^{-/-}$ as well as *TP53*$^{-/-}$*+POLE*^{P286R} cell lines 4 million cells were seeded per flask (Figure 11). Notably, for the tetraploid *TP53*$^{-/-}$*+POLE*^{P286R} clone T16 only 3 million cells were seeded per flask. 10 ml of cell suspension containing the respective cell numbers were added to the 5-layer flasks and 200 ml of DMEM/F-12 medium was added. Cells were expanded for 72 h at 37 °C and 5% CO_2 inside a cell incubator. Then, fresh media was prepared containing 63 nM GEX1A and added to the flask. The control flask received normal media without drug. After 24 h of treatment, cells were harvested. The media was discarded, and flasks were rinsed with 50-100 ml PBS. Cells were detached using 32 ml enzyme-free cell dissociation buffer. After 10-15 min incubation at 37 °C and 5% CO_2 inside a cell incubator, 18 ml medium were added and the whole suspension was transferred into

a T75 flask. 5-layer flasks were rinsed with additional 50 ml PBS and the suspension was transferred into the same T75 flask. All cell suspensions were resuspended thoroughly to prevent cell clumps and then transferred into 2x 50 ml falcons and chilled on ice.

Aliquots of 2 ml of each sample were taken for Western blot analysis. The aliquots were centrifuged for 5 min at 300 x *g* at 4 °C, supernatants were aspirated, and the pellets washed with ice-cold PBS. After pelleting the cells again and aspirating supernatants, pellets were snap-frozen in liquid nitrogen and stored at -80 °C. For all replicates, additional aliquots of 8 ml were taken as input for RNA-seq analysis which were processed as described in Section 2.4.1. The remaining cell suspension samples (50 ml falcons) were centrifuged for 5 min at 300 x *g* at 4 °C. Supernatants were aspirated and pellets were resuspended in 7.5 ml ice-cold PBS. Since samples were split between two falcons, both cell suspensions of paired falcons were pooled by transferring cell suspensions into a new 15 ml falcon. Eventually, falcons were centrifuged for 5 min at 300 x *g* at 4 °C, supernatants were aspirated, and cell pellets were snap-frozen in liquid nitrogen and stored at -80 °C.

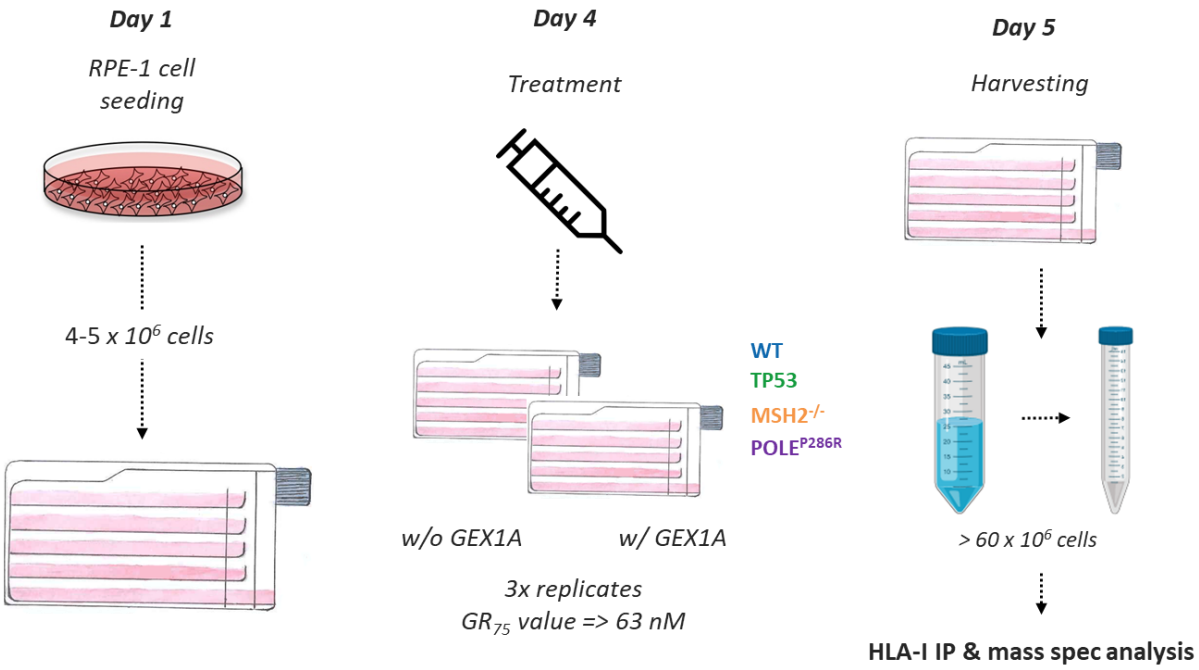


Figure 11 | Cell culture workflow for generating RPE-1 samples as input for HLA-I IP.

2.1.8 Additional K562 samples for RNA-seq

Since RNA-seq aliquots were only processed for replicate #3 of samples prepared for HLA-I IP, additional samples were generated to serve as input for RNA-seq. This follow-up experiment was designed to reproduce the experimental conditions of the bulk experiment in a 6-well plate format. For K562 WT and *U2AF1*^{S34F} cell lines, 500,000 cells (500,000 cells/ml) were seeded per well. For K562 *SF3B1*^{K700E} cells, 600,000 cells (600,000 cells/ml) were seeded per well to compensate the slower proliferation rate of this cell line. Cells were stimulated with 50 ng/ml IFN-γ. After 24 h, half of the samples were treated with GEX1A while the other half served as non-treated controls. After 24 h of treatment, cells were harvested, lysed in RNAzol for subsequent RNA extraction. In addition, 1 ml aliquot was taken for western blot analysis.

2.2 HLA-I immunopurification (HLA-I IP)

2.2.1 Antibody purification

Supernatants of HB-95 hybridoma cells producing anti-HLA-I antibody W6/32 were obtained from Jonas Becker. W6/32 antibody was purified from the supernatants using Sepharose A beads in gravity flow columns as described previously (Bassani-Sternberg, 2018). First, empty Poly-Prep columns (Bio-Rad, Hercules, USA) were rinsed with 9 ml of 1% SDS solution followed by 40 ml of dH_2O. Next, 4 ml of 50% Sepharose-Protein A beads solution were added to the column and the preservation buffer was allowed to drain. The beads were washed once using 9 ml of 100 mM Tris-HCl (pH 8.0) buffer. After preparation of the columns, column tips were closed, and cell culture supernatant was added. The supernatant was allowed to flow through the column enabling binding of W6/32 antibodies to Sepharose Protein A beads. Subsequently, beads were washed once using 9 ml of 100 mM Tris-HCl (pH 8.0) buffer and then a second time using 9 ml of 20 mM Tris-HCl (pH 8.0). Then, W6/32 antibodies were eluted with 8x 1 ml of 0.1 N acetic acid (pH 3.0) from the beads into eight individual tubes containing 300 µl of 1 M Tris-HCl (pH 8.0) to neutralize the solution. Tubes were gently vortexed, and the IgG concentration of each fraction was measured using a NanoDrop 2000/2000c spectral photometer (Thermo Fisher Scientific, Waltham, USA). High IgG-containing fractions were pooled, and aliquots were prepared containing 10 mg of W6/32 antibody per tube. These tubes were subsequently used for antibody cross-linking procedures. Low antibody-containing fractions were pooled and used as primary antibodies for western blot analysis. Purified W6/32 antibody aliquots were stored at -20 °C.

2.2.2 Antibody cross-linking

W6/32 antibodies were covalently coupled to Sepharose-Protein A beads for HLA-I IP experiments. Cross-linking was supposed to reduce co-elution and potential clogging of columns during the immunopurification process. The cross-linking with dimethyl pimelimidate (DMP; Thermo Fisher Scientific, Waltham, USA) was performed in gravity flow columns following the instructions of a previously published study (Bassani-Sternberg, 2018). First, empty Poly-Prep columns (Bio-Rad, Hercules, USA) were rinsed with 9 ml of 1% SDS solution followed by 40 ml of dH_2O. Next, 4 ml of 50% Sepharose-Protein A beads solution were added to the column and the preservation buffer was allowed to drain. The beads were washed once using 9 ml of 100 mM Tris-HCl (pH 8.0) buffer. After preparation of the column, the column tip was closed, and 10 mg W6/32 antibody diluted in 5 ml of 100 mM Tris-HCl (pH 8.0) buffer were added. Beads were resuspended in antibody solution and the mixture was incubated by slowly rotating the column at 8 rpm for 30 min at RT. Subsequently, the buffer was allowed to drain from the column and an aliquot of the flowthrough was collected and its protein concentration determined a NanoDrop 2000/2000c spectral photometer (Thermo Fisher Scientific, Waltham, USA) to ensure W6/32 antibody coupling to the beads. Next, beads were washed with 0.2 M sodium borate buffer (pH 9.0) to remove residual amines from the Tris buffer before the column tip was closed again. The beads were resuspended in 2 ml of a freshly prepared 40 mM DMP solution, resulting in a final DMP concentration of 20 mM. This mixture incubated by slowly rotating the column at 8 rpm for 30 min at RT. After cross-linking the antibody to the beads, unreacted DMP was quenched by washing the beads with 5 ml of 0.2 M ethanolamine (pH 8.0). The column tip was closed, another 5 ml of of 0.2 M ethanolamine (pH 8.0) were added and remaining unreacted DMP was further quenched by slowly rotating the column at 8 rpm for 2 h at RT. Then, beads were washed with 9 ml of PBS containing 0.02% sodium azide. Finally, beads were covered with 2 ml of PBS containing 0.02% sodium azide and stored at 4 °C until further usage.

2.2.3 High-throughput immunopurification of HLA-I:peptide complexes

The high-throughput HLA-I IP was performed following the instructions for the immunopeptidomics workflow described in a previous study (Chong et al., 2018). Lysates from frozen pellets were directly prepared before the HLA-I IP was performed. 1 ml of HLA lysis buffer (Table 2) was added to frozen pellets and samples were thawed on ice. After 15-20 min, additional 1 ml lysis buffer was added, cells were resuspended and transferred into fresh 2 ml tubes. The tubes were incubated for 45 min at 4 °C before the lysates were cleared by centrifugation at 21,000 x *g* for 30 min at 4 °C.

For the high-throughput immunopurification of HLA-I-bound peptides from the samples, an empty 96-well filter microplate with 3 µm glass fiber and 10 µm polypropylene membranes (Agilent, Santa Clara, USA) was loaded with W6/32-coupled Sepharose-Protein A beads. First, the filter plate was washed with 1 ml of 100% acetonitrile (I; Sigma Aldrich, St. Louis, USA) per well, followed by 1 ml of 0.1% trifluoracetic acid (TFA; Sigma Aldrich, St. Louis, USA) per well. Solutions were prepared by diluting reagents in Optima LC/MS Grade Water (Thermo Fisher Scientific, Waltham, USA) and washes were performed by applying 3-5 psi using the Positive Pressure-96 Processor (Waters, Milford, USA). Then, the filter plate was equilibrated by washing the columns (psi = 3-5) with 2 ml of 0.1 M Tris-HCl (pH 8.0) per well. The cross-linked Sepharose Protein A-W6/32 beads were resuspended using low-retention ART filter tips and 150 µl were loaded onto each column. Beads were slowly washed (psi = 1-2) by adding 1 ml 0.1 M Tris HCL (pH 8.0). Next, beads were conditioned with 400 µl HLA lysis buffer per well. After preparation of the filter plate, the cleared cell lysates were loaded onto the columns and allowed to flow through by gravity at 4 °C. The flow-through was collected in a special 96-deep well collection plate (Agilent, Santa Clara, USA). Then, columns were washed (psi = 3) several with buffers containing different salt concentrations:

1. 8x 1 ml/well 150 mM sodium chloride in 20 mM Tris-HCl (pH 8.0)
2. 8x 1 ml/well 400 mM sodium chloride in 20 mM Tris-HCl (pH 8.0)
3. 8x 1 ml/well 150 mM sodium chloride in 20 mM Tris-HCl (pH 8.0)
4. 4x 1 ml/well 20 mM Tris-HCl (pH 8.0).

For the elution of HLA-I:peptide complexes, a Sep-Pak tC18 96-well plate with 100 mg sorbent per well and 37-55 µm particle size (Waters, Milford, USA) was prepared. Columns were conditioned by washing (psi = 3) with 1 ml 80% ACN in 0.1% TFA per well and then with 2 ml 0.1 % TFA per well. Then, the filter plate was stacked on top of the Sep-Pak tC18 96-well plate and HLA-I: peptide complexes were eluted from the filter plate by slowly washing (psi = 1-2) 2x with 500 µl 1% TFA. The complexes were captured by the Sep-Pak tC18 96-well plate which was washed (psi = 2-3) 2x with 1 ml 0.1% TFA per well. For final elution of HLA-I-bound peptides, the plate was stacked on top of a 96-deep well collection plate and eluted 2x with 400 µl 28% ACN in 0.1% TFA. Eluted peptides were transferred into 1.5 ml tubes, dried by vacuum centrifugation, and stored at -80 °C until mass spectrometry analysis.

Table 2 | Composition of the HLA-I lysis buffer (adapted from Chong et al., 2018).

Reagent	Supplier	Stock	MW [Da]	Solubility	Final concentration	Dilution
Sodium deoxycholate	Thermo Fisher Scientific, Waltham, USA	5 g	414.6	50 mg/ml in H2O	0.25 % / 0.125 mg/ml	1:20
Iodoacetamide	Sigma Aldrich, St. Louis, USA	5 g	184.96	10-100 mM in H2O	0.2 mM	1:100
Ethylenediamine-tetraacetic acid (EDTA)	Sigma Aldrich, St. Louis, USA	0.5 M	292.24	-	1 mM	1:50
Octyl-β-D-glucopyranoside	Sigma Aldrich, St. Louis, USA	250 mg	292.37	100 mg/ml in H2O	1.0 % / 1 mg/ml	1:10
Halt Protease and Phosphatase Inhibitr Cocktail	Thermo Fisher Scientific, Waltham, USA	-		-	1x	1:100

2.3 Biochemical and molecular biology methods

2.3.1 Western blot analysis

2.3.1.1 Cell lysis & sample preparation

Cell suspensions were centrifuged for 5 min at 300 x *g* at 4 °C, supernatants were aspirated, and the pellets washed with ice-cold PBS. Cells were pelleted again, and supernatants were aspirated before pellets were snap-frozen in liquid nitrogen and stored at -80 °C.

Cells lysis was accomplished by adding HLA-I lysis buffer (Table 2) which stabilized HLA-I molecules (Chong et al., 2018) in a concentration of 100 µl per 1 million cells. Lysis buffer was always prepared fresh from reagent stocks. During cell lysis, samples were kept under constant agitation (1000 rpm) for 30-60 min at 4 °C. Lysed samples were centrifuged at 14,000 x *g* for 20 min at 4 °C. Samples were chilled on ice and 90% of the supernatant was transferred into a fresh tube. Protein content was determined by BCA assay, or samples were directly prepared for SDS-PAGE.

2.3.1.2 Bicinchoninic acid (BCA) assay

Protein concentrations of samples were determined following the user guide instructions of the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, Waltham, USA). A serial dilution of bovine serum albumin (BSA) in HLA-I lysis buffer was used as standard curve. Samples were pre-diluted 1:5 in HLA-I lysis buffer to fit in the dynamic range of the assay. 25 µl of sample or standard was to 96-wells in duplicates. BCA working reagent solution as prepared by mixing Reagent A with B in a 50:1 mix. 200 µl working reagent solution was added to each well and the plate was incubated for 30 min at 37 °C. The plate was brought back to RT and absorbance was measured at 562 nm using the GloMax Microplate Reader (Promega, Madison, USA). Background absorbance was subtracted from the results and the sample protein concentrations were determined based on absorbance in reference to the BSA standard curve.

2.3.1.3 Protein gel electrophoresis & Western blot

For protein separation by gel electrophoresis, the NuPAGE system and reagents were used (Thermo Fisher Scientific, Waltham, USA). Samples were denatured by adding 4x NuPAGE LDS Sample Buffer + 10x NuPAGE Sample Reducing Agent containing 500 mM dithiothreitol (DTT) and boiled for 10 min at 70 °C. The NuPAGE system was used to separate proteins by molecular mass. The gel chamber was filled with NuPAGE MES SDS Running Buffer (20x) diluted 1:20 in dH_2O, and 500 µl NuPAGE Antioxidant was added. Then, samples containing 10-40 µg protein were loaded on NuPAGE 4-12% Bis-Tris gels (1.5 mm x 10/15 wells). In addition, 5 µl of PageRuler Plus Prestained Protein Ladder (10 to 250 kDa) was loaded as reference. Proteins were separated at constant 150 V for 90 min.

After successful protein separation, the gel was dismounted and stored in running buffer. For transferring proteins onto a membrane, the user guide instruction for the Bio-Rad Trans-Blot Turbo Transfer System (Bio-Rad, Hercules, USA) were followed. Trans-Blot Turbo Transfer Pack containing a pre-activated PVDF membrane (0.2 µm) and the gel were mounted onto the cassette. Proteins were transferred from the gel to the membrane using the program 'High MW' (1.3 A constant current, up to 25 V) of the Trans-Blot Turbo Transfer device. After the run, the PVDF membrane was stored in TBS-T.

2.3.1.4 Membrane detection

PVDF membranes were blocked using 5% (w/v) milk powder (Carl Roth, Karlsruhe, Germany) in TBS-T for 1 h at RT to prevent unspecific binding of antibodies. After rinsing the membranes with TBS-T, membranes were cut around the 70 kDa marker and specific proteins were probed by incubation with

primary antibodies. The upper part of the membrane containing proteins >70 kDa was incubated with rabbit polyclonal anti-Vinculin antibody (Cell Signaling Technology, Danvers, USA) diluted 1:2000 in 1% BSA (w/v) in TBS-T. The bottom part was incubated with mouse monoclonal anti-HLA-ABC antibody (W6/32) diluted 1:2000 in 1% BSA (w/v) in TBS-T. After overnight incubation, membranes were washed three times with TBS-T for 10 min and incubated with secondary antibody (anti-mouse or anti-rabbit, see Table 10) diluted 1:5000 in 1% (w/v) milk powder solution for 1 h. Unbound antibodies were washed off by rinsing the membranes three times for 5 min with TBS-T. For enhanced chemiluminescence (ECL) detection via horseradish peroxidase (HRP) SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Fisher Scientific, Waltham, USA) was used in a 1:1 mix of Solution A and B. The signal was detected by the FUSION FX Imaging System (Vilber Lourmat, Eberhardzell, Germany) using the High Sensitivity program.

2.3.2 HLA-I haplotyping

gDNA were isolated from cells following the user guide of the GeneJET Genomic DNA Purification Kit (Thermo Fisher Scientific, Waltham, USA). Sample concentrations of isolated gDNA were determined by the NanoPhotometer NP80 (Implen, Munich, Germany). Aliquots with 25 ng/µl with a total volume of 120 µl were prepared and sent for sequencing by DKMS Life Science Lab GmbH (Dresden, Germany) to determine the specific HLA-I allele types of each cell line.

2.3.3 RT-PCR validation of splice events

Computationally predicted differential splice events identified from RNA-seq data were experimentally validated by reverse transcription polymerase chain reaction (RT-PCR).

RPE-1 cells were seeded in 10 cm culture dishes so that they reached 95% confluency on the following day of treatment (RPE-1 WT = 1.9 M; *TP53*$^{-/-}$ cells 1.75 M cells). Cells were cultured at 37 °C and 5% CO_2 inside a cell incubator and treated with different concentrations of GEX1A (0, 10, 30, 60 nM). After 24 h of treatment, cells were lysed by adding 2 ml of RNAzol RT (Sigma Aldrich, St. Louis, USA) to each plate and stored at -80 °C. RNA was isolated from the samples as described in Section 2.4.2. RNA concentrations were determined by measuring the absorbance of samples at 260 nm and 280 nm using the NanoPhotometer NP80 (Implen, Munich, Germany). Samples were diluted to a target concentration of 150 ng/µl and 0.75 µg RNA was used as input for cDNA synthesis.

Reverse transcription of sample RNA into cDNA was performed by following the protocol for the ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs, Ipswich, USA). First, primer mix was added to the RNA sample and the sample mixes were denatured at 65 °C for 5 min. Then,

Protoscript II reaction mix as well as RNase inhibitor were added. After 1 h incubation of the samples at 42 °C, the enzyme was heat inactivated at 80 °C for 5 min. cDNA samples were stored at -20 °C.

cDNA samples were used as input for RT-PCR. The presence of candidate splice event was evaluated by amplifying the region of interest resulting in PCR products with different lengths. For exon skipping events, forward primers were designed binding the upstream exon and reverse primers binding the downstream exon. In case of full intron retention, primers were designed to bind the upstream and downstream exon. All primers fulfilled following properties: length of 20-24 bases, 40-60% G/C content, 1-2 G/C pairs at start and end, and a similar melting temperature (T_m) of 58-60 °C.
Splice events region were amplified following the protocol for the protocol for OneTaq Quick-Load 2X Master Mix with Standard Buffer (New England Biolabs, Ipswich, USA). As input for RT-PCR, 75 ng or 100 ng of sample cDNA were used and amplified using the Thermocycler (Bio-Rad, Hercules, USA).

Table 3 | Thermocycling conditions for amplification of splice event regions.

Step	Temperature	Time
Initial denaturation	94 °C	30 sec
30 cycles	94 °C	20 sec
	54 °C (optimized for all primers)	30 sec
	68 °C	1 minute per kb
	68 °C	5 min
Hold	4 °C	∞

2.4 RNA sequencing & data processing

2.4.1 Sample preparation

Cell suspensions were centrifuged for 5 min at 300 x *g* at 4 °C, supernatants were aspirated, and pellets were resuspended in 1 ml of RNAzol RT (Sigma Aldrich, St. Louis, USA) for cell lysis. After 20 min incubation at RT, samples were stored at -20 °C.

2.4.2 RNA isolation

RNA was isolated from lysed samples following the user manual for RNAzol RT RNA Isolation Reagent (Sigma-Aldrich, St. Louis, USA) for sample preparation and isolation of total RNA. For homogenization of samples, 0.4 ml of water per 1 ml RNAzol was added and samples were shaken for 15 sec before

incubation for 15 min at RT. DNA, protein, and polysaccharides were precipitated by centrifugation of samples at 12,000 x *g* for 15 min at RT. 1 ml of the supernatant was transferred to a 2 ml tube and RNA was precipitated by adding 1 ml of isopropanol per 1 ml supernatant. After storing samples for 10 min at RT, they were centrifuged at 12,000 x *g* for 10 min at RT. The supernatant was decanted, and the RNA pellet was washed 2x by adding 75% ethanol (v/v) and centrifugation at 8,000 x *g* for 3 min at RT. The ethanol was carefully removed, and the RNA pellets were resuspended in nuclease-free water (New England Biolabs, Ipswich, USA). Samples were stored at -80 °C.

2.4.3 RNA concentrations & quality measurements

As first estimation, RNA concentrations were determined by measuring the absorbance of samples at 260 nm and 280 nm using the NanoPhotometer NP80 (Implen, Munich, Germany). Based on these results, samples were diluted, and precise RNA concentrations were determined following the instructions of the Qubit RNA BR Assay Kit (Thermo Fisher Scientific, Waltham, USA). In brief, the Qubit working solution was prepared by diluting the RNA BR Reagent 1:200 in RNA BR Buffer. 198 µl of working solution were added to sample tubes, 190 µl to standard tubes. 10 µl of each standard A and B were added to respective standard tubes and 2 µl of diluted RNA samples to sample tubes. Tubes were briefly vortexed and incubated for 2 min at RT before they were measured by the Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, USA). Standard curve was determined by acquiring Qubit standards. Then, samples tubes were acquired to determine the RNA content. Final RNA concentrations were calculated based on sample dilution factors.

Besides, RNA quality was measured using the 2100 Bioanalyzer (Agilent, Santa Clara, USA) following the instructions of the RNA 6000 Nano Kit Guide (Agilent, Santa Clara, USA). First, the electrodes were decontaminated with 350 µl RNaseZAP (Thermo Fisher Scientific, Waltham, USA) and cleaned with 350 µl of nuclease-free water (New England Biolabs, Ipswich, USA). For the preparation of the gel, 550 µl of Agilent RNA 6000 Nano gel matrix were loaded into the top receptacle of a spin filter and centrifuged at 1500 x *g* for 10 min. Next, 1 µl of RNA 6000 Nano dye concentrate was added to a 65 µl aliquot of filtered gel. After vortexing, the gel-dye mix was centrifuged at 13,000 x *g* for 10 min. 9 µl of the mix was added at the specific well bottom of an RNA chip placed on the chip priming station. The chip was pressurized by using the syringe to cast the gel. After loading the gel-dye mix, 5 µl RNA 6000 Nano Marker was loaded into each sample well and in the ladder well. Then, the 1 µl of previously diluted samples were loaded. In addition, 1 µl of RNA ladder was added to the ladder well. After vortexing the chip, it was inserted in the Agilent 2100 Bioanalyzer. RNA quality was measured using the EukaryoteTotal RNA Nano setup of the 2100 Expert Software (Agilent, Santa Clara, USA).

2.4.4 RNA sample preparation for RNA-seq

For each sample, 3000 ng RNA was prepared in a 100 ng/µl dilution in nuclease-free 1.5 ml tubes based on concentrations measured with QUBIT. RNA spike-ins were added as control probes. Iso Mix E2 of the SIRV Set 1 (Lexogen, Vienna, Austria) was resuspended 1:10 in 10 µl nuclease-free water and then diluted 1:10 in nuclease free water, obtaining a 1:100 dilution (0.303 ng/µl). For each sample containing 3000 ng RNA, 2.97 µl of diluted spike-ins were added to each sample (Table 4). Final RNA concentrations of the sample were measured by QUBIT. Samples were stored at -80 °C before sample submission for sequencing.

Table 4 | Calculations for RNA spike-ins added to RNA-seq samples.

Formulas							
m SIRV =	0.01 (F SIRV reads) x 0.03 (F target RNA) x X ng (m RNA input)						
V SIRV =	30 pg (m SIRV) / 25.2 pg/µl (C SIRV)						
Parameters					**Calculation**		
m RNA input	total RNA conc.	3000	ng		**m SIRV**	0.9	
F target RNA	mRNA fraction	0.03	= 3%		**V SIRV**	2.97	
F SIRV reads	SIRV fracton of mRNA	0.01	= 1%				
C SIRV 1:10	*Stock resuspended in 10 µl*	3.03	ng/µl				
C SIRV 1:100	*1:10 of resuspended*	0.303	ng/µl		**--> Prepare 40 µl of a 1:100 dilution**		
C SIRV 1:1000	*1:10 of 1:100*	0.0303	ng/µl		**for sufficient spike-in for 12 samples**		

2.4.5 RNA sequencing

The generated K562 libraries were sequenced pair-end with read length of 75 bp using a HiSeq2000 V4 (Illumina, San Diego, USA) at the Genomics Core Facility at European Molecular Biology Laboratory (EMBL, Heidelberg, Germany). RPE-1 libraries were shipped to Novogene (Cambridge, UK) and sequenced pair-end with read length of 150 bp using a NovaSeq 6000 system (Illumina, San Diego, USA).

RNA-seq data summarized in SRA files were transformed to fastq format in bash using the command: fasterq -dump SRRxxxxxx -e 10 (e = number of threads). Adapter trimming for each sample pair was performed using the bash command: trim_galore --illumina --fastqc --paired $NAME1 $NAME2 (NAME = file name).

2.4.6 *STAR* alignment

RNA-seq reads were aligned to the human genome using the *Spliced Transcripts Alignment to a Reference* (STAR) software (v2.7.0; Dobin et al., 2013). As reference, the human genome release 32 (GRCh38.p13) was used containing the comprehensive gene annotation originally created on the GRCh38 reference chromosomes, mapped to the GRCh37 primary assembly. Reference annotation files were obtained from GENCODE (Frankish et al., 2019). *STAR* alignment was performed for paired-end reads executing the following the bash scripts in the '01_RNAseq_analysis' folder on the RSC GitHub repository:

1. Reads were aligned to transcriptome (1st pass)
   Script: 01A_STAR_GRCh37_1st-pass_paired-end_ACK.sh
2. Reference genome was indexed
   Script: 01B_IndexingGenome2ndpass.sh
3. Reads were aligned to indexed transcriptome and quantified
   Script: 01C_STAR_GRCh37_2nd-pass_paired-end.ACK.sh

Importantly, overhangs were adjusted according to read length (K562 = 74 nt, RPE-1 = 149 nt).

2.4.7 Gene expression analysis

Gene expression analysis based on quantified reads was performed using the *DESeq2* package (Love et al., 2014). The whole workflow is summarized in the R script '02_Differential_gene_expression_analysis.R' available in the '01_RNAseq_analysis' folder on the RSC GitHub repository. In brief, 'ReadsPerGene.out.tab' files obtained from *STAR* alignment were imported into R Studio and a count matrix was generated. In addition, an annotation data frame was generated to include information about the different sample groups. Count matrix and annotation data frame served as input for gene expression analysis. Gene identifiers and features were obtained from the *Ensembl* repository using the *biomaRt* package (Durinck et al., 2009). Based on *DESeq2* results, samples were clustered based on gene expression in a principal component analysis (PCA) using the *factoextra* package (Kassambara and Mundt, 2020). Gene set enrichment analysis was performed using the *clusterProfiler* package (Yu et al., 2012). To show enrichment of specific gene sets in the treated condition, gene set enrichment (GSEA) plots were produced using the *enrichmentplot* package (Yu et al., 2012). The *pathfindR* package (Ulgen et al., 2019) was utilized to highlight enrichment of genes of specific pathways.

2.5 Mass spectrometry

2.5.1 Mass spectrometry data acquisition

Lyophilized samples were reconstituted in 10 µl of 0.1% formic acid and 9 µl were injected as input for tandem mass spectrometry analysis. The liquid chromatography-mass spectrometry directly combined with a mass spectrometry (LC-MS/MS) was performed on an UltiMate 3000 RSLCnano System (Thermo Fisher Scientific, Waltham, USA) directly coupled to an Orbitrap Fusion Lumos (Thermo Fisher Scientific, Waltham, USA). MS analysis was accomplished following the instructions for higher-energy collisional dissociation (HCD) measurements described in a previous study (Becker et al., 2021). Sample measurements were executed by the EMBL Proteomics Core Facility, Heidelberg, Germany.

2.5.2 Peptide sequence validation

To confirm peptide sequences of neoepitope candidates, mass spectra of experimentally determined peptide sequences were compared to mass spectra of synthetic peptide standards. Synthetic peptides were ordered as SpikeTides serving as reference peptides for targeted proteomics (JPT Peptide Technologies, Berlin, Germany). Lyophilized synthetic peptides were delivered as pool in a single cryo vial containing 19 nmol of each of the 24 candidate peptides. For mass spectrometry data acquisition, lyophilized peptide mix were diluted to a concentration of 100 fmol/µl in 0.1 % LC-MS grade formic acid and 3 µl were used per injection. The same acquisition parameters were used as applied before for acquisition of the experimental datasets.

After mass spectrometry data acquisition, raw data was analyzed using PEAKS Studio X applying the same parameters as used for experimental data described in Section 2.7.1. The data for matched fragment ions ('PSM-ions.txt') for synthetic standards as well as for spectra from previous experimental runs were exported and imported into R Studio. All spectra of experimental replicates of a candidate peptide were compared to all spectra from the synthetic standards control by following the R Script 'Peptide_sequence_validation.R' adapted from Daniel Zolg (Zolg et al., 2017) and Jonas Becker (Becker et al., 2021). Next, experimental mass spectra were compared to all spectra from their synthetic counterpart. To validate spectra similarity between biological and synthetic measurements, Pearson's correlation coefficients were determined for every pair of matched fragment ions. Besides, the peptide retention time differences were calculated. The intensities of the spectra pair with the highest correlation were normalized and visualized using the *ggplot2* package (Wickham, 2016).

2.6 Generation of custom reference peptide libraries

For the detection of non-canonical splice junction-derived peptides, custom reference libraries were generated based on RNA-seq information to complement peptide information found in UniProt. All scripts are available in the '02_Custom_ref_DBs' folder on the RSC GitHub repository: 'https://github.com/MauerLab'.

2.6.1 Splice Junction-derived Database (SJDB)

The SJDB contained all possible short peptides sequences originating from splice junctions annotated by *STAR* alignment. Splice junction regions were extended to capture surrounding nucleotide sequence and was converted into bed12 files executing the bash script 'SJDB_01_convertSJ2junctionregion.sh'. In brief, coordinates of splice junctions annotated in 'SJ.out.tab' files generated by *STAR* were extended for each region of interest. Coordinates were extended 29 nucleotides upstream of start coordinates and 29 nucleotides downstream of end coordinates. Length of 29 nucleotides was chosen since most HLA-I binders have a length between 8-10 amino acids/24-30 nucleotides (Gfeller et al., 2018b) and at least one nucleotide was required to be junction-spanning. Based on the extended coordinates, nucleotide sequences for regions surrounding splice junctions were retrieved using the 'getfasta' command from *BEDTools* (Quinlan and Hall, 2010). All splice junction information was summarized in one bed12 file per sample. The files were filtered based on read counts applying the bash script 'SJDB_02_filter_bed12_5reads.sh. Only junctions were selected that had robust read evidence (>= 5 reads).

Next, the genomic information was translated into peptide sequences using the R script 'SJDB_03_SJDB_generation.R'. The nucleotide sequences were split into individual characters and then translated into all three reading frames using the R package *SeqinR* (Charif and Lobry, 2007). Sequences after stop codons and peptide sequences < 8 aa were eliminated. Duplicates were removed because an accumulation of similar peptide sequences in the reference library would affect peptide identification during *PEAKS* analysis. All peptide sequences originating from splice junction across all samples were saved in one master FASTA file that was imported as new reference library in *PEAKS Studio X*.

2.6.2 TransDecoder reference library

The workflow for this *de novo* proteome assembly was adapted from *ProteomeGenerator* (Cifani et al., 2018). First, a sample-specific transcriptome was assembled using *StringTie* (Pertea et al., 2015). Sorted bam files from previous *STAR* alignment served as input for the bash script 'TransDecoder_01_stringtie_generate.sh'. The threshold of the minimum read coverage allowed for predicted transcript was varied (-c 1/5) as well as the length of the side bases of junction-spanning reads (-a 1/5/10). Resulting gtf files of individual samples were combined by running the bash script 'TransDecoder_02_stringtie_merge.sh' applying default parameters. Similar to *STAR* alignment, human genome release 32 (GRCh38.p13) obtained from GENCODE was used as reference.

Based on the results from *StringTie*, a *de novo* proteome was generated using *TransDecoder* (Haas, 2018). The instructions are summarized in 'TransDecoder_03_TransDecoder_workflow.docx' The transcript structure GTF file was converted to an alignment GFF3-formatted file using the 'gtf_to_alignment_gff3.pl' command in bash. Then, a transcript FASTA file was generated based on the reference genome and the merged gtf file from *StringTie* using the 'gtf_genome_to_cdna_fasta.pl' command in bash. Then, 'TransDecoder.LongOrfs' was executed to predict coding regions from the transcript FASTA file. As an additional step, homology searches as open reading frame (ORF) retention criteria were included. A reference proteome containing SwissProt and TrEMBL protein sequences was obtained from UniProt (Bateman et al., 2021). This proteome was indexed and the ORF sequences were searched against this reference using *BLASTp* (Camacho et al., 2009). Only peptide sequences from the longest ORFs were kept. Besides, additional sequences were included by predicting sequences likely originating from coding regions using the bash command 'TransDecoder.Predict'. The results were saved as FASTA and formatted to serve as reference library for peptide identification in *PEAKS Studio X*.

2.7 Peptide identification from mass spectrometry data

All scripts are available in the '03_Peptide_identification' folder on the RSC GitHub repository: 'https://github.com/MauerLab'.

2.7.1 Peptide identification by *PEAKS*

Peptide sequence identification from mass spectrometry data was performed using *PEAKS Studio X* (v10.0; Bioinformatics Solutions, Waterloo, Canada). Default refinement options were applied to raw mass spectrometry data. Peptide identification was achieved by *de novo* sequence-assisted database search selecting the 'HCD fragmentation' method and setting thresholds for precursor mass error tolerance to 10 ppm and fragment mass error tolerance to 0.02 Da. The spectra were searched against the UniProt database (20,659 entries, February 2019) or against custom reference libraries. To include peptides with posttranslational modifications, terminal acetylation (+42.01 Da), oxidation of methionine (+15.99 Da), and carbamidomethylation of cysteine (+57.02 Da) were selected as variable modification options. Since protein samples were not enzymatically digested before, 'no enzyme' was selected for enzyme specificity. In addition, contaminations were filtered out by searching spectra against a database containing sequences of common contaminants. '*De novo* spectra' exhibiting a local confidence score > 50% and were not detected in the initial search against UniProt were used as input for multi-round searches against custom reference libraries.

Before data export, results were filtered by contaminants and decoy sequences. Additionally, only sequences with a length between 8-15 amino acids were kept. Identified peptide sequences were then exported with an FDR cut-off of 5%. Results were later also filtered in R Studio for peptide sequences passing an FDR cut-off of 1%.

PEAKS analysis was performed with support from Jonas Becker at the Molecular Medicine Partnership Unit of the Heidelberg University Hospital and European Molecular Biology Laboratory (EMBL, Heidelberg, Germany).

2.7.2 PEAKS data processing

PEAKS data was loaded into R Studio and modified as described in the R script '01_PEAKS_data_processing.R'. Column names were adjusted to the respective sample names and only peptides sequences with 8-12 amino acids were kept since this is the expected length of HLA-I binders (Gfeller et al., 2018b). The processed files were merged to a single file containing all peptides identified from the different reference libraries. Duplicates and peptides without any accession were removed. Peptides with post-translational modifications were only kept if the peptide did not occur in the unmodified version. After these filtering steps, the results were saved as csv file. In addition, the

data was subset to keep only peptide information of a specific cell line (e.g. WT samples). Besides, peptides were filtered based on FDR cut-offs (1% or 5%).

2.7.3 Quality control of immunopeptidomics data

The processed data was subjected to different quality control measures commonly used in the field of immunopeptidomics.

2.7.3.1 Peptide hydrophobicity indices

Peptide hydrophobicity is an established and orthogonal parameter for validation of peptide identifications (Rolfs et al., 2019). Sequence-specific hydrophobicity indices of peptides are supposed to correlate with their experimentally observed retention times (Krokhin et al., 2006). Hydrophobicity indices of predicted peptides were calculated using the web-based *Sequence Specific Retention Calculator* (*SSRCalc* version Q; Krokhin et al., 2006). Processed data from *PEAKS* output was adapted to the input format of the *SSRCalc* online tool and saved as FASTA file as described in the R script '02A_Hydropobicity_vs_rentention_time.R'. *SSRCalc* was executed with following parameters: 100 Å C18 column, 0.1% formic acid (2015 Model), none label deltas, free cysteine (no protection). The results were imported, combined, and visualized in a scatter plot using *ggplot2* (Wickham, 2016). Pearson's correlation coefficient between hydrophobicity indices and retention times was calculated to validate peptide identification.

2.7.3.2 *GibbsCluster* analysis of peptide sequences

Shared peptide sequence motifs characteristic for binding to HLA-I subtypes were detected by k-means clustering of peptide sequences using the web tool *GibbsCluster* (v2.0; Andreatta et al., 2017) with default parameters for HLA-I binders. Processed *PEAKS* data was adapted to the input format of the online tool using the R script '02B_GibbsCluster_analysis.R'. In the *GibbsCluster* online tool, numbers of clusters were set to 1-5, motif length to 9, penalty factor for inter-cluster similarity to 0.8, and weight on small clusters to 5. Pre-calculated amino acids frequencies from UniProt were used as background. The resulting clusters were visualized by sequence logos generated using the web tool *Seq2Logo* (v2.0; Thomsen and Nielsen, 2012). The Kullback-Leibler logo type was selected in which amino acids are represented as their probability, enriched amino acids as positive, and depleted amino acids as negative. The sequence motifs were compared to computational predicted motifs for the respective HLA-I allele subtype obtained from the *NetMHCpan* website.

2.7.3.3 HLA-I binding prediction

The quality of identified peptides was further determined by performing an HLA-I binding prediction using the command line version of *NetMHCpan* (v4.0; Jurtz et al., 2017). The workflow is summarized in the R script '02C_HLA_binding_prediction.R. First, a FASTA file of identified peptide sequences was generated. In addition, a text file with HLA-I allele information was created. HLA binding prediction was performed by executing the bash script '02C_Run_netMHCpan.sh'. *NetMHCpan* was executed with standard parameters considering 8-12mer peptides as potential binders and the rank threshold for strong binders was set to < 0.5% and for weak binders to < 2%. The results were saved as xls files and imported into R Studio. In case several versions of the same peptide were identified as binders of the same HLA-I allele, only the best-ranked peptide sequence with its corresponding binding affinity in nM was kept. Likewise, in case peptide sequences were predicted to bind multiple HLA-I alleles, the peptide was matched to the allele with the highest binding affinity. The results of the HLA-I binding prediction were visualized in a stacked bar plot using *ggplot2* (Wickham, 2016).

In addition, peptide-binding affinities were ranked for specific HLA-I subtypes and visualized. The data for the binding prediction towards the specific HLA-I subtype were selected and peptide entries were ranked based on their binding affinity values (K_D in nM) and then plotted on a log scale using *ggplot2* (Wickham, 2016).

2.7.4 Immunopeptidome overlap analysis

Numbers of shared peptides between replicates were determined by comparing identified peptides of each sample group following the R script '03A_Peptide_overlap_between_replicates.R. A binary data frame was generated for qualitative analysis, transforming the values for mass spectra areas into 1. In case no peptide was detected, the non-existing value was transformed into 0. Then, a sample list was created for each sample group. The total number of peptides for the sample group and each replicate were determined. Peptide overlap between replicates was visualized in Venn diagrams using the R packages *gplots* (v3.0.3, Warnes et al., 2020). In addition, heatmaps were generated to visualize the sample distance using the *ComplexHeatmap* package (v1.20.0; Gu et al., 2016). Replicates were assigned to their respective sample group and peptides found across all replicates were excluded (background subtraction). Next, size factors for each sample were generated to normalize column entries by column sum, i.e. total number of peptides per replicate. Normalized values were used to visualize sample distances in a heatmap.

In addition, the immunopeptidome overlap between whole sample groups was determined following a strategy similar to the one explained above ('03B_Peptide_overlap_between_conditions.R').

Moreover, the overlap of identified peptides between the K562 and the RPE-1 datasets was determined using the R script '03C_Peptide_overlap_between_datasets.R'.

2.7.5 Gene expression of identified peptides

To evaluate whether identified peptides originated from highly expressed genes, peptide information was combined with gene expression data following the R script '04_Expression_of_peptide_genes.R'. First, gene names of identified peptides were obtained from *biomaRt* search (Durinck et al., 2009). In addition, the normalized gene counts from *STAR* alignment were imported into R Studio and also subjected to *biomaRt* search to translate Ensembl identifiers into gene names. Only protein-coding genes were selected. Peptide and gene expression information was merged and the expression of peptide-matched genes and genes without peptide evidence was visualized in histograms using *ggplot2* (Wickham, 2016).

2.8 Differential alternative splicing analysis

All scripts are available in the '04_Splicing_analysis' folder on the RSC GitHub repository: 'https://github.com/MauerLab'.

2.8.1 *Salmon* analysis

Salmon (v1.6.0; Patro et al., 2017) was used as a tool to quantify the expression of transcripts/isoforms from RNA-seq data. As reference, a decoy-indexed transcriptome was generated based on the human genome release 32 (GRCh38.p13) obtained from GENCODE by executing the bash script '03A_Generate_salmon_index.sh'. Reads from fastq files were mapped to this reference by running Salmon ('03B_Run_Salmon_GRCh37_gencode_paired-end.sh'). The 'quant.sf' files were assigned to respective samples and then imported into R Studio using the *tximport* package (Soneson et al., 2016). Based on the Salmon results, sample groups were clustered by PCA using the *factoextra* package (Kassambara and Mundt, 2020) to visualize group differences. The workflow is summarized in the R script '03C_Salmon_results_PCA.R' available in the '01_RNAseq_analysis' folder on the RSC GitHub repository.

2.8.2 *rMATS* analysis

For the identification of the different types of alternative splice events from RNA-seq data, *replicate Multivariate Analysis of Transcript Splicing* (rMATS, v4.0.2) was performed (Shen et al., 2014).

Reference text files were created to be able to compare fastq files of different sample groups. *rMATS* analysis was carried out in *Python* (van Rossum, 2015). As genomic reference, the human genome release 32 (GRCh38.p13) from GENCODE was used, which was previously indexed by *STAR* alignment to include the annotation of novel splice junction (ACK_index). The executed bash commands can be found in the bash script 'A01_python_run_rMATS.sh'.

After performing *rMATS* analysis, the JCEC results files were imported into R Studio including both junction-spanning counts (JC) and exon counts (EC) as final output. Splice events with an FDR value of ≥ 0.05 and inclusion level difference of ≥ 0.1 were excluded from the results. Numbers of significantly differentially spliced events were determined for individual splice event type and the results were visualized as box plots using *ggplot2* (Wickham, 2016). Instructions for data visualization are summarized in the R script 'A02_rMATS_analysis.R'.

2.8.3 *JunctionSeq* analysis

Differential alternative splice events were identified by determining differential usage of exons and splice junctions between the sample groups. This exon/junction expression analysis from RNA-seq data was achieved by following the instruction of the *JunctionSeq* pipeline (Hartley and Mullikin, 2016). First, raw counts were generated from RNA-seq data via *QoRTs* (Hartley and Mullikin, 2015). Annotation files were created for samples groups and the sorted bam files were processed executing the bash script 'B01.1_bam_processing.sh. The *--stranded* option was applied since the RNA-seq data was strand-specific. The original GTF annotation file human genome release 32 (GRCh38.p13) from GENCODE was used as genomic reference.

After read alignment, results were loaded into R Studio and library size factors were extracted. Next, gene and (novel) splice junction counts were generated via *QoRTs* by executing the bash script 'B01.2_create_flat_annotation_with_novel_junctions.sh'. Here, also read counts for novel splice junctions were produced but low-coverage junctions (< 5 reads) were filtered out by executing the command *–minCount 5*. The resulting GFF flat annotation file and the count tables, along with the sample decoder files, were loaded into R Studio. *JunctionSeq* analysis was performed to detect differentially expressed splice events between the sample groups. The whole workflow is summarized in the R script 'B01.3_JunctionSeq_analysis.R.

The significantly differentially expressed splice events were loaded into R Studio and the number and type of splice events were determined. Splice events were considered significant if they passed

following threshold: FDR ≤ 0.05 and logFC ≥ 2 The results of the *JunctionSeq* analysis were visualized in a volcano plot using *ggplot2* (Wickham, 2016).

2.8.4 *DJExpress* analysis

The *DJExpress* workflow was used as an alternative method to determine differential splice junction expression between sample groups (Paez & Mauer, 2022). The SJ.out.tab files from previous *STAR* alignment were loaded into R Studio. The human genome release 32 (GRCh38.p13) reference file from GENCODE was modified by annotating novel splice junctions based on read alignment information. Next, samples were prepared for analysis by defining the control samples and filtering splice junctions based on read coverage (> 5 reads). Junctions were considered significantly expressed when they passed the cut-offs: FDR ≤ 0.05 and logFC ≥ 2. The whole workflow is summarized in the R script 'C01_DJExpress_analysis.R'.

2.8.5 Alternative splicing peptide libraries

Results from *DJExpress* analysis were loaded into R Studio and peptide sequences were retrieved from genomic coordinates of significantly altered splice events. The result files were transformed into bed file format and the splice events were classified into upregulated and downregulated junctions. The genomic coordinates of upregulated splice events were each extended by 29 nucleotides since most HLA-I binders have a length between 8-10 amino acids/24-30 nucleotides (Gfeller et al., 2018b) and at least one nucleotides was supposed to be junction-spanning. After adapting the coordinates, one bed file was saved containing the updated start coordinates and the other containing the updated end coordinates. For both files genomic sequences were retrieved by *bedtools* (Quinlan and Hall, 2010) producing two sequences for each splice event. Obtained sequences were merged by matching their feature identifiers.

The obtained bed files containing the combined genomic sequence information of differential splice events were then translated into peptide sequences. Genomic sequences were split into individual characters and then translated into all three reading frames using the R package *SeqinR* (Charif and Lobry, 2007). Sequences after stop codons and peptide sequences < 8 amino acids were eliminated. The data frame was prepared to serve as input for matching neoepitopes to differential splice events. The whole workflow is summarized in the R script 'C03_DJExpress_peptide_library.R'.

2.8.6 Sashimi plots

The command line tool *rmats2sashimiplot* (Xie et al., 2015) was utilized to visualize read coverage of candidate splice events based on sorted bam files obtained from *STAR* alignment. To compare splice events between sample groups, the read information of all replicates was combined. Sashimi plots were produced by providing genomic coordinates of a splice event (instead of rMATS output files) and a GFF annotation file matching of the human genome release 32 (GRCh38.p13) which was previously used for *STAR* alignment. Splice events were displayed in Sashimi plots when they passed the filter for read coverage (> 5 reads). The bash command script 'Sashimi_plots_RPE1.WT_grouped.sh' is available in the folder '06_Candidate_validation' on the RSC GitHub repository.

2.9 Identification of splicing-derived neoepitopes

All scripts are available in the '05_Neoepitope_identification' folder on the RSC GitHub repository: 'https://github.com/MauerLab'.

2.9.1 Neoepitope identification

Peptides were considered neoepitopes when they were exclusively and robustly detected in samples treated with splicing inhibitor GEX1A. To identify peptides specific for each sample group, a binary data frame was generated, transforming the values for mass spectra areas into 1. In case no peptide was detected, the non-existing value was transformed into 0. Then, a sample list was created containing the two sample groups of interest. Immunopeptidomes of the two sample groups were compared and only peptides were selected that exclusively appeared in the GEX1A-treated condition. Of these GEX1A-specific peptides, only those were kept, which were detected in at least 2 out of 3 replicates. The remaining peptides were saved as csv for downstream analysis. The workflow for neoepitope selection from peptide data is explained in the R script '01_Neoepitope_identification_from_PEAKS.R'.

2.9.2 Splicing-derived neoepitope identification

After neoepitope identification, peptide candidates were compared against the alternative splicing peptide libraries to determine which neoepitopes could be matched to aberrant splice events. Peptide libraries generated in Section 2.8.5 and previously identified neoepitope candidates from Section 2.9.1 were imported into R Studio. Here, neoepitope candidates were matched against all possible peptide sequences derived from differential splice events. The information about successfully matched neoepitope candidates was saved as csv files. The workflow is described in the R script '02B_Find_neoepitopes_in_DJExpress_results.R'. Subsequently, all matched neoepitope candidates

identified from different sample groups were merged using the R script '03B_Merge_all_DJExpress_neoepitopes.R'.

In addition, a GSEA was performed to determine whether neoepitopes originate from differential gene expression due to GEX1A treatment following the R script '04_Gene_expression_of_neoepitopes.R'. Gene identifiers of neoepitopes were obtained from the *Ensembl* repository using the *biomaRt* package (Durinck et al., 2009). The peptide information was combined with gene expression data of protein-coding genes obtained from *DESeq2* analysis (Section 2.4.7). Gene set enrichment analysis was performed using the *clusterProfiler* package (Yu et al., 2012) and visualized in enrichment plots.

2.10 Immunogenicity assays of neoepitope candidates

Immunogenicity assays were primarily performed by Alejandro Hernandez-Sanchez at the Applied Tumor Biology Department at Heidelberg University Hospital's Institute of Pathology.

2.10.1 Mouse model

HLA-A2.1/HLA-DR1-transgenic H-2 class I-/class II-knockout mice (A2.DR1 mice; Pajot et al., 2004) were provided by the Institute Pasteur (Paris, France). The mice were hosted by the Animal Care Facilities at German Cancer Research Center (DKFZ, Heidelberg, Germany) and received a standard chow diet and water ad libitum. For the experiments, age-matched groups of male mice were used.

2.10.2 *In vivo* immunization

Peptides for *in vivo* immunization experiments were synthesized by the 'GMP and T Cell Therapy Unit' at the German Cancer Research Center (DKFZ, Heidelberg, Germany) following GMP standards. HPLC-purified peptides were reconstituted in DMSO, obtaining a target concentration of 20 µg/µl. A2.DR1 mice were immunized to a total of 100 µg peptide. Candidate peptides were vaccinated in a group of two with a 50 µg injection amount for each candidate peptide (Table 5). 100 µg of viral *HPV16* epitope E7 11-19 peptide served as positive control. Peptide pools or single peptides were mixed with 20 µg of the adjuvant CpG ODN 1826 (TIB Molbiol, Berlin, Germany) reconstituted in sterile PBS. Mice were injected intradermally three times with peptide mixes at seven-day intervals. Seven days after the last immunization, mice were euthanized by CO_2 inhalation.

Table 5 | Vaccination groups for immunization experiments of A2.DR1 mice.

Peptide	Function	Origin	Group (n = 6 mice)
SLAKALYEA	candidate	ANXA1	2
KLNEYLLQY	candidate	PSMA4	2
ILDNGEWTV	candidate	CCNB1	3
GMIIGPPRV	candidate	UBE2V2	3
KILDNGEWTV	candidate	CCNB1	4
GMLPDPKNV	neg. control	ANT1	4
YMLDLQPET	pos. control	HPV16-E7	1

2.10.3 Splenocyte isolation

Mice were dissected under aseptic conditions and the spleens were transferred to sterile ice-cold PBS supplemented with 10% FBS. The tissue was disrupted by passing through a 40 µm cell strainer to obtain a cell suspension. Cell suspensions were pelleted by gentle centrifugation at 300 x *g* for 5 min at 4 °C. Supernatants were discarded and pellets were resuspended in 1 ml of 1X Red Blood Cell (RBC) lysis buffer (Sigma-Aldrich; St. Louis, USA). Samples were incubated for 5 min at RT to eliminate erythrocytes before 9 ml Gibco RPMI 1640 [+] L-Glutamine supplemented with 10% FBS and 1% Penicillin-Streptomycin were added. Cells were pelleted by gentle centrifugation at 300 x *g* for 5 min at 4 °C. The cell pellet was resuspended in 1 ml Gibco RPMI 1640 [+] L-Glutamine supplemented with 10% FBS and 1% Penicillin-Streptomycin and kept on ice. Cell suspensions were counted and adjusted to $2x10^6$ cells/ml. Isolated splenocytes were either directly used in ELIspot assays or were co-cultured in the presence of candidate peptides at 37 °C, 5% CO_2 to serve as input for T cell killing assays.

2.10.4 *Ex vivo* IFN-γ ELISpot

IFN-γ ELISpot assays were performed *ex vivo* using isolated splenocytes. The assays were performed with support from Alejandro Hernandez-Sanchez of the Applied Tumor Biology at the University Hospital Heidelberg, Germany.

First, MultiScreen$_{HTS}$-IP plates (Merck, Darmstadt, Germany) were activated with 70 µl of 70% ethanol per well for 5 min and washed five times with sterile PBS. The plates were coated with 100 µl purified rat anti-mouse IFN-γ antibody (1:200 in PBS; BD Biosciences, Franklin Lakes, USA) and incubated overnight at 4 °C. The next day, coated plates were washed four times with PBS and blocked with 200 µl RPMI Medium 1640 supplemented with 10% FBS and 1% Gibco Penicillin-Streptomycin for 1 h at 37

°C. Next, 2 µg of a specific peptide were diluted in 50 µl assay medium and directly added to respective wells. To each well $2x10^6$ splenocytes in 50 µl assay medium were added and incubated for 16-20 h at 37 °C. After incubation, the splenocyte-peptide mix was removed, and the plates were washed five times with PBS. Next, the plates were coated with 100 µl biotinylated rat anti-mouse IFN-γ antibody (1:500 in PBS; BD Biosciences, Franklin Lakes, USA) for 1 h at RT. Subsequently, plates were washed seven times with PBS. To label bound biotinylated antibodies, 100 µl AKP Streptavidin (1:500 in PBS; BD Biosciences, Franklin Lakes, USA) were added and the plates were incubated for 30 min at RT in the dark. This was followed by six washing steps with PBS before 1-Step NBT/BCIP Substrate Solution (100 µl/well; Thermo Fisher Scientific, Germany) was added to start the enzymatic reaction. Plates were incubated up to 15 min depending on the color development and the reaction was eventually stopped by adding dH_2O. The plates were dried overnight and analyzed using the CTL ImmunoSpot Reader (ImmunoSpot, Cleveland, USA). As positive control for IFN-γ production, the known T cell activator Concanavalin A (2 µg/ well; Merck, Darmstadt, Germany) was used in all performed experiments. Results were visualized in GraphPad Prism (v8.43, Graphpad Software Inc, San Diego, USA).

2.10.5 T cell killing assay

Remaining splenocytes isolated of each animal of a vaccination group (Table 5) which were not used for ELISpot assays were pooled to have a sufficient amount of input material for T cell killing assays. The splenocytes were cultured in lymphocyte media consisting of RPMI media supplemented with 10 mM HEPES, 10% FBS, 2 mM glutamine, 1% sodium pyruvate, 1% non-essential amino acids, 50 µM β-mercaptoethanol, and combined antibiotics. Peptides for which a strong T cell observation was observed in ELISpot assays were added to the cells at 100 nM final concentration. Splenocytes of the control Group 1 were divided and either stimulated with *HPV16* epitope E7 as positive control or the *MYH11.1* peptide as negative control. Group 2 splenocytes were challenged with the *ANXA1* peptide and Group 4 with the *CCNB1 #2* peptide. Splenocytes of Group 3 were incubated with both *CCNB1 #1* and *UBE2V2* peptides. After 72 h, splenocytes were stimulated with 10 IU/ml interleukin 2 (BioLegend, San Diego, USA) for 96 h. After one week in culture, $CD8^+$ T cells were isolated following the instructions of the $CD8a^+$ T Cell Isolation Kit, mouse (Miltenyi, Bergisch Gladbach, Germany).

In parallel, target cells derived from the A2.DR1 cell line were labeled and loaded with peptides. Cell labeling was achieved using either CellTrace CFSE Cell Proliferation Kit or CellTrace FarRed Cell Proliferation Kit (both Thermo Fisher Scientific, Waltham, USA). The 10 mM CFSE stock was pre-diluted 1:100 in PBS and the 1mM FarRed stock was pre-diluted 1:10 to obtain a working concentration of 1 mM each. Cell were stained using a concentration of $1x10^6$ cells/ml by incubation with 20 µl of CFSE

and 2.5 µl of the FarRed dilution for 15 min at 37 °C. After staining, cells were washed twice with 1 ml RPMI 10% FBS-supplemented per $1x10^6$ cells and finally resuspended in 1 ml RPMI 10% FBS-supplemented media per $1x10^6$ cells. 1 ml per well of labeled cell suspension was transferred into a 6-well plate. 2 ml of RPMI Medium 1640 supplemented with 10% FBS and 1% Gibco Penicillin-Streptomycin were added per well and cells were incubated for 24 h at 37 °C, 5% CO_2. The following day, peptides were loaded onto target cells. Cells were processed and resuspended in 1 ml peptide solution diluted in RPMI 10% FBS-supplemented media obtaining a final concentration of 10 µM peptide per sample. Target cell peptide mix was incubated for 90 min at 37 °C, 5% CO_2 and subsequently washed and resuspended in RPMI 10% FBS-supplemented media. 1,000 of both CFSE- and FarRed-labeled cells were transferred per sample onto a 96-well plate.

Isolated $CD8^+$ T cells were prepared in RPMI media supplemented with 10% and 20 IU/ml IL-2 and were added at different concentrations to the target cells. Effector and target cells were co-cultured at 37 °C, 5% CO_2 for 48 h. After incubation time, supernatants were transferred onto a new 96-well plate and cells were pelleted at 300 x *g* for 5 min. In parallel, adherent cells were detached by trypsinization and added to the new 96-well plate. Cells were again pelleted and then resuspended in 100 µl FACS buffer per well. T cell killing of labeled target cells was evaluated by flow cytometry using the FACSCanto II (BD Biosciences, Franklin Lakes, USA).

2.11 Software, Programs, and R packages

Table 6 | R packages used for this study.

Package	Version	Source
biomaRt	2.38.0	Durinck et al., 2009
clusterProfiler	3.10.0	Yu et al., 2012
ComplexHeatmap	1.20.0	Gu et al., 2016
DESeq2	1.22.2	Love et al., 2014
dplyr	1.0.0	Wickham et al., 2020
enrichplot	1.2.0	Yu et al., 2012
factoextra	1.0.7	Kassambara and Mundt, 2020
fgsea	1.8.0	Sergushichev, 2016
ggplot2	3.0.3	Wickham, 2016
gplots	3.0.3	Warnes et al., 2020
JunctionSeq	1.5.4	Hartley and Mullikin, 2016
magrittr	1.5	Wickham & Bache, 2014
msigdbr	7.1.1	Dolgalev, 2020
pathfindR	1.6.3	Ulgen et al., 2019
pheatmap	1.0.12	Kolde, 2019
RColorBrewer	1.1.2	Neuwirth, 2014
rlang	0.4.6	Henry & Wickham, 2020
SeqinR	1.0.2	Charif and Lobry, 2007
tidyr	1.1.0	Henry & Wickham, 2020
tximport	1.10.0	Soneson et al., 2016

Statistical data analysis and visualization were performed in R (v3.5.1, R Core Team, 2018). All R packages used are listed in Table 7. R and bash scripts are available on the RSC GitHub repository: 'https://github.com/MauerLab'.

Table 7 | Programs and software used for this study.

Software	Version	Reference
2100 Expert Software	B.02.11 (SR1)	Agilent, Santa Clara, USA
BEDTools	2.17.0	(Quinlan and Hall, 2010)
BLASTp	2.2.31+	(Camacho et al., 2009)
GibbsCluster	2.0	(Andreatta et al., 2017)
GraphPad Prism	8.43	Graphpad Software Inc, San Diego, USA
NetMHCpan	4.0	(Jurtz et al., 2017)
PEAKS Studio X	10.0	Bioinformatics Solutions Inc., Waterloo, Canada
Python	2.7	(van Rossum, 2015)
QoRTs	1.3.6	(Hartley and Mullikin, 2016)
R	3.5.1	(R Core Team, 2018)
rMATS	4.0.2	(Shen et al., 2014)
rmats2sashimiplot	2.0.4	(Xie et al., 2015)
Salmon	1.6.0	(Patro et al., 2017)
Seq2Logo	2.0	(Thomsen and Nielsen, 2012)
Sequence Specific Retention Calculator (SSRCalc)	Q	(Krokhin et al., 2006)
STAR software V2.7.0	2.7.0	(Dobin et al., 2013)
StringTie	2.1.1	(Pertea et al., 2015)
TransDecoder	5.5.0	(Haas, 2018)

2.12 Equipment and reagents

Table 8 | Important materials used for this study.

Material	Supplier
10 µm polypropylene membranes	Agilent, Santa Clara, USA
5-layer Rectangular Straight Neck Cell Culture Multi-Flasks	Thermo Fisher Scientific, Waltham, USA
96-deep well collection plate	Agilent, Santa Clara, USA
96-well filter microplate with 3 µm glass fiber and 10 µm polypropylene membranes	Agilent, Santa Clara, USA
Acclaim PepMap 100 C18, 5µm, 300 µm i.d. x 5 mm, 100 Å	Thermo Fisher Scientific, Waltham, USA
MultiScreen$_{HTS}$-IP plates	Merck, Darmstadt, Germany
nanoEase MZ HSS T3 column, 100 Å, 1.8 µm, 75 µm x 250 mm	Waters, Milford, USA
Poly-Prep columns	Bio-Rad, Hercules, USA
Sep-Pak tC18 96-well Plate, 100 mg Sorbent per Well, 37-55 µm Particle Size	Waters, Milford, USA

Table 9 | Equipment used for this study.

Device	Supplier
2100 Bioanalyzer	Agilent, Santa Clara, USA
CTL ImmunoSpot Reader	ImmunoSpot, Cleveland, USA
FACSCanto II	BD Biosciences, Franklin Lakes, USA
FUSION FX Imaging System	Vilber Lourmat, Eberhardzell, Germany
GloMax Microplate Reader	Promega, Madison, USA
Heracell VIOS 160i CO2 incubator	Thermo Fisher Scientific, Waltham, USA
Herasafe cell culture hood	Thermo Fisher Scientific, Waltham, USA
NanoDrop 2000/2000c spectral photometer	Thermo Fisher Scientific, Waltham, USA
NanoPhotometer® NP80	Implen, Munich, Germany
Orbitrap Fusion Lumos	Thermo Fisher Scientific, Waltham, USA
PCR machine	Bio-Rad, Hercules, USA
Positive Pressure-96 Processor	Waters, Milford, USA
Qubit 3.0 Fluorometer	Thermo Fisher Scientific, Waltham, USA
Trans-Blot Turbo Transfer System	Bio-Rad, Hercules, USA
UltiMate 3000 RSLCnano System	Thermo Fisher Scientific, Waltham, USA

Table 10 | Antibodies used for this study.

Antibody	Source	Catalogue number
Vinculin Antibody (rabbit)	Cell Signaling Technology, Danvers, USA	#4650, RRID = AB_10559207
AKP Streptavidin	BD Biosciences, Franklin Lakes, USA	#554065
Biotin Mouse Anti-Human IFN-γ	BD Biosciences, Franklin Lakes, USA	#554550, RRID = AB_395472
Purified Rat Anti-Mouse IFN-γ	BD Biosciences, Franklin Lakes, USA	#559065, RRID = AB_2123177
Anti-rabbit IgG, HRP-linked Antibody #7074	Cell Signaling Technology, Danvers, USA	#7074, RRID: AB_2099233
Anti-mouse IgG, HRP-linked Antibody	Cell Signaling Technology, Danvers, USA	#7076, RRID = AB_330924
Mouse monoclonal anti-HLA-ABC antibody (W6/32)	(Barnstable et al., 1978)	self-supplied

Table 11 | Reagents used for this study.

Reagent	Source
1-Step™ NBT/BCIP Substrate Solution	Thermo Fisher Scientific, Waltham, USA
1X Red Blood Cell (RBC) lysis buffer	Sigma-Aldrich; St. Louis, USA
Acetonitrile	Sigma Aldrich, St. Louis, USA
CD8a+ T Cell Isolation Kit, mouse	Miltenyi, Bergisch Gladbach, Germany
CellTiter-Glo 2.0 Assay	Promega, Madison, USA
CellTrace CFSE Cell Proliferation Kit	Thermo Fisher Scientific, Waltham, USA
CellTrace FarRed Cell Proliferation Kit	Thermo Fisher Scientific, Waltham, USA
Concanavalin A	Merck, Darmstadt, Germany
CpG ODN 1826	TIB Molbiol, Berlin, Germany
DMEM/F-12 GlutaMAX	Thermo Fisher Scientific, Waltham, USA
Ethylenediaminetetraacetic acid	Sigma Aldrich, St. Louis, USA
fetal bovine serum (FBS)	Thermo Fisher Scientific, Waltham, USA
GeneJET Genomic DNA Purification Kit	Thermo Fisher Scientific, Waltham, USA
Gibco Penicillin-Streptomycin	Thermo Fisher Scientific, Waltham, USA
Halt Protease and Phosphatase Inhibitor Cocktail, EDTA-free	Thermo Fisher Scientific, Waltham, USA
Human TruStain FcX	BioLegend, San Diego, USA
Iodoacetamide	Sigma Aldrich, St. Louis, USA
Milk powder	Carl Roth, Karlsruhe, Germany
Nuclease-free water	New England Biolabs, Ipswich, USA
NuPAGE 4-12% Bis-Tris gels (1.5 mm x 10/15 wells)	Thermo Fisher Scientific, Waltham, USA
NuPAGE Antioxidant	Thermo Fisher Scientific, Waltham, USA
NuPAGE LDS Sample Buffer (4x)	Thermo Fisher Scientific, Waltham, USA
NuPAGE MES SDS Running Buffer (20x)	Thermo Fisher Scientific, Waltham, USA
NuPAGE Sample Reducing Agent (10x)	Thermo Fisher Scientific, Waltham, USA
Octyl-β-D-glucopyranoside	Sigma Aldrich, St. Louis, USA
OneTaq Quick-Load 2X Master Mix with Standard Buffer	New England Biolabs, Ipswich, USA
Optima LC/MS Grade Water	Thermo Fisher Scientific, Waltham, USA
PageRuler Plus Prestained Protein Ladder (10 to 250 kDa)	Thermo Fisher Scientific, Waltham, USA
Pierce BCA Protein Assay Kit	Thermo Fisher Scientific, Waltham, USA
ProtoScript II First Strand cDNA Synthesis Kit	New England Biolabs, Ipswich, USA
Qubit 3.0 Fluorometer	Thermo Fisher Scientific, Waltham, USA
Qubit RNA BR Assay Kit	Thermo Fisher Scientific, Waltham, USA
Recombinant Mouse IL-2	BioLegend, San Diego, USA
RNA 6000 Nano Kit	Agilent, Santa Clara, USA
RNaseZAP	Thermo Fisher Scientific, Waltham, USA
RNAzol RT	Sigma Aldrich, St. Louis, USA
RPMI 1640 (ATCC modification)	Thermo Fisher Scientific, Waltham, USA
SIRV Set 1	Lexogen, Vienna, Austria
Sodium deoxycholate	Thermo Fisher Scientific, Waltham, USA
SpikeTides	JPT Peptide Technologies, Berlin, Germany
SuperSignal West Femto Maximum Sensitivity Substrate	Thermo Fisher Scientific, Waltham, USA
Trans-Blot Turbo Transfer Pack with PVDF membrane (0.2 µm)	Bio-Rad, Hercules, USA
Trifluoracetic acid	Sigma Aldrich, St. Louis, USA

2.13 Primers and synthetic peptide sequences

Table 12 | Primers used for validation of alternative splice variants by RT-PCR. *CDS = coding DNA sequence*

Gene	forward primer	reverse primer
ANXA1	CTGAAGAGAGATCTGGCCAAAGAC	GGATAGCTTCTGGTGGTAAGGATG
ANXA1 CDS	CAGAAGCCCAAGTCTCCACTG	CTGATCATAGCTTGAGACCATCAAG
CCNB1	GATGTCGAGCAACATACTTTGGC	GAACTAGTGCAGAATTCAGCTGTG
PSM4A	CCAAGTTGAATATGCCATGGAAGC	GCTGTAACCAACTGCTCACAAG
TUBA1A	CTGGTGTCCAGATTGGCAATG	CCTCGGGCATAGTTATTGGCAG
UBE2V2	GTAAAAGTCCCTCGCAATTTCCG	GAAGCTCTTGCAGGACAACTTTG

Table 13 | Synthetic peptides used for peptide sequence validation and immunogenicity assays.

Gene	Peptide sequence	Supplier for peptide sequence validation	Supplier for immunogenicity assays
ANXA1	SLAKALYEA	JPT Peptide Technologies, Berlin, Germany	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany
PSMA4	KLNEYLLQY	JPT Peptide Technologies, Berlin, Germany	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany
CCNB1 #1	ILDNGEWTV	JPT Peptide Technologies, Berlin, Germany	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany
UBE2V2	GMIIGPPRV	JPT Peptide Technologies, Berlin, Germany	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany
CCNB1 #2	KILDNGEWTV	JPT Peptide Technologies, Berlin, Germany	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany
ANT1	GMLPDPKNV	JPT Peptide Technologies, Berlin, Germany	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany
HPV16-E7	YMLDLQPET	-	GMP and T Cell Therapy Unit, German Cancer Research Center, Heidelberg, Germany

3. Results

3.1 HLA-I typing of model cell lines

Self-peptides and neoepitopes are presented on the cell surface by HLA-I molecules. However, the immunopeptidome of patients and cell lines can be very individual due to different HLA haplotypes. Based on their sequence, HLAs can be classified into various subtypes by gene, allele group, and specific HLA protein (Figure 12A). These features contribute to the extreme diversity of HLA molecules. For example, the IPD and IMGT/HLA database reported that 23,002 different HLA-I alleles were identified as of September 2021 (www.hla.alleles.org; Robinson et al., 2015). The HLA-I subtypes occur in varying frequencies, with some rare and others very common. For example, a study by Deutsche Knochenmarkspenderdatei (DKMS) revealed that the HLA-A*02:01 subtype was most common and occurred at a frequency of 26.7% in the German bone marrow registry including 39,689 donors (Figure 12B). In contrast, 127 of the 140 detected HLA-A alleles in the dataset occurred with a frequency of less than 1% (Allele Frequency Net Database; Gonzalez-Galarza et al., 2020). The differences in the sequences of HLA subtypes influence peptide binding because ligands need to match a specific binding motif to be able to be presented on a respective HLA receptor (Figure 12C). Consequently, the immunopeptidome composition heavily depends on the HLA haplotype.

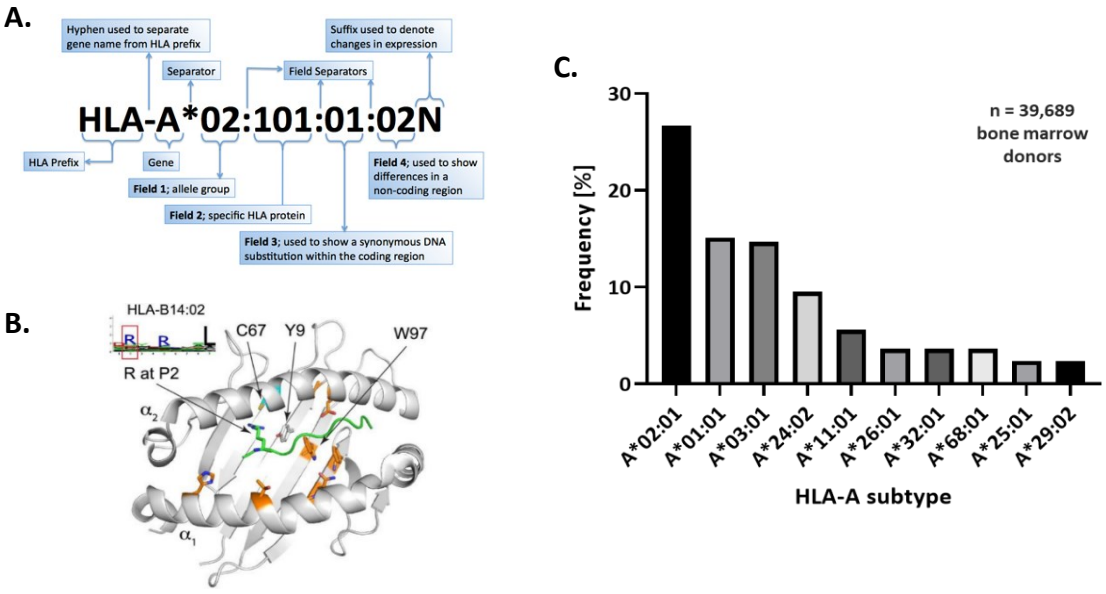


Figure 12 | Diversity of HLA-I alleles influences peptide binding. (A) *HLA nomenclature* (Marsh et al., 2010). ***(B)*** *Top 10 most occurring HLA-A alleles found in the German bone marrow registry (n = 39,689) provided by the DKMS and available in the Allele Frequency Net Database* (Gonzalez-Galarza et al., 2020). ***(C)*** *Structural view of HLA-B14:02 in complex with a peptide* (Bassani-Sternberg et al., 2017).

To allow antigen matching to specific HLA-I alleles, I determined the HLA-I haplotype of model cell lines used in this study by genomic DNA haplotyping. All K562 cell lines exhibited the same haplotype independent of splicing factor mutations (Table 14). Here, the haplotyping results were compared to predicted haplotypes annotated in the TRON Cell Line Portal. The HLA typing results for K562 cells matched the HLA-A alleles predicted from the NCI-60 Human Tumor Cell Lines Screen by high-resolution sequence-based typing (Adams et al., 2005). Also, the results for HLA-B and HLA-C allele groups matched, but there were non-resolved ambiguities for the identification of the specific HLA proteins in the literature reference. Besides, I compared the haplotyping results to HLA typing from *seq2HLA* based on RNA-seq data (Boegel et al., 2014). This delivered matching results for HLA-C alleles, but seq2HLA predictions for HLA-A and HLA-B alleles did not match other haplotyping results from DKMS or NCI-60. Moreover, I determined the HLA-I haplotype of RPE-1 cells since I could not find any information about it in publicly available datasets. Both tested RPE-1 cell lines carried the same HLA-I subtypes (Table 14). To summarize, the DKMS haplotyping of our model cell lines resolved existing ambiguities in K562 cells and provided novel information for RPE-1 cells.

Table 14 | HLA-I haplotypes of K562 and RPE-1 cell lines. *Haplotyping was performed by Deutsche Knochenmark-spenderdatei (DKMS) using isolated gDNA samples and was compared to haplotypes found in public databases. * = Non-resolved ambiguities; ** = Ambiguity on 4-digit level, chosen allele had the highest number of reads.*

Cell line	HLA-A Allele 1	HLA-A Allele 2	HLA-B Allele 1	HLA-B Allele 2	HLA-C Allele 1	HLA-C Allele 2	Source	Reference
K562 WT	*11:01:01*	*31:01:02*	*18:01:01*	*40:01:02*	*03:04:01*	*05:01:01*	*DKMS*	*own experiment*
K562 *SF3B1^{K700E}*	*11:01:01*	*31:01:02*	*18:01:01*	*40:01:02*	*03:04:01*	*05:01:01*	*DKMS*	*own experiment*
K562 *U2AF1^{S34F}*	*11:01:01*	*31:01:02*	*18:01:01*	*40:01:02*	*03:04:01*	*05:01:01*	*DKMS*	*own experiment*
K562 WT	*11:01:01*	*31:01:02*	*18**	*40**	*03**	*n/a*	NCI-60	(Adams et al., 2005)
K562 WT	*24:02***	*26:02***	*81:01*	*35:18*	*03:04*	*05:01*	seq2HLA	(Boegel et al., 2014)
RPE-1 WT	*01:01:01*	*02:01:01*	*07:02:01*	*08:01:01*	*07:01:01*	*07:02:01*	*DKMS*	*own experiment*
RPE-1 *TP53$^{-/-}$*	*01:01:01*	*02:01:01*	*07:02:01*	*08:01:01*	*07:01:01*	*07:02:01*	*DKMS*	*own experiment*

3.2 HLA-I expression in cell models

Apart from carrying different HLA alleles, HLA expression levels vary across cell lines (Boegel et al., 2014). Thus, the HLA-I expression was determined across all cell line models to ensure peptide presentation on the cell surface for HLA-I immunoaffinity purification (HLA-I IP) experiments.

3.2.1 HLA-I expression in K562 cell lines

K562 cells do not naturally express HLA-I at baseline, but HLA-I expression can be induced by IFN-γ stimulation (Sutherland et al., 1985). Indeed, I did not detect HLA-I expression at protein level of non-stimulated K562 cells as revealed by western blot analysis (Figure 13A). Stimulation of K562 WT cells with 50 ng/ml IFN-γ resulted in a robust HLA-I expression after 24 hours of treatment. HLA-I expression levels further increased when the IFN-γ stimulation was prolonged to 48 hours. After removing the IFN-γ stimulation, HLA-I expression abated in a time-dependent manner (Figure 13A). This demonstrates that HLA-I expression can be induced and maintained by IFN-γ stimulation.
Furthermore, I examined the effect of the SF3B1^{K700E} splicing factor mutation on HLA-I expression. At low IFN-γ concentrations, the HLA-I expression of the SF3B1^{K700E} cell line was lower compared to WT cells after 24 hours of INF-γ stimulation (Figure 13B). However, this effect was abolished when K562 cell lines were treated with higher doses of IFN-γ (Figure 13C). Accordingly, HLA-I expression levels did not differ in samples treated with 50 or 100 ng/ml IFN-γ, suggesting saturation of HLA-I expression at higher IFN-γ concentrations. Concluding from these results, a concentration of 50 ng/ml IFN-γ to achieve a consistent HLA-I expression across all K562 cell lines.

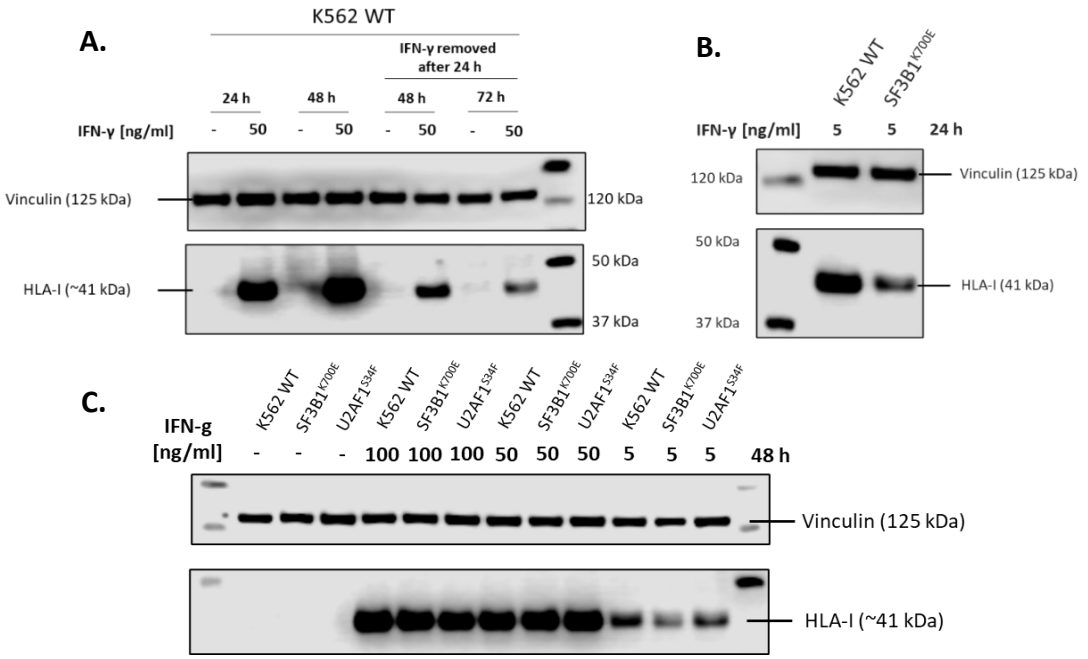


Figure 13 | Induction of HLA-I expression in K562 cells by IFN-γ stimulation. (A) *Western blot detection of HLA-I protein levels in K562 cells before and after stimulation with 50 ng/ml IFN-γ.* ***(B)*** *Comparison of HLA-I expression levels in K562 SF3B1^{K700E} cells to K562 WT cells in response to 50 ng/ml IFN-γ stimulation for 24 h.* ***(C)*** *Detection of HLA-I protein levels in K562 cell lines in response to 0, 5, 50, and 100 ng/ml IFN-γ stimulation.*

After successfully confirming HLA-I expression by western blot, I used flow cytometry analysis to assess HLA-I presentation on the cell surface of K562 cell lines. This analysis revealed a shift in the signal of IFN-γ-stimulated K562 WT cells compared to unstimulated cells, indicating an increase in HLA-I surface expression (Figure 14A). All three K562 cell lines showed similar intensities of signal shift, although the response of SF3B1^{K700E} cells was slightly attenuated (Figure 14A-C). Overall, the IFN-γ-induced shift did not result in a separated population as the signals of both groups overlapped. In contrast, HCT-116 cells naturally expressing HLA-I clearly distinguished from the K562 negative control (Figure 14D). This suggested that IFN-γ stimulation was able to induce HLA-I expression in K562 cells, but to a lower extent compared to cells naturally expressing HLAs.

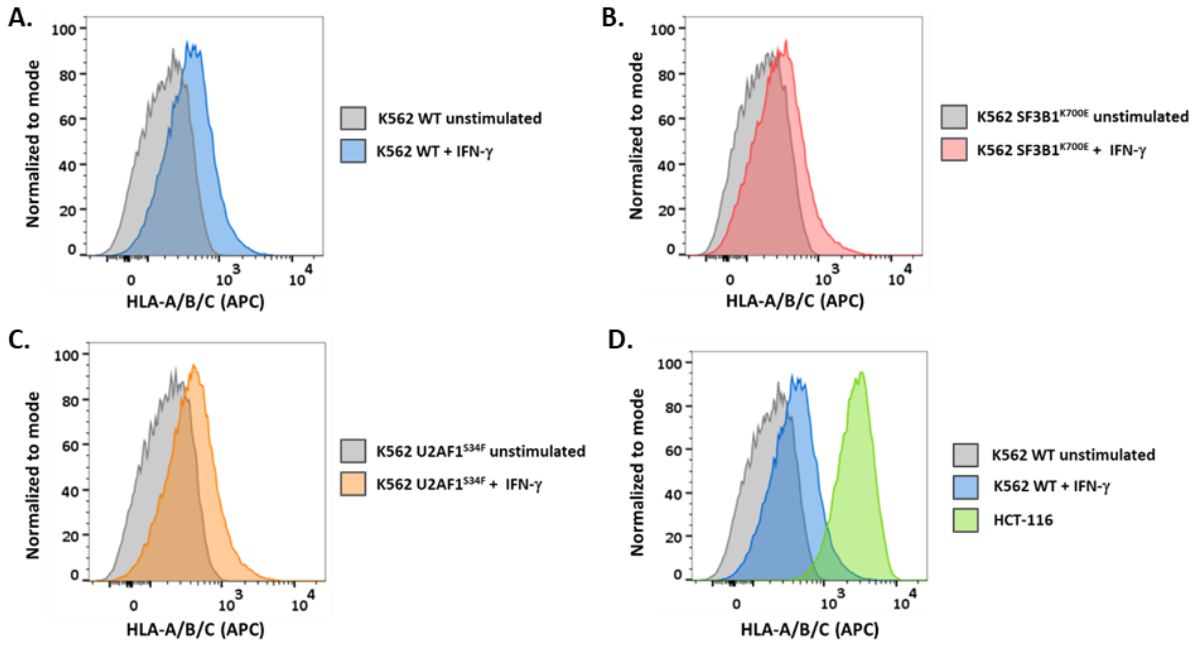


Figure 14 | Induction of HLA-I surface expression of K562 cell lines in response to IFN-γ stimulation. (A-C) *Flow cytometry analysis of HLA-I surface expression of K562 cell lines in response to IFN-γ stimulation (50 ng/ml for 48 h).* ***(D)*** *Comparison of HLA-I surface expression between K562 cells stimulated with 50 ng/ml IFN-γ for 48 h and HCT-116 cells naturally expressing HLA-I.*

3.2.2 HLA-I expression of RPE-1 cell lines

Next, I investigated HLA-I levels in RPE-1 cell lines included in the study, which all naturally expressed HLA-I (Figure 15). Furthermore, treatment with the splicing inhibitor GEX1A had no impact on HLA-I expression levels since no difference in protein level was detected between non-treated and GEX1A-treated samples (Figure 15). Similar to the different K562 cell lines, HLA-I expression levels slightly differed between RPE-1 cell lines (Figure 15). The expression was more pronounced in WT and *MSH2*$^{-/-}$ cells. In addition, two bands instead of a single band were detected in all samples. The double bands were more prominent in the aged knockout/mutation cell lines. These were likely caused by different HLA-I fragments resulting from sample degradation, which were detected by the pan-HLA-I W6/32 antibody (Wiesch and Meyers, 2001). Thus, the data suggested that HLA-I expression was not affected by GEX1A treatment.

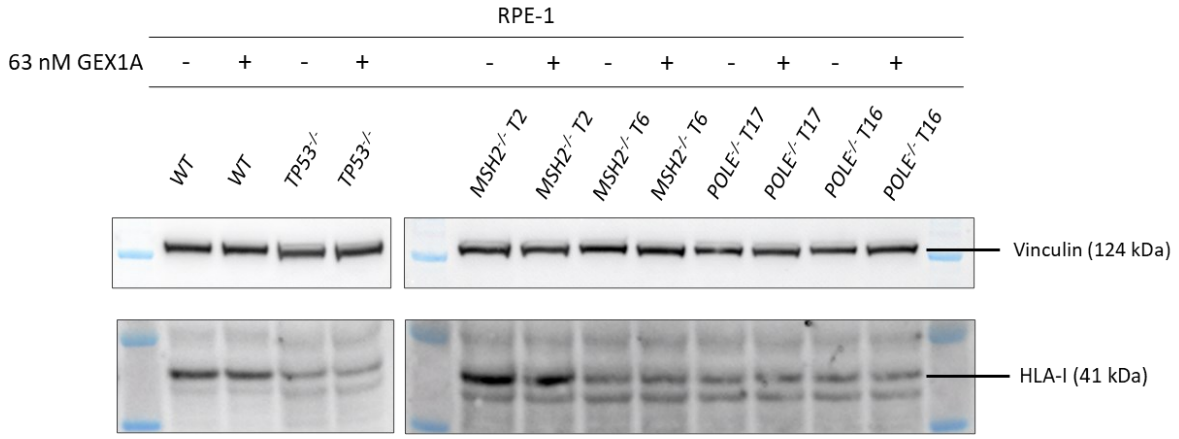


Figure 15 | HLA-I expression levels in RPE-1 cell lines. *Western Blot detection of HLA-I protein expression in RPE-1 cell lines treated with 0 nM or 63 nM GEX1A for 24 h.*

To investigate the impact of mutations/knockouts of components of the DDR pathway on HLA-I surface expression, I analyzed RPE-1 cell lines by flow cytometry. All cell lines demonstrated a strong HLA-I surface expression reflected by the positive signal shift of labeled samples (Figure 16A). Differences in the intensities of the signal shift between wild-type and knockout/mutation cell lines might be explained by different cell morphologies influencing the signal measurement. Especially the RPE-1 *POLE*^{P286R} T16 cell line exhibited a broader signal range, most likely resulting from the different cell morphology caused by its tetraploidy. Accordingly, the increased cell size of RPE-1 *POLE*^{P286R} heavily affects cell scattering (Figure 16B). Nevertheless, all RPE-1 cell lines exhibited a strong HLA-I surface expression and were thus suitable cell models for following HLA-I IP experiments.

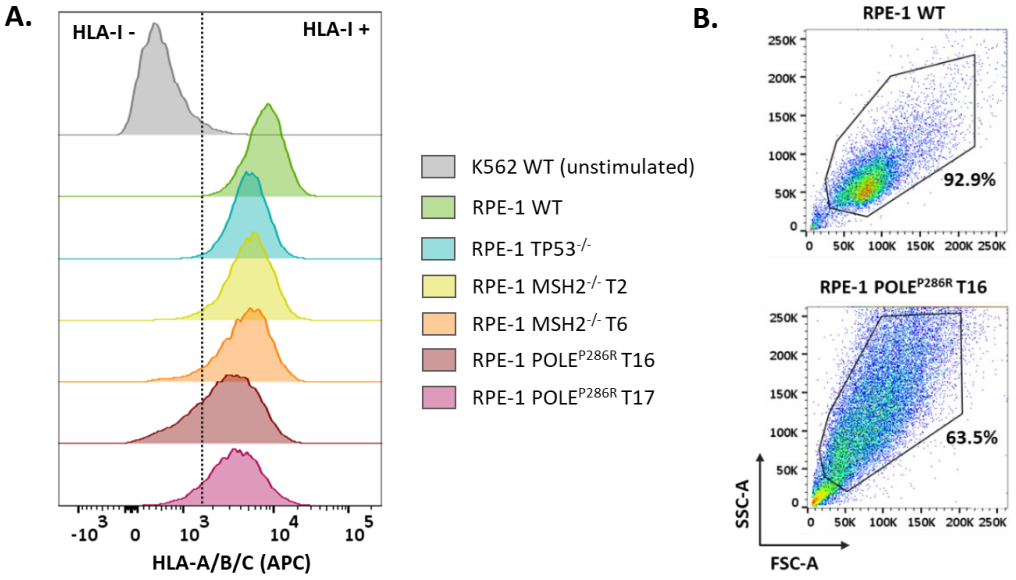


Figure 16 | Differences in HLA-I surface expression and cell morphology of RPE-1 cell lines. (A) *Flow cytometry analysis of HLA-I surface expression of RPE-1 cell lines.* ***(B)*** *Cell morphology differences of tetraploid RPE-1 POLE*^{P286R} *T16 cell line compared to RPE-1 WT cells visualized in cell scatter plots.*

3.3 Enzyme-free detachment of adherent cells

To harvest RPE-1 cell lines for HLA-I IP experiments, I used trypsinization as a standard procedure for cell detachment from culture dishes. However, this enzymatic digestion process could damage cell surface proteins such as HLA-I (Zhang et al., 2012). Thus, a previous study used a mechanical scraping strategy to detach cells for HLA-I IP (Becker et al., 2021). To estimate how scraping affected cell viability, I measured RPE-1 and HCT-116 cell viability by flow cytometry. Following a stringent gating strategy to identify viable cells, 88.7% of RPE-1 cells detached with TrypLE were viable (Figure 17). In contrast, only 35.4% of cells detached by scraping were viable. To eliminate the possibility that the RPE-1 cell line is particularly sensitive to scraping, they were compared to HCT-116 cells. Accordingly, much higher cell viability was obtained when HCT116 cells were detached using TrypLE instead of cell scraping (76.7% vs. 45.6 %). The flow cytometry results demonstrated that mechanical cell scraping caused a decrease in cell viability and consequently, less input material would be available for HLA-I IP experiments.

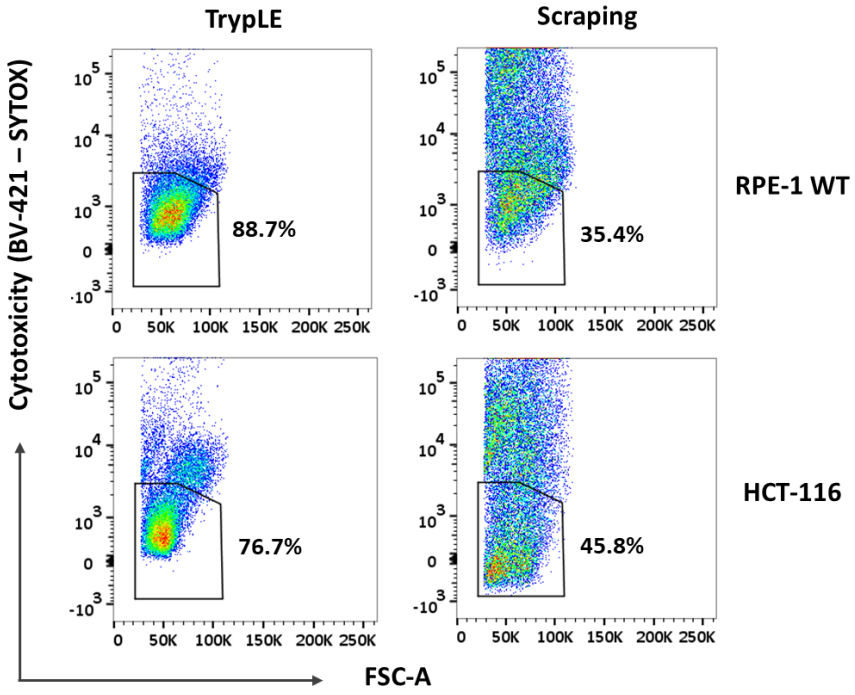


Figure 17 | Impact of cell detachment method on cell viability. *RPE-1 WT and HCT-116 cells were detached by incubation in TrypLE for 5 min or by cell scraping. Detached cells were stained with SYTOX Blue Dead Cell Stain allowing the quantification of viable cells via flow cytometry analysis.*

Additionally, I determined the effect of cell detachment on the immunopeptidome diversity. Similar numbers of RPE-1 cells obtained from mechanical detachment, TrypLE, or enzyme-free buffer were processed by HLA-I IP and subsequently analyzed by mass spectrometry (Figure 18A). Identification of HLA-I-bound peptides revealed the highest immunopeptidome diversity in the sample detached using enzyme-free buffer (4850 peptides, Figure 18B). In samples where cells were detached by scraping and TrypLE, lower numbers of unique peptides were identified (3844 and 3456 peptides). Although cell scraping might not have a considerable impact on immunopeptidome diversity compared to TrypLE detachment, it still reduced the number of viable cells serving as input for immunopeptidomics experiments.

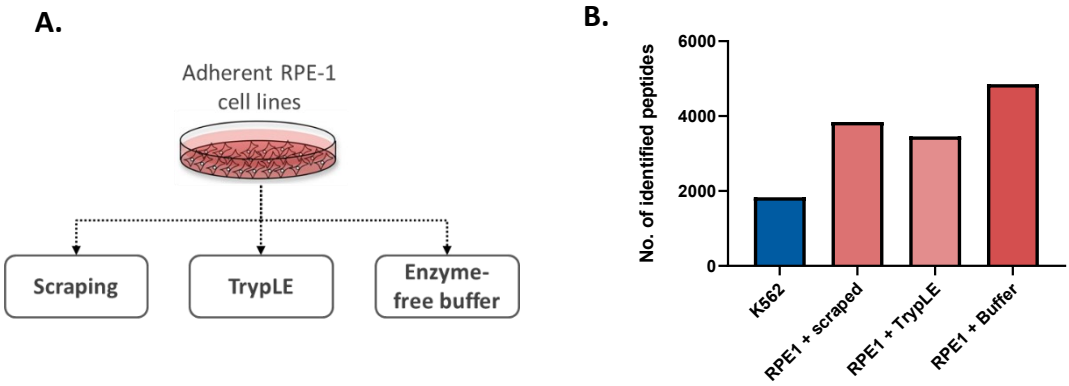


Figure 18 | Impact of cell detachment method on immunopeptidome diversity. (A) *Adherent RPE-1 WT cells were detached by cell scraping or by incubation in TrypLE or enzyme-free buffer for 10 min. Samples were processed by HLA-I IP and captured peptides subsequently analyzed by mass spectrometry.* ***(B)*** *Number of unique identified HLA-I-bound peptides discovered by PEAKS search against the UniProt database.*

Concluding from these results, I used enzyme-free buffer to detach adherent RPE-1 cells for HLA-I IP experiments for the following reasons:

i) Enzyme-free buffer eliminated the possibility of enzymatic cleavage of surface proteins

ii) Cell scraping caused cell death and thereby decreased the input amount for HLA-I IP and further could not be applied in 5-layer flasks

iii) Samples treated with enzyme-free buffer showed the highest immunopeptidome diversity in test experiments

3.4 Pharmacological splicing inhibition

The phenotype of aberrant alternative splicing in tumor cells was recapitulated by treating cells with small molecule inhibitors of the splicing machinery. Such splicing inhibitors limit the availability of active core spliceosomal components, thereby mimicking overburdening of the splicing machinery, a frequently observed condition in cancer cells (Kahles et al., 2018). To determine a suitable treatment dose for immunopeptidomics experiments, growth rate (GR) inhibition was determined for each cell line.

3.4.1 Splicing inhibition in K562 cell lines

First, I evaluated the dose-response of GEX1A treatment for K562 WT cells and cell lines carrying mutations in splicing factors *SF3B1* and *U2AF1*. Dose-response curves revealed that splicing factor mutant cell lines were slightly more susceptible to splicing inhibition by GEX1A (Figure 19). This was reflected in a decreased cell survival after 24 hours of treatment. Consequently, the GR_{50} value of K562 WT cells (27.95 nM) was higher than those of K562 $SF3B1^{K700E}$ cells (20.88 nM) and K562 $U2AF1^{S34F}$ cells (23.02 nM). However, these differences did not reach statistical significance. Thus, for HLA-I IP experiments, all cell lines were treated with the GR_{25} value of K562 WT cells (10 nM) highlighted in Figure 19.

I chose the final concentration based on the results of WT cells. The other cell lines were treated with the same dose of GEX1A to resemble treatment conditions in patients.

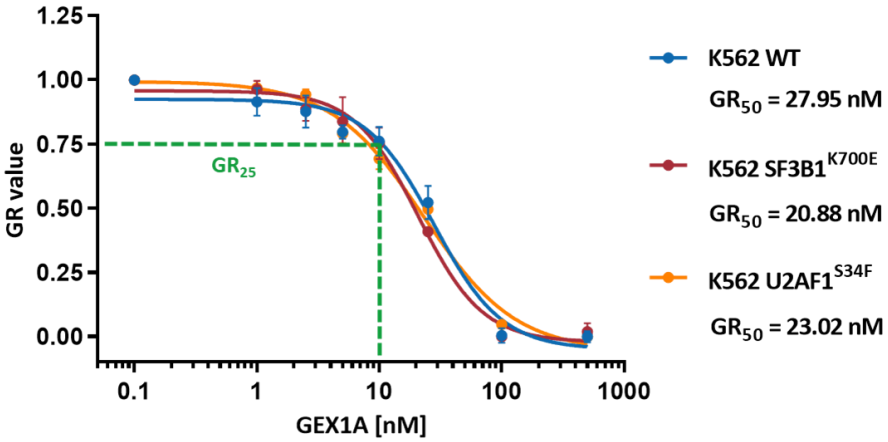


Figure 19 | Dose response curve of K562 cell lines treated with GEX1A for 48 h. *Cell viability was determined by CellTiter-Glo measurements at start and end of drug treatment.*

3.4.2 Splicing inhibition in RPE-1 cell lines

Moreover, I evaluated the dose-response of GEX1A treatment for RPE-1 WT cells, RPE-1 cells with loss of tumor suppressor *TP53*, and RPE-1 cells with genetic deficiencies in the DNA damage repair pathway (Figure 20). Loss of *TP53* made RPE-1 cells more susceptible to GEX1A treatment indicated by the lower GR_{50} value of *TP53* knockout cells compared to WT cells (21.48 nM vs. 31.15 nM). Along with this, *TP53* knockout cells carrying additional defects in the DNA damage repair machinery were even more susceptible to GEX1A treatment (*MSH2*$^{-/-}$ = 17.03 nM and *POLE*^{P286R} = 13.78 nM).

In previous HLA-I IP experiments with K562 cells, I observed no cell death after 24-48 h of treatment, even at higher GEX1A concentrations. Instead, the differences in cell numbers of treated samples resulted from affected cell proliferation due to GEX1A treatment. Thus, I chose the GR_{75} value of RPE-1 WT cells (63 nM) for RPE-1 HLA-I IP experiments to obtain an even more prevalent aberrant splicing phenotype in these RPE-1 samples.

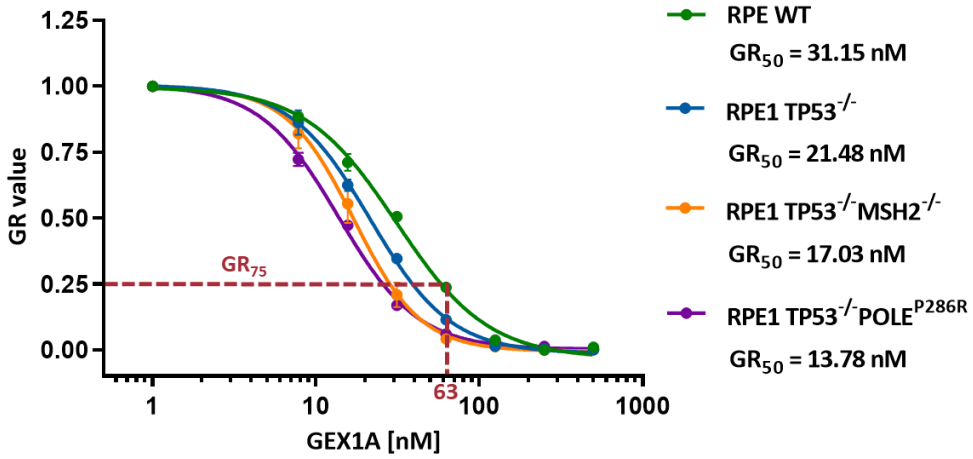


Figure 20 | Dose-response curve of RPE-1 cell lines treated with GEX1A for 48 h. *Cell viability was determined by CellTiter-Glo measurements at start and end of drug treatment.*

3.5 Recovery of cell immunopeptidomes

To explore the immunopeptidome of cells with splicing perturbations, I recovered HLA-I-presented peptides by HLA-I immunopurification (HLA-I IP). This directed approach is the only unbiased methodology allowing to examine the repertoire of naturally presented peptides (Bassani-Sternberg, 2018). Following this procedure, I isolated HLA-I molecules from samples by immunoaffinity-based purification using anti-HLA-I antibodies. Subsequently, I extracted HLA-I peptides, which were then analyzed by mass spectrometry.

3.5.1 Antibody purification & cross-linking

To obtain enough material for all HLA-I IP assays, I needed to have sufficient amounts of W6/32 HLA-I antibody available. Thus, I purified W6/32 HLA-I antibody at large scale from cell culture supernatants of HB-95 hybridoma cells, which are engineered for antibody production (Bassani-Sternberg, 2018; provided by Jonas Becker). After following the antibody purification protocol, eluates were pooled and tested for IgG content. In total, the purification yielded 20 mg of W6/32 HLA-I antibody from 200 ml cell supernatant, which was sufficient to perform all HLA-I IP experiments carried out in this study.
To capture HLA-I molecules during the HLA-I IP, I cross-linked W6/32 HLA-I antibodies w to Sepharose beads. I chose Sepharose A-coated beads since protein A possesses a strong binding affinity towards IgG-type antibodies such as W6/32. Since protein A specifically recognizes the Fc region of antibodies, the beads also ensured correct orientation of bound antibodies and thereby improved immunopurification efficiency. To determine cross-linking efficiency, I compared the protein content of W6/32 antibody aliquots and flow-through samples. The IgG content of flow-through samples after the cross-linking procedure showed complete protein depletion, indicating efficient antibody coupling to Sepharose A beads (Figure 21).

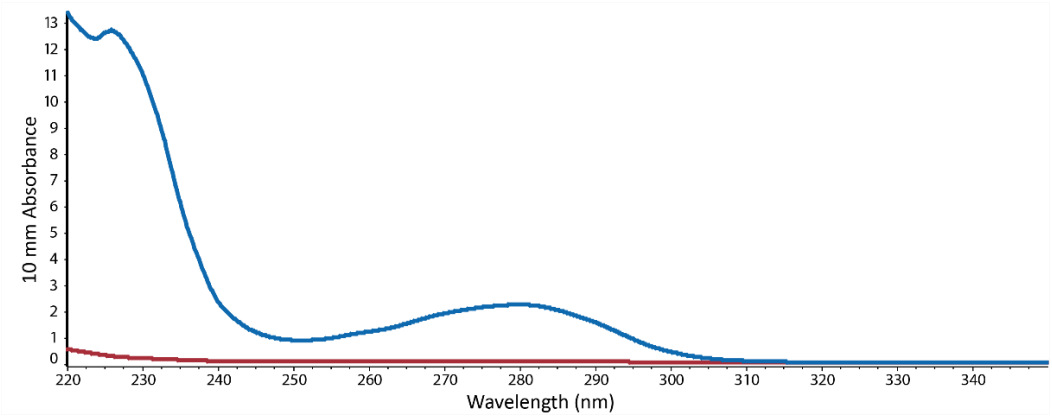


Figure 21 | Quality control of chemical cross-linking of W6/32 HLA-I antibodies. *W6/32 antibody concentration of aliquots measured before (blue) and after (red) chemical cross-linking to Sepharose A beads.*

3.5.2 HLA-I immunopurification (HLA-I IP)

Next, I used the W6/32-coupled Sepharose beads to purify HLA-I molecules from protein lysates. I loaded the beads on the column, added protein lysates, and let it run through by gravity. Afterwards, I determined the HLA-I content of the flow-throughs to examine the efficacy of HLA-I binding to the beads. In K562 cells, only faint bands of HLA-I were detected in flow-through samples, whereas the signal was much stronger in the positive control (Figure 22A). Similarly, the HLA-I content in flow-through of RPE-1 samples was drastically depleted compared to the positive control (Figure 22B). Notably, all samples showed partial protein degradation since the workflow did not allow sample processing directly after cell lysis. The western blot results demonstrated that the HLA-I molecules were successfully captured by the W6/32-cross-linked beads. Still, a small amount of HLA-I was still detected in all flow-through samples, which suggests that the amount of input protein oversaturated the columns. Taken together, the strategy was successful in extracting HLA-I molecules from protein lysates. Subsequently, HLA-I-bound peptides were eluted, and their sequences were analyzed by mass spectrometry.

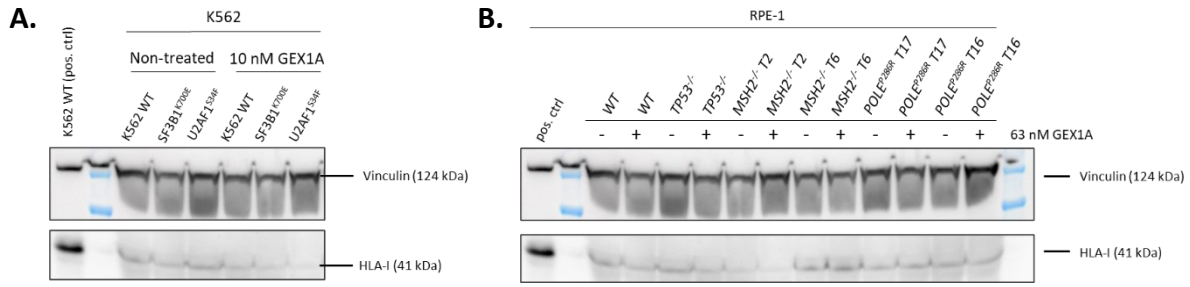


Figure 22 | HLA-I content of cell lysates after HLA-I IP. *Western blot detection of HLA-I in flow-through samples of* ***(A)*** *K562 cell lines and* ***(B)*** *RPE-1 cell lines. Positive control samples were not subjected to HLA-I IP.*

3.6 Peptide identification by PEAKS search

Mass spectrometry measurements provide raw peptide spectra which subsequently need to be converted into peptide sequences. Since my experimental approach aimed to identify non-canonical peptides, it was essential to choose a strategy that did not rely on canonical reference databases to match the experimental mass spectrometry data. Otherwise, it would only be possible to detect annotated peptide sequences but not novel peptides originating from aberrant alternative splice events. Thus, I chose PEAKS Studio X (Bioinformatics Solutions Inc., Waterloo, Canada) because it uses *de novo* peptide sequencing to determine a peptide sequence without using a reference database (Tran et al., 2019). Here, peptide sequences were first derived from mass spectra by *de novo* peptide sequencing. The obtained *de novo* peptide sequences were then searched in common or custom reference library databases to identify matching peptide entries.

3.6.1 Peptide identification quality control

For PEAKS analysis, I followed the 'decoy-fusion' approach, which allowed me to obtain a very accurate false discovery rate (FDR) for peptide identification. The decoy database contained random peptide sequences retrieved from the reference database. The FDR value indicated the proportion of detected decoy sequences during peptide identification via peptide-spectrum matches (PSMs) (Figure 23A). The proportion of detected decoy sequences grew exponentially, and the peptide identification was abrupted after 281,647 discovered PSMs included 5% sequences from the decoy database.

The quality of identified peptides was represented by the PEAKS peptide score (-10lgP). Target PSMs exhibited a high PEAKS score, whereas lower-scoring PSMs contained a larger proportion of identified decoy sequences (Figure 23B). I further confirmed the quality of PSMs by the deviation of precursor mass. Target PSMs exhibited a precursor mass error between -2.5 and 2.5 parts per million (ppm) which increased with lower scores (Figure 23C). Here, peptides exhibiting high PEAKS scores were centered around the mass error of zero, whereas lower-scoring peptides scattered similar to decoy peptides. Thus, the PEAKS score reflects the quality of identified peptides since low-scoring peptides exhibited similar properties to decoy peptides.

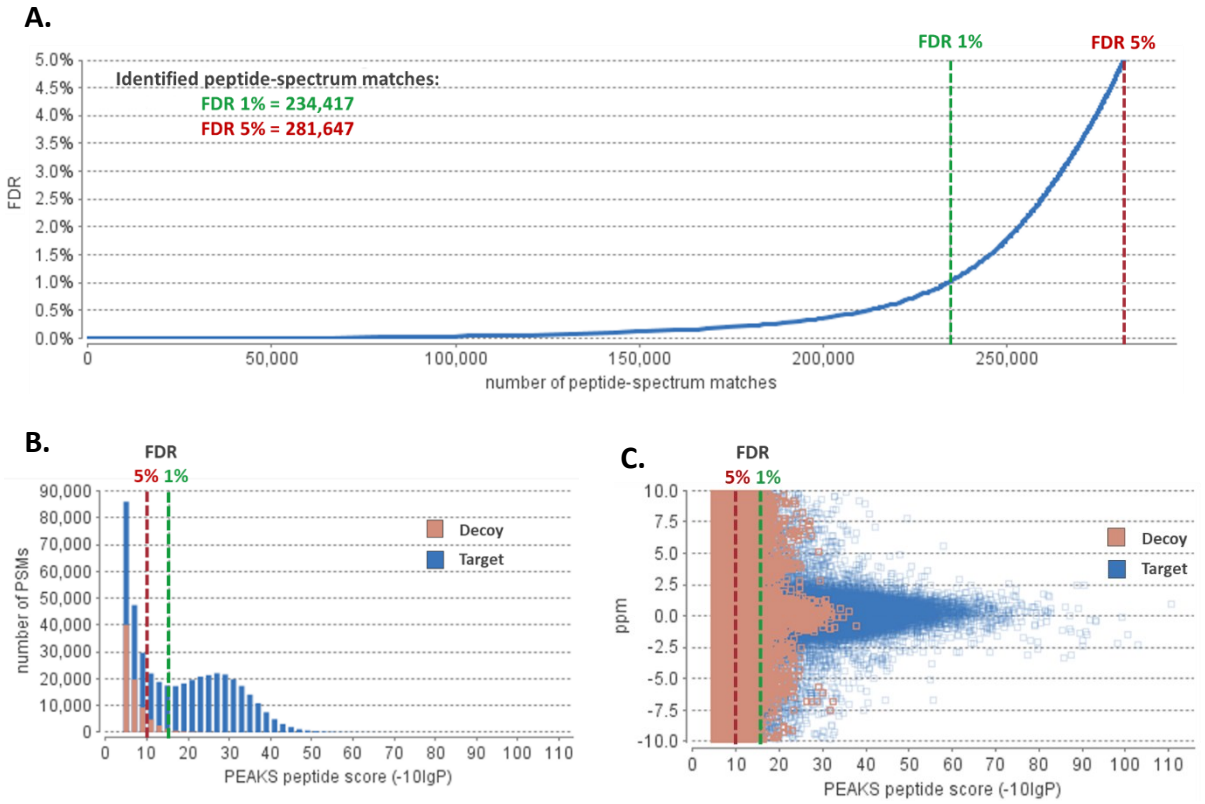


Figure 23 | Impact of FDR cut-offs on the identification of PSMs from PEAKS search. (A) *FDR curve of PSMs with X axis describing the number of PSMs and Y axis showing the corresponding FDR.* ***(B)*** *Distribution of PEAKS peptide score of PSMs (-10lgP)* ***(C)*** *Scatter plot of PEAKS peptide score vs. precursor mass error (ppm).*

To identify the best FDR threshold, I exported the resulting PSMs from the search against the UniProt database with 5% and 1% FDR value cut-offs. I plotted the number of identified peptides against the PEAKS peptide score (Figure 24A+B), resembling the distribution pattern observed for PSMs (Figure 24B). Here, I noticed that the number of peptides drastically increased with lower FDR cut-off values.

To validate identified peptides obtained from different FDR cut-offs, I determined peptide hydrophobicity indices as this is an established and orthogonal parameter to evaluate the quality of peptide identification (Rolfs et al., 2019). The 13,066 peptides identified from the PSMs with a 1% FDR exhibited a tight correlation (R^2 = 0.95) between peptide hydrophobicity and their observed retention time during mass spectrometry measurements (Figure 24C). Only a few dozen peptides did not follow this correlation. By relaxing the FDR cut-off to 5%, additional 4,864 peptides were obtained. However, this increase in peptide numbers was accompanied by a poorer correlation between predicted hydrophobicities and retention times of peptides (R^2 = 0.85). Hundreds of peptides did not correlate, indicating incorrect peptide identification (Rolfs et al., 2019). These findings suggested that the substantial increase in the number of identified peptides obtained from relaxed FDR thresholds (Figure

24B) partly resulted from the identification of false-positive peptide sequences. Consequently, the 1% FDR threshold was applied for multi-round searches to only assign PSMs with a high PEAKS score to reference entries from the UniProt database. I then searched the remaining unassigned PSMs against the custom reference libraries, as later explained in section 3.6.5.

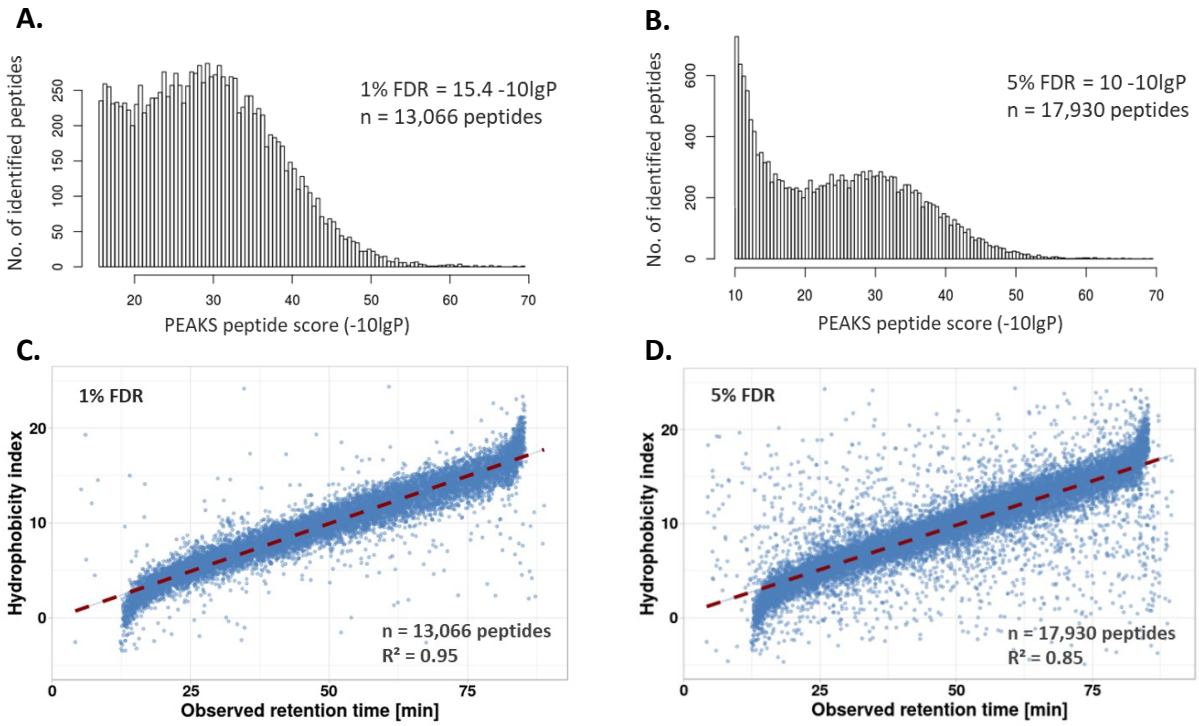


Figure 24 | Impact of FDR cut-offs during PEAKS search on the quality of identified peptides. (A) *Distribution of peptides by PEAKS score identified with 1% FDR (n = 13,066).* ***(B)*** *Distribution of peptides by PEAKS score identified with 5% FDR (n = 17,930).* ***(C)*** *Scatter plot of peptide retention time vs. hydrophobicity index for peptides identified with 1% FDR (R^2 = 0.95).* ***(D)*** *Scatter plot of peptide retention time vs. hydrophobicity index for peptides identified with 5% FDR (R^2 = 0.85).*

3.6.2 GibbsCluster analysis

GibbsCluster analysis is another quality control step to confirm the accuracy of identifying peptides from PEAKS search. Here, I evaluated whether sequence motifs of identified peptides matched the predicted binding motifs of HLA-I receptors. Every HLA allele subtype has a distinct binding pattern that peptides need to fulfill to be able to be presented as ligand on HLA receptors (Figure 12C). HLA-I ligands have a length between 8-11 amino acids and are typically 9-mers (Gfeller et al., 2018a). Peptides interact with the HLA binding pocket through crucial residues at positions 2 or 3 and/or 9, leading to a characteristic binding sequence motif.

I clustered peptides based on sequence similarity and the three largest groups were represented by sequence logos (Figure 25). The top three sequence logos identified for K562 samples matched the predicted binding motifs of HLA*A31:01, HLA*B:40:01, and HLA*C05:01, which all three matched the HLA-I subtypes detected by HLA haplotyping (Figure 25A). Similarly, being binders of HLA*A01:01, HLA*A02:01, and HLA*B07:02, peptide motifs of the three largest clusters of RPE-1 samples matched the RPE-1 haplotype (Figure 25B). Hence, GibbsCluster analysis supported the accuracy of PEAKS peptide identification from PSMs since detected peptide-binding patterns were in line with predicted sequence motifs.

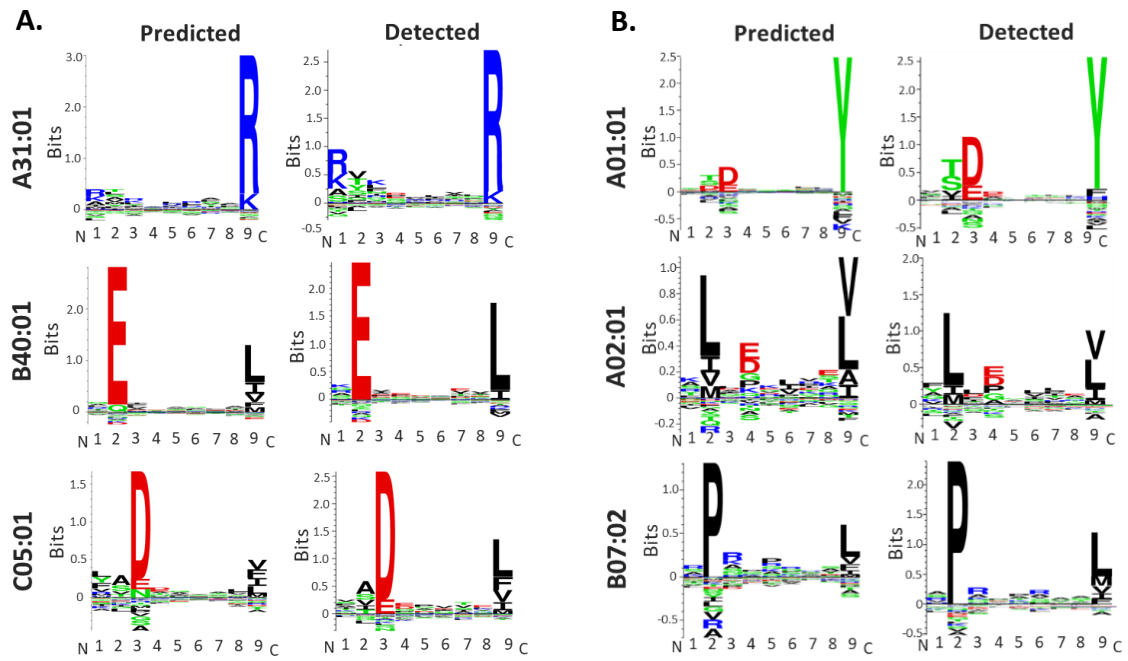


Figure 25 | HLA-I binding motifs of identified peptides revealed by GibbsCluster analysis. (A) *Predicted binding motifs of top three clusters of peptides identified from K562 cells compared to predicted binding motifs of given K562 HLA-I subtypes.* ***(B)*** *Predicted binding motifs of top three clusters of peptides identified from RPE-1 cells compared to predicted binding motifs of given RPE-1 HLA-I subtypes.*

3.6.3 HLA-I binding prediction

After identifying common peptide motifs from the peptide datasets, I performed HLA-I binding prediction to model binding of identified peptides to HLA-I receptors. Here, NetMHCpan uses artificial neural networks to match short peptide sequences to binding profiles of specific HLA-I subtypes of a cell line.

Indeed, I was able to confirm that the majority of identified peptides were predicted HLA-I binders (Figure 26). For RPE-1 WT samples, 83.1% of total peptides (1% FDR) were matched to HLA-I subtypes of RPE-1 cells, while only 16.9% of peptides were non-predicted binders (Figure 26A). Changing the FDR threshold to 5% resulted in the identification of additional 4,864 peptides, but also the proportion of non-predicted binders increased to 27.2% (Figure 26B). From the additional 4,864 peptides, only 2,188 peptides (45%) were predicted binders. Accordingly, more than half of the additionally identified peptides were non-predicted binders, suggesting the detection of false-positive peptide sequences. Similar results were obtained for K562 WT samples. Here, 89.8% of peptides identified with 1% FDR were predicted HLA-I binders (Figure 26C), whereas only 82.1% of peptides identified with 5% FDR were predicted HLA-I binders (Figure 26D).

Notably, I detected a higher proportion of predicted binders in K562 samples compared to RPE-1 samples. This could be explained by the different haplotypes of the cell lines. The accuracy of HLA-I binding prediction heavily depends on the HLA-I subtype. For example, RPE-1 cells express the HLA*A:02:01 subtype, one of the most frequent alleles found in humans (Figure 26B). Accordingly, binding prediction can draw on many data and is thus more stringent than for other alleles. In addition, HLA*A:02:01 is known to present a variety of different antigens. Given the 2% rank threshold, peptides needed a minimum binding constant (K_D) of 272 nM to be identified as predicted binder of HLA*A:02:01, which was very stringent compared to all other alleles (Figure 26F). For instance, peptides were predicted binders of HLA*A:01:01 when exhibiting a binding constant of 6668 nM or lower – almost a 25-fold difference in binding affinity (Figure 26E). This highlights the limitations of the computational HLA-I binding prediction approach. Consequently, also non-predicted binders could be HLA-I ligands.

Nevertheless, HLA-I binding prediction provided a confident indication about the quality of identified peptides. Concluding from the results, I chose the 1% FDR threshold for PEAKS peptide identification from UniProt, confirming the previous observations made for hydrophobicity correlations (Section 3.6.1).

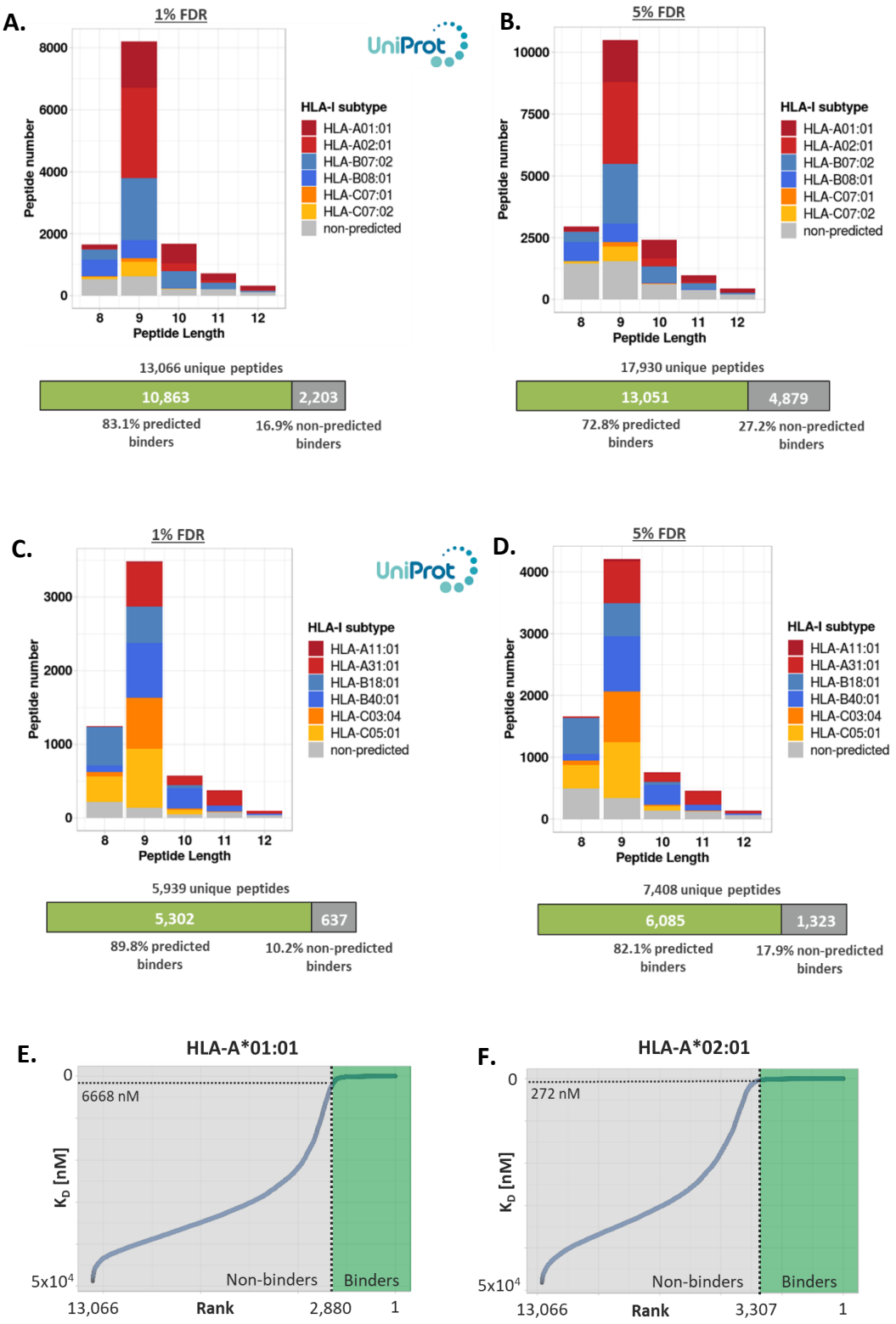


Figure 26 | Computational HLA-I binding prediction of identified peptides. (A) *Distribution of peptide length and HLA-I subtype clustering of peptides identified from RPE-1 mass spectra data searched against UniProt with 1% FDR cut-off and* ***(B)*** *5% FDR cut-off.* ***(C)*** *Distribution of peptide length and clustering to HLA-I subtypes of peptides identified from K562 mass spectra data searched against UniProt with 1% FDR and* ***(D)*** *5% FDR.* ***(E)*** *Ranking of predicted peptide binding affinities for the RPE-1 HLA-A*01:01 subtype.* ***(F)*** *Ranking of predicted peptide binding affinities for the RPE-1 HLA-A*02:01 subtype. K_D = Dissociation constant.*

3.6.4 Similarity of immunopeptidomes between replicates, sample groups, and datasets

To provide information about peptide diversity and how robust peptides were detected across different replicates, I explored the composition of immunopeptidomes of sample groups. Given the limited sensitivity of mass spectrometry, only a certain number of mass spectra per time can be identified from peptide samples. For immunopeptidomics, it is especially challenging to detect a wide range of individual peptides because HLA-I-bound peptides are very similar in their sequence length and amino acid composition.

To explore the diversity of recovered immunopeptidoms, I processed the PEAKS search results and plotted the numbers of unique identified peptides of each replicate (Figure 27).

In K562 replicates, I identified 2,791 peptides on average (Figure 27A). A variance of a few hundred peptides was observed between replicates of a sample group. *SF3B1* sample groups exhibited the highest variance. In one replicate of the GEX1A-treated *SF3B1* mutant sample group, there were almost twice as many peptides identified as in the other two replicates (Figure 27A). Despite using similar cell numbers as input for the HLA-I IP, the peptide diversity of *SF3B1* sample groups was lower compared to WT or *U2AF1* sample groups.

RPE-1 cell lines generally exhibited a greater immunopeptidome diversity than K562 cell lines (Figure 27B). Here, I detected 3,759 unique peptides in cancer-like RPE-1 sample groups on average. The RPE-1 WT sample groups with around 6,000 unique peptides obtained per replicate showed exceptionally high immunopeptidome diversity. The variance between replicates of *MSH2*$^{-/-}$ and *POLE*^{P286R} cell lines was higher. Especially the number of unique peptides identified from replicates of the *MSH2*$^{-/-}$ T6 sample groups differed by thousands of peptides (Figure 27B). Overall, peptide identification yielded different numbers of peptides both between sample groups and between replicates.

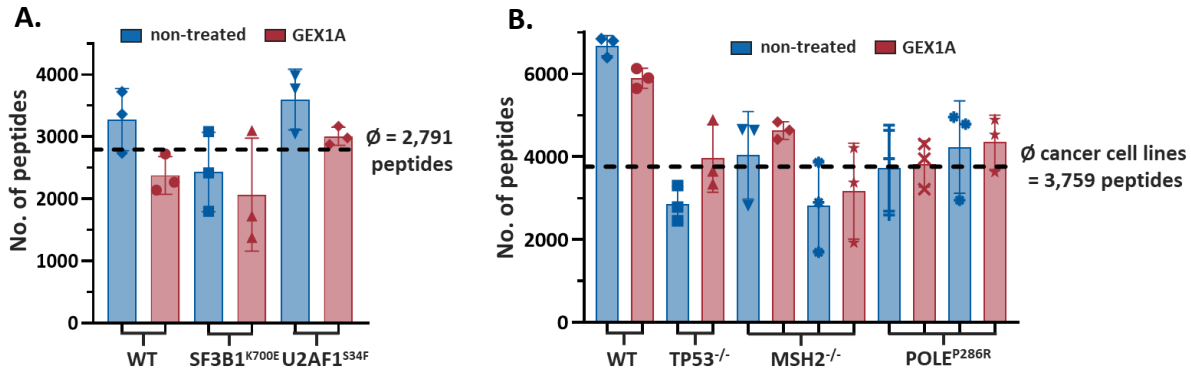


Figure 27 | Number of unique peptides identified per sample from immunopeptidomics. (A) *Number of HLA-I-binding peptides of K562 cell lines identified from independent biological and technical replicates (n = 3).* ***(B)*** *Number of HLA-I-binding peptides of RPE-1 cell lines identified from independent biological and technical replicates (n = 3). Dashed lines represent the average number of peptides found in cancer(-like) cell lines.*

To evaluate how the different sample sizes impacted the similarity of independent biological replicates, I compared the recovered immunopeptidomes in a Venn diagram. The overlap of non-treated K562 WT replicates revealed that 47.4% of all unique identified peptides were discovered in all three replicates (Figure 28A). Similarly, 45.3% of all peptides identified from non-treated RPE-1 WT samples were detected in all three replicates (Figure 28B). The major factor influencing replicate overlap was the variance in the number of detected peptides. For example, non-treated *MSH2*$^{-/-}$ T6 samples only shared 28.6% of unique peptides identified in all replicates because many peptides were not detected in replicate #3 with exceptionally low peptide numbers (Figure 28C). These findings indicate that similar peptide numbers were critical for a high similarity between immunopeptidomes.

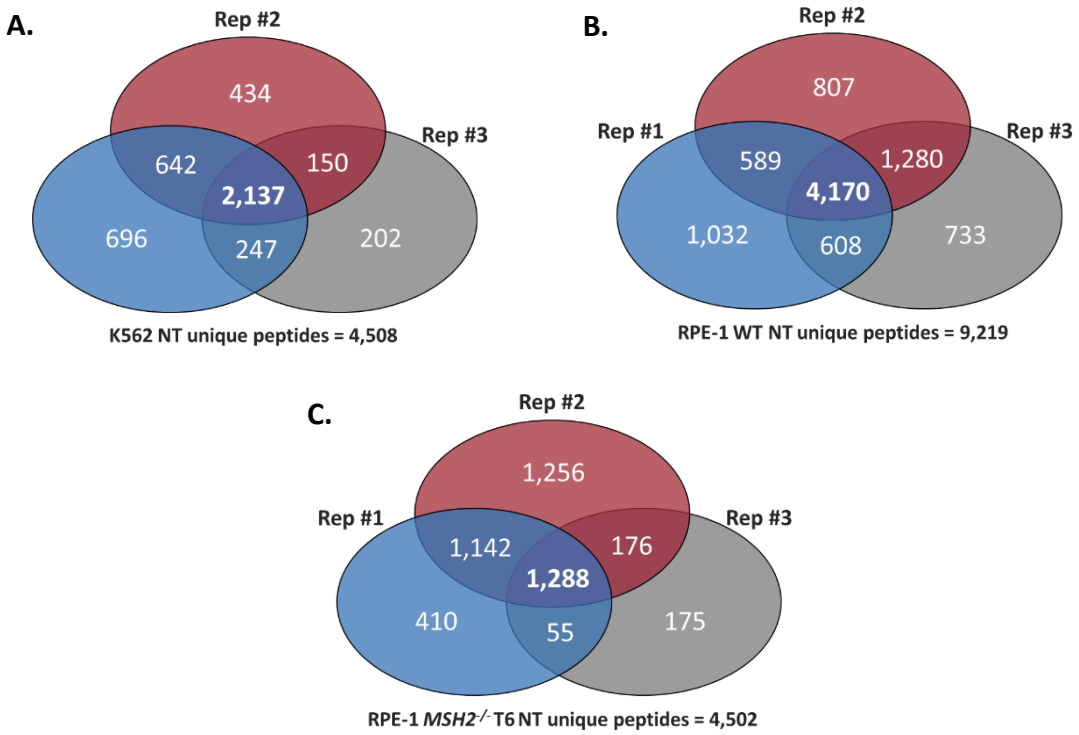


Figure 28 | Replicate overlap of identified peptides recovered by immunopeptidomics. (A) *Overlap of K562 WT NT replicates.* ***(B)*** *Overlap of RPE-1 WT NT replicates.* ***(C)*** *Overlap of RPE-1 MSH2*$^{-/-}$ *T6 NT replicates. Venn diagrams illustrate replicate overlap and do not reflect actual proportions. Peptide overlaps of all 16 sample groups are displayed in supplement figure 1 in the appendix of this thesis.*

For a better visualization of the differences between individual replicates, I compared the immunopeptidome of all samples by hierarchical clustering. I considered peptides found in all replicates as background and eliminated from this analysis. This comprised 679 of total 13,066 identified peptides. Then, I normalized replicates by their size factors and the sample differences were visualized in a distance matrix (Figure 29).

Notably, the WT sample groups clustered together, while the GEX1A treatment had no clear impact on clustering of the samples. Instead, the similarity between samples largely depended on the number of identified peptides. Replicates with high peptide diversity showed a high similarity to one another. In contrast, hierarchical clustering revealed that both samples of *MSH2*$^{-/-}$ T6 replicate #2, the two samples with the lowest number of identified peptides, were very distinct from all other samples (Figure 29). These results suggest that a comparable number of peptides is required for robust peptide detection across samples.

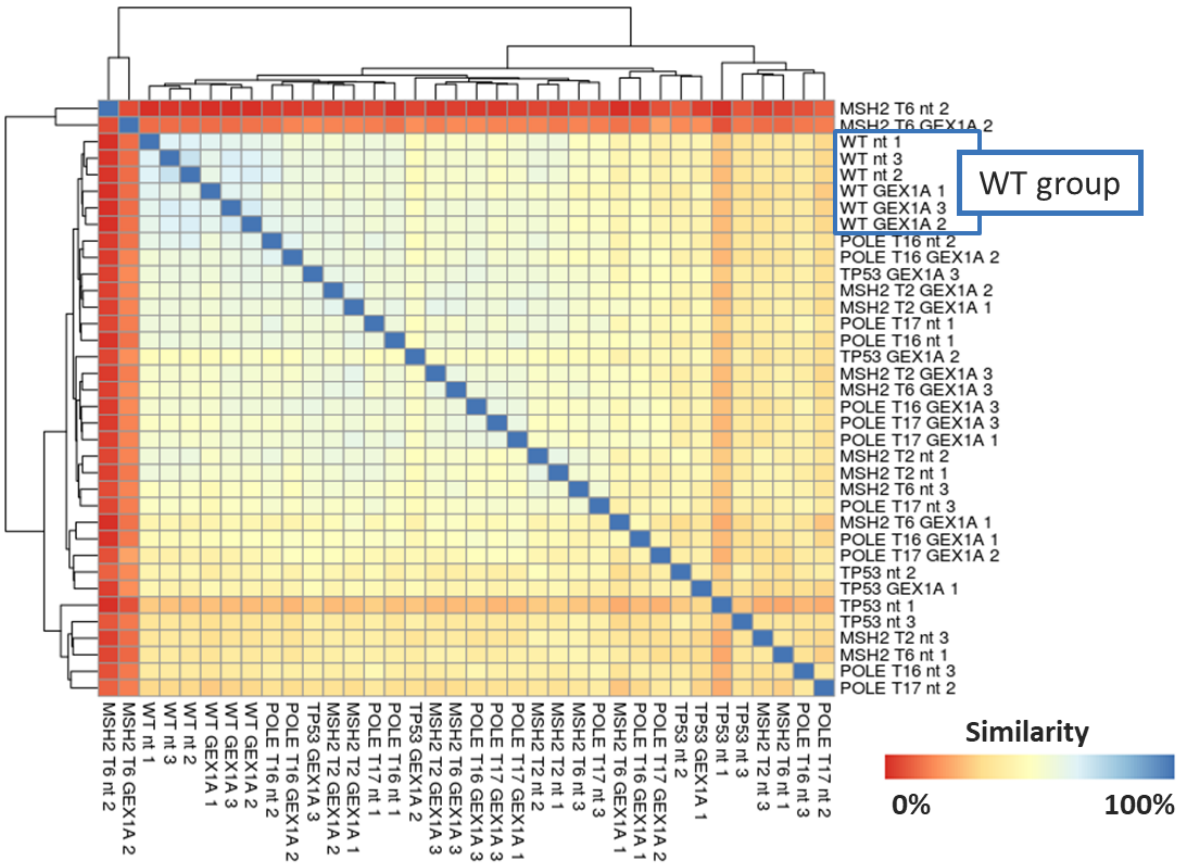
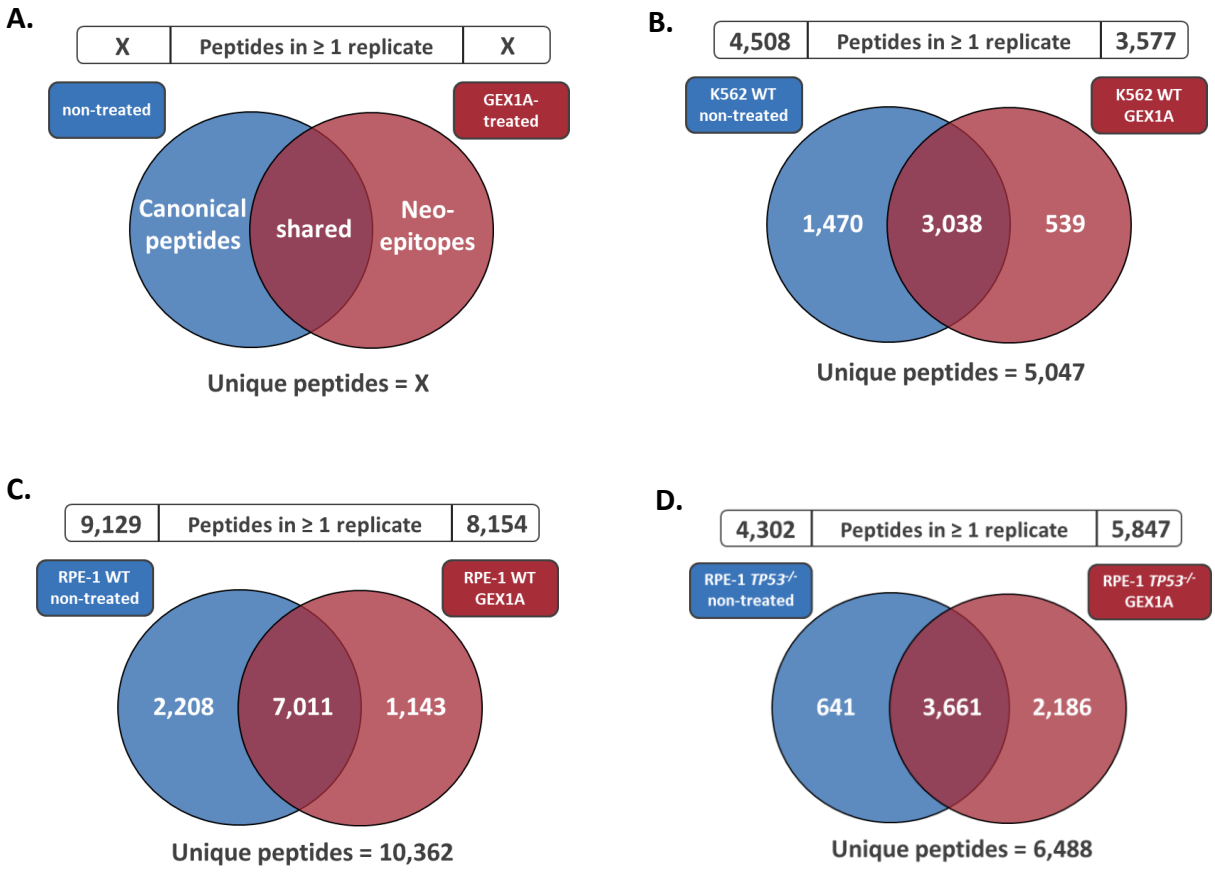


Figure 29 | Similarity of immunopeptidomes of different RPE-1 sample groups. *Peptides of each sample groups were normalized by sample size factors and compared by hierarchical clustering. Peptides shared among all sample groups were considered as background.*

In addition to comparing individual replicates, I explored the proportion of peptides shared between treatment conditions (Figure 30A). Combined, I identified 5,047 unique peptides in non-treated and GEX1A-treated K562 WT cells (Figure 30B). More than 60% of all unique peptides were shared between the two sample groups. Notably, 84.9% of peptides found in GEX1-treated samples were also identified in the non-treated sample condition. Vice versa, only 67.4% of peptides from non-treated samples were detected in the GEX1A-treated condition because the non-treated sample group contained almost thousand more peptides. In RPE-1 WT samples, I identified 10,362 unique peptides from which 2 out of 3 peptides (67.6%) were shared between non-treated and GEX1A-treated sample groups (Figure 30C). Again, I detected more unique peptides in the non-treated condition and consequently 86% of peptides from treated samples were found in non-treated samples, but only 76.8% vice versa. In general, the overlap between treatment conditions was substantial but there were also hundreds of peptides exclusively found in the GEX1A-treated condition (Figure 30D-H). Notably, these treatment-specific peptides were of special interest represent the pool of potential neoepitope candidates.



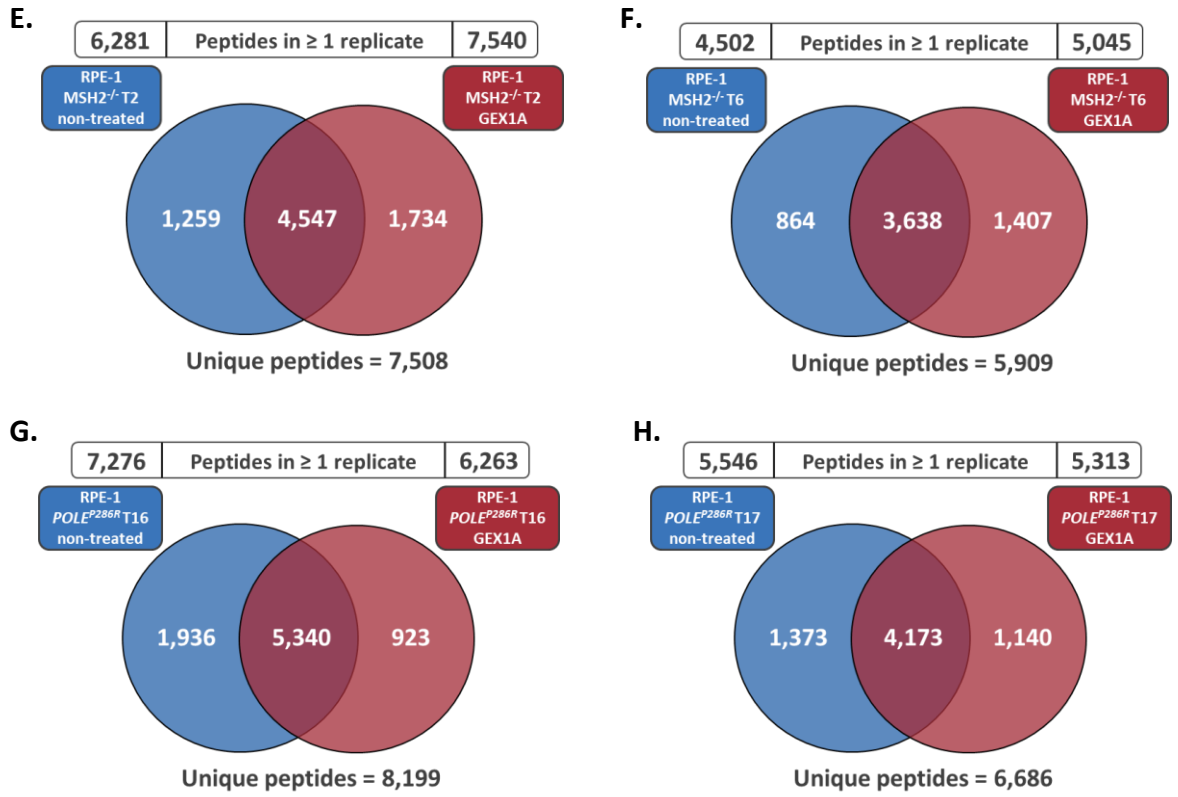


Figure 30 | Overlap of identified peptides between treatment conditions. (A) *Peptides not shared between conditions were considered 'canonical peptides' whereas peptides only present in the treated condition were considered 'neoepitopes'.* ***(B)*** *Overlap of K562 WT non-treated and GEX1A-treated sample groups.* ***(C)*** *Overlap of RPE-1 WT non-treated and GEX1A-treated sample groups.* ***(D-H)*** *Overlap of non-treated and GEX1A-treated sample groups of RPE-1 cell lines harboring mutations and defects in the DNA repair pathway. Venn diagrams illustrate replicate overlap and do not reflect actual proportions.*

To evaluate the specificity of recovered immunopeptidomes for a cell line, I explored the overlap of discovered peptides between the K562 and the RPE-1 dataset. Notably, only 77 (0.7%) out of total of 10,993 unique peptides were shared between K562 and RPE-1 WT samples (Figure 31A). The tremendous difference in the immunopeptidomes between these cell lines can be explained by their different haplotypes, resulting in the presentation of peptides with very different sequence motifs (Figure 31B+C). This observation highlights that the HLA-I IP approach facilitated the accurate recovery and identification of peptides specific for the haplotype of a respective cell line.

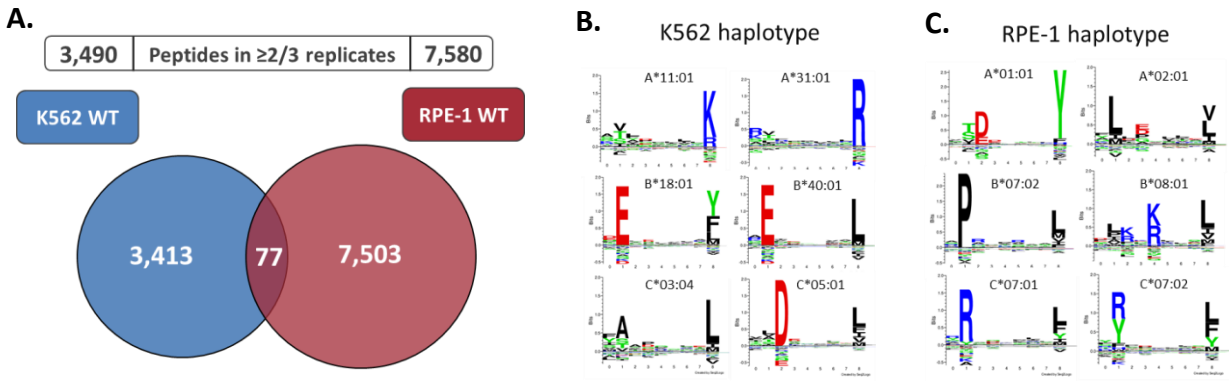


Figure 31 | Differences in immunopeptidome compositions between K562 and RPE-1 datasets. (A) *Overlap between identified peptides in K562 WT and RPE-1 WT samples.* ***(B)*** *HLA-I binding profiles of the K562 haplotype.* ***(C)*** *HLA-I binding profiles of the RPE-1 haplotype.*

3.6.5 Custom reference databases

Canonical reference databases such as UniProt only contain annotated protein information. However, the investigated samples likely included novel peptide sequences originating from non-annotated splice events caused by spliceosomal perturbations. Thus, custom references libraries were generated based on matching RNA-seq information. These additional reference databases also contained peptide sequences produced by non-canonical splice events, complementing the protein information found in UniProt.

Two approaches were followed to design sample-specific reference databases. First, the TransDecoder workflow included the assembly of a *de novo* proteome based on RNA-seq data. Second, the splice junction database (SJDB) contained all short junction-spanning peptides that could be produced from splice junctions identified from RNA-seq data.

Searching mass spectra against different reference databases yielded different numbers of unique peptides (Figure 32A). PEAKS search against UniProt, the most comprehensive reference of the three databases, detected 9,142 unique peptides in RPE-1 non-treated WT samples applying a 1% FDR threshold. Using the TransDecoder database, 6,831 unique peptides were identified. The PEAKS score threshold (-10lgP) for the 1% FDR cut-off slightly increased, which could be explained by the different library sizes. Consequently, the PEAKS score threshold of the SJDB search was extremely high because this library was limited to short peptide sequences, which affected quality control during PEAKS search. The SJDB search only delivered 813 peptides using the same settings as for the other approaches. Notably, 91.6% of peptides identified by the SJDB were also found by UniProt (Figure 32B). Even a higher proportion (98.2%) of peptides identified by TransDecoder were also found by UniProt, indicating the similarity of the *de novo* assembled proteome and the UniProt reference. Thus, the custom reference libraries could only slightly increase the peptide search space.

Notably, there was also a substantial overlap between both custom databases. However, only nine peptides were shared from the peptides exclusively detected by the custom libraries. Taken together, these data suggest that the best strategy for peptide matching is to combine all reference libraries in a concatenated, multi-round PEAKS search to allow peptide identification from the largest possible search space.

A.

Reference database	-10lgP for 1% FDR	Identified peptides
UniProt	15.4	9,142
TransDecoder	17.3	6,831
SJDB	37.2	813

B.

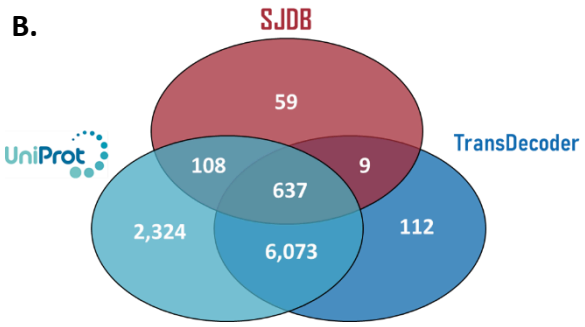


Figure 32 | Comparison of different reference libraries used for peptide identification. (A) *Number of peptides detected in non-treated RPE-1 WT samples identified from different reference databases samples (FDR cut-off = 1%).* ***(B)*** *Overlap of identified peptides from different reference databases. -10lgP = PEAKS peptide score.*

3.6.6 Quality control of custom database peptides

Peptides identified from these multi-round searches also underwent quality control steps. Peptides obtained from PEAKS search with a 1% FDR threshold exhibited a tight correlation ($R^2 = 0.95$) between their observed retention times and their predicted hydrophobicity indices (Figure 33A). Particularly peptides detected by custom reference databases showed an excellent correlation ($R^2 = 0.98$). Relaxing the FDR threshold to 5% during PEAKS search resulted in a weaker correlation between the two parameters ($R^2 = 0.85$), especially for peptides identified from UniProt (Figure 33B). In contrast, the custom library peptides still exhibited a tight correlation ($R^2 = 0.94$), suggesting that low numbers of false-positive peptides were identified from TransDecoder and SJDB databases. Consistently, HLA-I binding prediction confirmed the quality of identified peptides (Figure 33C). From the 12,127 unique peptides, 87.1% of UniProt peptides (1% FDR) were predicted HLA-I binders. Accordingly, 85.2% of custom database peptides (5% FDR) were predicted binders of RPE-1 HLA-I receptors. Consequently, UniProt peptides were filtered by a 1% FDR cut-off, while custom database peptides were selected with a more relaxed 5% FDR threshold. These data highlight that multi-round searches allow accurate peptide identification from custom reference databases.

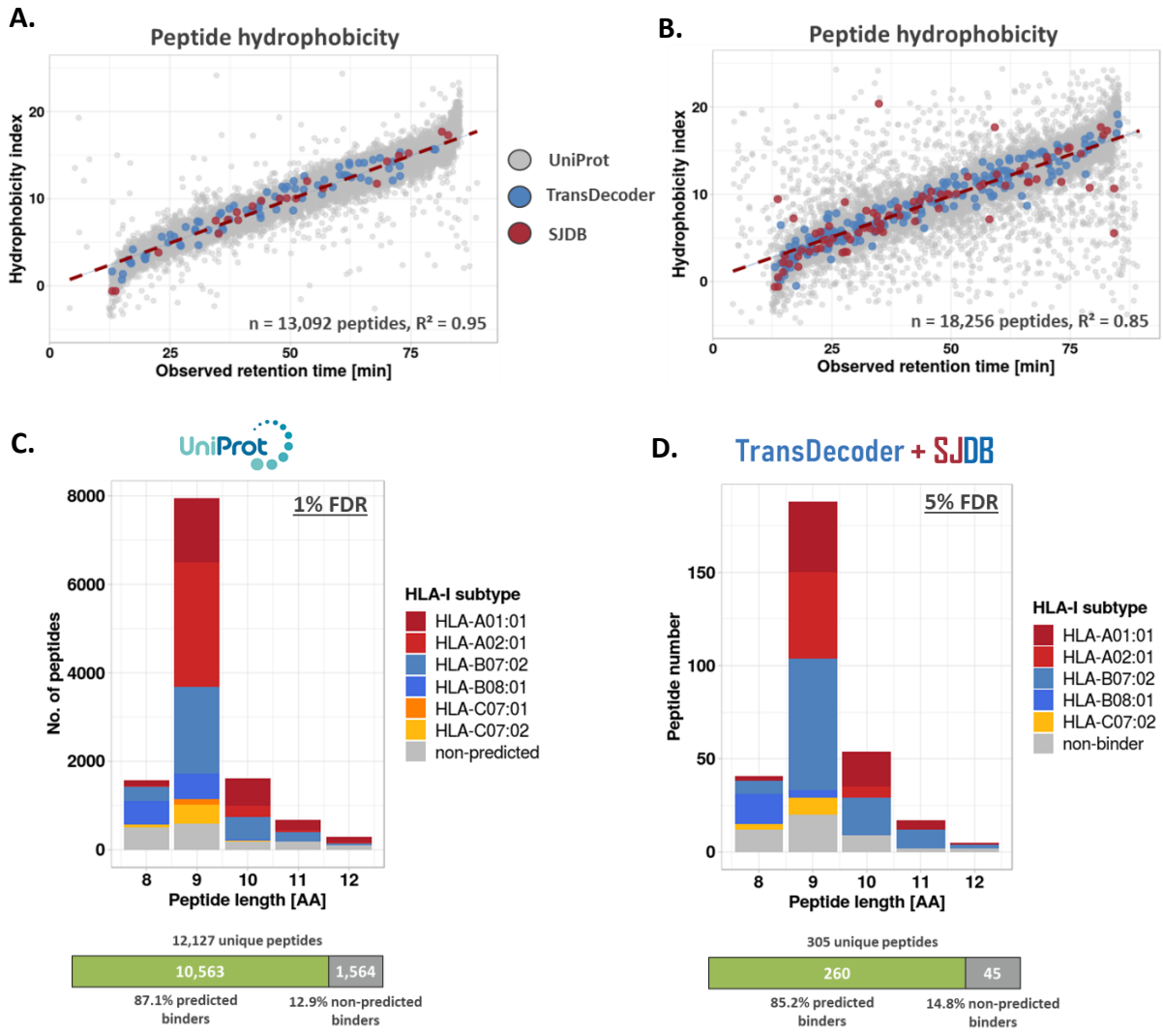


Figure 33 | Quality control of peptides identified from multi-round searches using custom reference databases. ***(A)*** *Scatter plot of peptide retention time vs. hydrophobicity index for peptides identified with 1% FDR cut-off (R^2 = 0.95) and* ***(B)*** *peptides identified with 5% cut-off (R^2 = 0.85).* ***(C)*** *HLA-I binding prediction revealing the distribution of peptide length and HLA-I subtype clustering of peptides identified from UniProt with 1% FDR cut-off and from* ***(D)*** *custom reference databases identified with 5% FDR cut-off.*

3.6.7 Neoepitope candidate identification

Potential neoepitopes candidates were identified by comparing the immunopeptidomes of different sample groups. All unique peptides found in all three replicates of RPE-1 WT non-treated and GEX1-treated samples were visualized in a barcode plot (Figure 34). As observed before, most of the total 10,380 peptides were present in both sample groups (n = 6,955). Peptides exclusively detected in the GEX1A-treated condition were considered as potential neoepitope candidates. To increase robustness, peptides were required to be detected in at least 2 out of 3 GEX1A-treated replicates to be considered neoepitope candidates. Applying these criteria, 366 neoepitope candidates were obtained from RPE-1 WT samples. Additionally, other sample groups were also filtered for neoepitopes yielding several hundreds of peptides that were robustly detected in the GEX1A-specific group (Table 15). Overall, several hundreds of peptides were identified as potential neoepitope candidates.

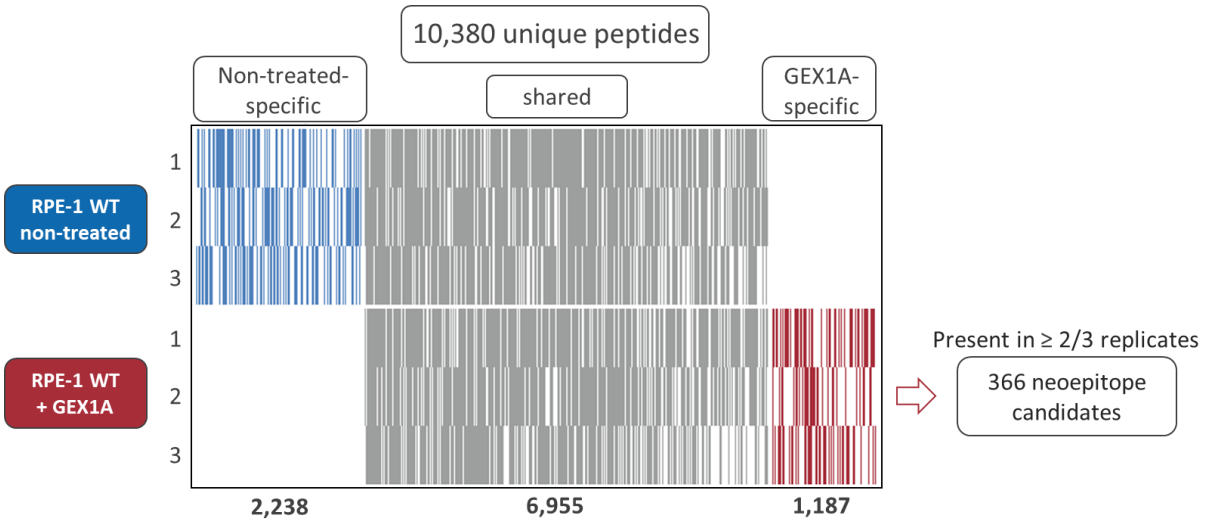


Figure 34 | Bar code plot highlighting the distribution of identified peptides between replicates and sample groups of RPE-1 WT samples. *Each bar represents a single peptide. Shared peptides were found at least once in both conditions, specific peptides were exclusively detected in one condition. For neoepitope identification, GEX1A-specific peptides were filtered for peptides which were detected in ≥ 2/3 GEX1A-treated RPE-1 replicates.*

Table 15 | Neoepitope candidates identified exclusively in ≥ 2/3 GEX1A-treated RPE-1 replicates.

Sample group	WT	TP53$^{-/-}$	MSH2$^{-/-}$ T2	MSH2$^{-/-}$ T6	POLE^{P286R} T16	POLE^{P286R} T17
No. of neoepitopes	366	728	751	382	217	402

3.7 Impact of gene expression on identified peptides

In parallel with the exploration of the immunopeptidome, I analyzed the transcriptome of cells with splicing perturbations. Before I explored the changes in the alternative splicing landscape, I determined how gene expression was affected by GEX1A treatment and how this impacted neoepitope identification.

3.7.1 Correlation between peptide frequency and gene expression

To determine the origin of previously identified neoepitope candidates, I next evaluated how many of them possibly resulted from expression changes of the respective protein. Since protein expression correlates well with protein expression (Liu et al., 2016), I used RNA-seq data sets to determine transcript abundance as a proxy for protein levels. Here, I assigned all peptides detected in the GEX1A-treated condition to their genes of origin. More than 99% of all peptides identified in GEX1A-treated RPE-1 WT samples were successfully matched to expressed protein-coding genes (Figure 35A). The remaining 1% unmatched peptides were assigned to pseudogenes or genes not found to be expressed in GEX1A-treated RPE-1 WT cells. Similar results were obtained when only peptides were taken into consideration that were robustly detected in at least 2 out of 3 replicates Figure 35B). To examine the effect of gene expression on peptide detection, the distribution of gene expression frequency was visualized (Figure 35C). For most highly expressed genes, a matching peptide was found in the immunopeptidomics dataset. The lower the expression, the lower the proportion of genes for which peptides were identified. Again, the same observations were made using a more stringent peptide cut-off (Figure 35D). Notably, for low-abundantly expressed genes ($\log_2 < 5$), only very few peptides were identified, highlighting the limited sensitivity of the mass spectrometry approach.

Moreover, I examined the expression changes of genes assigned to neoepitope candidates of RPE-1 WT samples (Figure 35E). All of the 1,187 neoepitopes identified in Figure 34 were assigned to 1,040 protein-coding genes expressed in RPE-1 WT cells. The expression changes between the GEX1A-treated and the non-treated sample group suggested an enrichment in neoepitope-producing genes in the GEX1A-treated condition. This trend became significant when only the 366 neoepitopes were considered, which were robustly detected in at least 2 out of 3 replicates (Figure 35F). The GSEA results indicate that most peptides that were exclusively detected in the treated condition potentially resulted from gene expression changes in response to GEX1A treatment. The analysis revealed that peptides detected by mass spectrometry mainly originated from abundantly expressed genes. Consequently, many neoepitope candidates likely result from gene expression changes between the sample groups and not from differential alternative splice events.

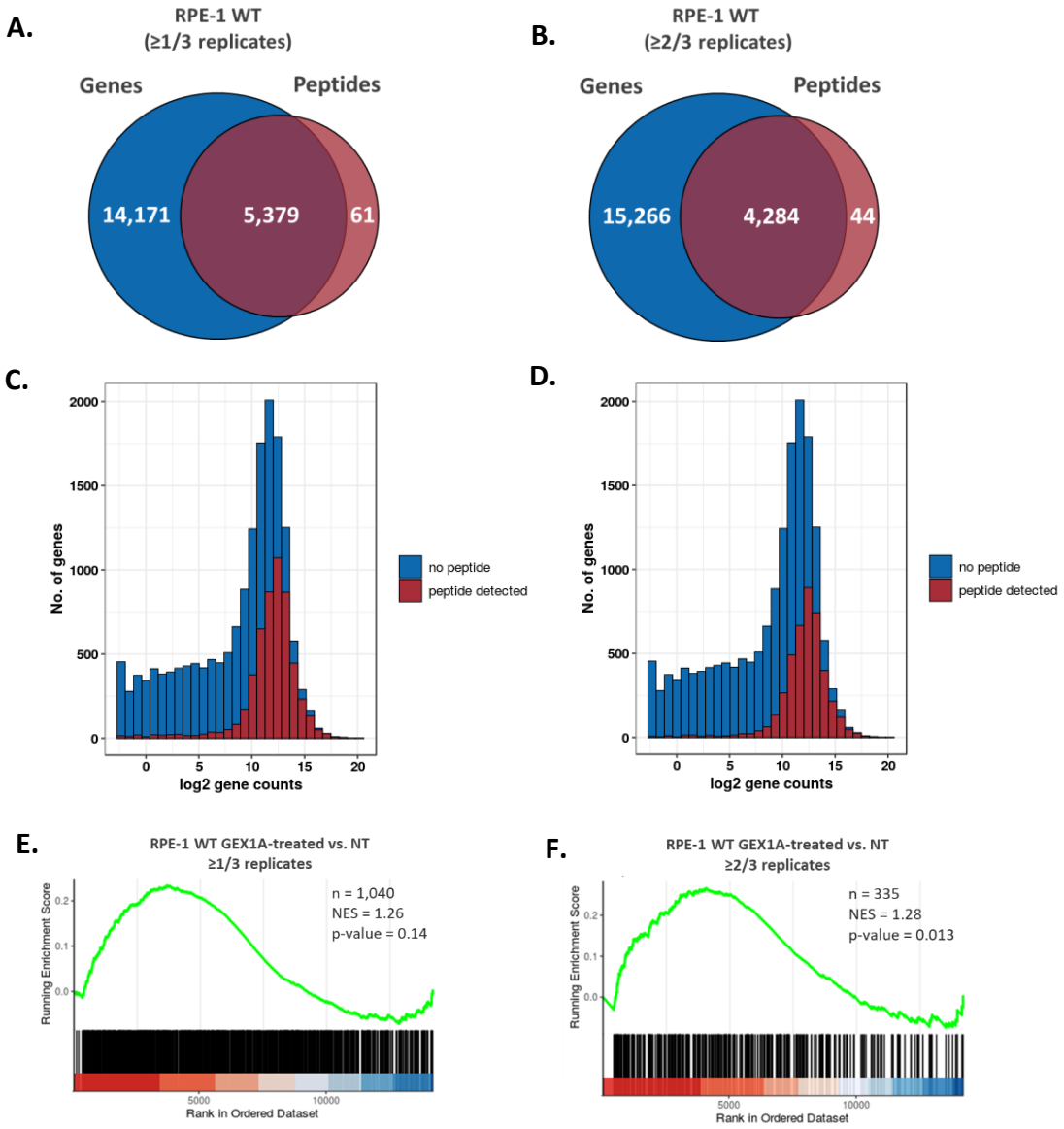


Figure 35 | Gene expression frequency of peptide-associated genes. (A) *Overlap of all identified HLA-I peptides matched to protein-coding genes expressed in the GEX1A-treated RPE-1 WT sample group.* ***(B)*** *Overlap of HLA-I peptides detected in ≥2/3 replicates matched to protein-coding genes expressed in the GEX1A-treated RPE-1 WT sample group.* ***(C)*** *Gene expression frequency of genes for which HLA-I peptides were found (red) and not found (blue) in ≥1/3 RPE-1 WT replicates or* ***(D)*** *in ≥2/3 RPE-1 WT replicates.* ***(E)*** *GSEA of genes assigned to identified neoepitopes present in ≥1/3 RPE-1 WT GEX1A replicates or* ***(F)*** *in in ≥2/3 RPE-1 WT GEX1A replicates.*

3.7.2 Gene expression analysis

The gene expression differences between the sample groups were further analyzed by *DESeq2* to obtain insights into the impact of splicing perturbations on sample group separation. Normalized gene expression data was used to cluster samples in a principal component analysis (PCA). Here, samples from K562 cell lines were clearly separated based on treatment conditions by the first principal component (Figure 36A). The influence of GEX1A treatment on differential gene expression was further investigated by a Gene Set Enrichment Analysis (GSEA). The GSEA revealed that inhibition of the core splicesome resulted in an enrichment of cellular processes associated with transcription and splicing itself (Figure 36B).

Following the observations made for K562 cells, RPE-1 sample groups were clearly separated from each other (Figure 36C). Here, RPE-1 WT cells were compared against RPE-1 *TP53*$^{-/-}$ cells to investigate whether the separation of sample groups was primarily influenced by the genotype or instead by GEX1A treatment. The first principal component clearly separated RPE-1 sample groups by GEX1A treatment (NT vs. GEX1A), while in the second principal component, sample groups were separated by genotype (WT vs. *TP53*$^{-/-}$). The PCA suggested that RPE-1 cells with different genotypes shared a higher similarity in their gene expression profiles than non-treated and GEX1A-treated RPE-1 cells. In line with the results obtained from K562 cells, GSEA between GEX1A-treated and non-treated RPE-1 sample groups revealed a significant enrichment of cellular processes associated with transcription and splicing (Figure 36D).

From these data, it was concluded that gene expression differences between treated and non-treated sample groups need to be considered for differential splicing analysis. Splice events could occur more frequently due to elevated gene expression while their inclusion levels remain unaltered.

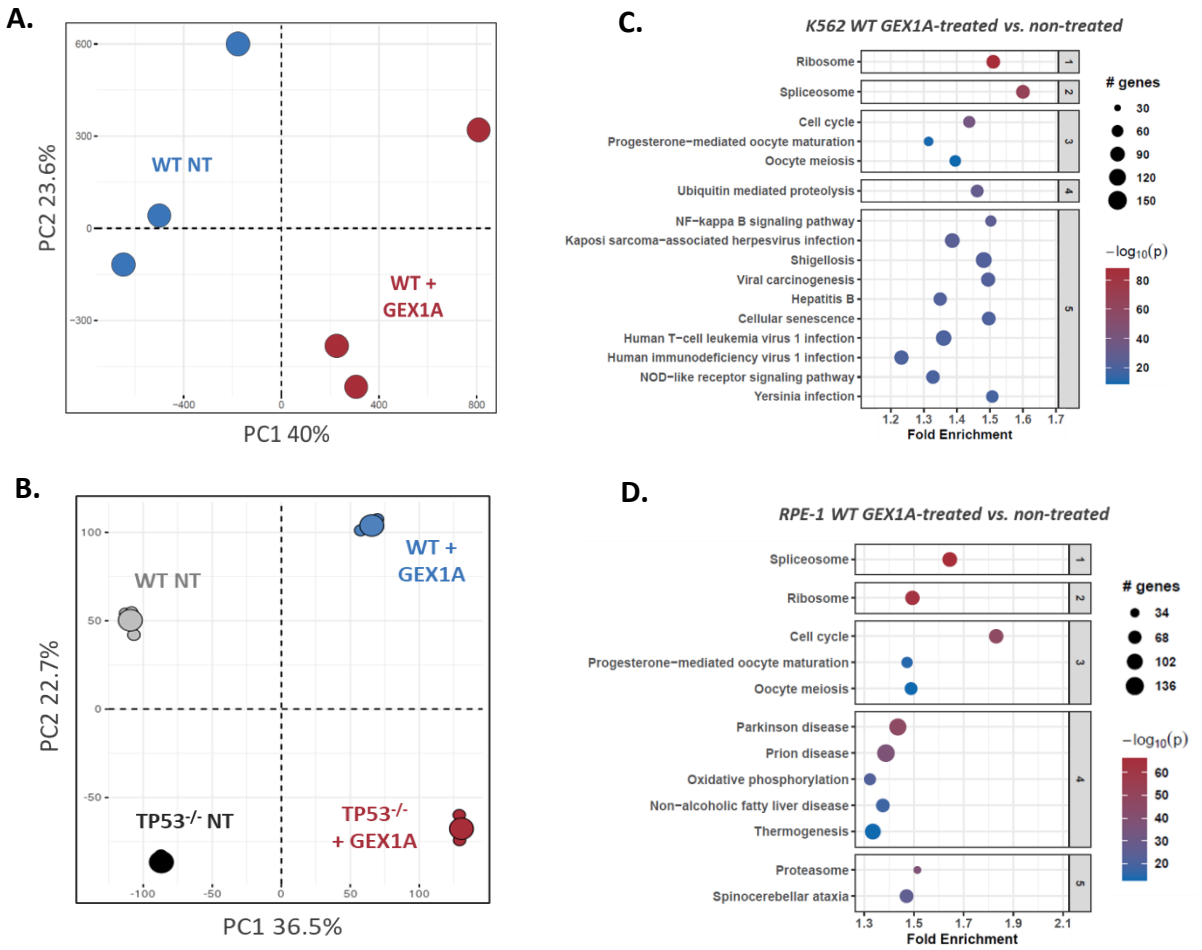


Figure 36 | Effect of splicing inhibition on gene expression profiles. (A) *PCA based on gene expression of non-treated and GEX1A-treated K562 WT cells separated samples based on treatment condition.* ***(B)*** *GSEA of non-treated vs. GEX1A-treated K562 WT cells.* ***(C)*** *PCA based on gene expression of non-treated and GEX1A-treated RPE-1 WT cells separated samples based on treatment condition and genotype.* ***(D)*** *GSEA of non-treated vs. GEX1A-treated RPE-1 WT cells. PCA was performed on normalized gene expression data obtained from DESeq2. GSEA was performed on ranked gene expression differences obtained from DESeq2. The percentage of a PCA variable describes how much it contributes to the variability in the dataset.*

3.7.3 Impact of IFN-γ stimulation on gene expression

In addition, the effect of IFN-γ stimulation on gene expression of K562 cells was investigated. IFN-γ stimulation successfully induced the expression of HLA-I genes (Figure 37A). Except for HLA-C, no baseline gene expression was detected in the absence of IFN-γ. Besides, IFN-γ stimulation affects the expression of a wide range of different genes. Thus, the IFN-γ-treated samples from the K562 cell lines used for HLA-I IP experiments were compared against previously published sequencing data of unstimulated K562 cell lines (Seiler et al., 2018b). PCA revealed the differences in gene expression profiles between the datasets as these were clearly separated by the first principal component (Figure 37B). In the second principal component, non-stimulated samples were separated based on their

SF3B1 genotype. This separation was absent in IFN-γ-stimulated samples indicating that the strong inflammatory gene expression could mask mutation-related expression changes. Indeed, GSEA between non-stimulated and IFN-γ-treated K562 WT samples revealed that the most enriched cellular processes were associated with pro-inflammatory pathways (Figure 37C). These findings were important to consider when interpreting the results of the alternative splicing analysis of K562 cell lines at a later stage. Here, the strong IFN-γ-induced inflammatory response could also mask the differential alternative splicing phenotype between the sample groups.

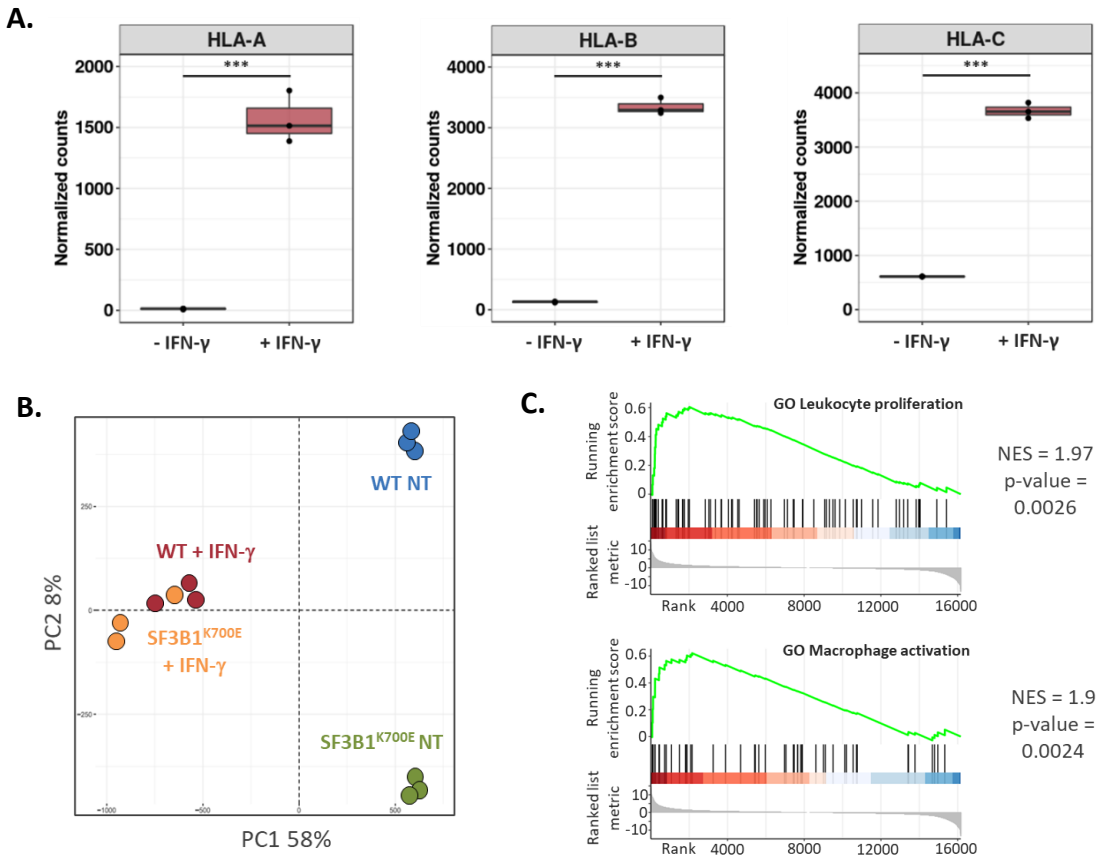


Figure 37 | Effect of IFN-γ stimulation on gene expression of K562 cells. (A) *Box plot of normalized read counts of the HLA genes detected in non-treated and IFN-γ-stimulated K562 WT cells. T-test was applied to evaluate statistical significance: *** = p-value < 0.001.* ***(B)*** *PCA of IFN-γ-stimulated and non-treated K562 cell lines. The percentage of a PCA variable describes how much it contributes to the variability in the dataset.* ***(C)*** *GSEA plots of gene ontology pathways found significantly enriched between IFN-γ-stimulated and non-treated K562 WT cells.*

3.8 Differential alternative splicing analysis

To unravel the effect of spliceosomal perturbation on the alternative splicing landscape, I performed a differential splicing analysis between the sample groups. This was crucial to identify differentially expressed splice events that could produce novel peptide sequences and thus contribute to the neoepitope repertoire of GEX1A-treated cells.

3.8.1 Isoform expression analysis by *Salmon*

For an initial analysis, I explored the changes in the alternative splicing landscape by performing an isoform expression analysis using *Salmon* and evaluated the differences between sample groups by PCA.

First, the effect of IFN-γ stimulation on isoform expression profiles between the K562 sample groups was investigated (Figure 38A). The sample groups were mainly separated in the PC1 due to INF-γ stimulation. The non-stimulated sample groups were separated by genotype, while this effect was absent in IFN-γ-stimulated K562 cells (Figure 38A). These data resembled the results obtained from gene expression analysis, where the strong inflammatory gene expression could mask mutation-related expression changes. In contrast, splicing perturbation induced by GEX1A had a more pronounced effect and affected isoform expression profiles in both IFN-γ-stimulated and non-stimulated K562 cells (Figure 38B).

Apart from K562 cell lines, I could also determine the treatment-induced changes in isoform expression, which separated RPE-1 cell lines in the PC1 (Figure 38C). In addition, the PC2 separated cell lines based on genotype. Notably, GEX1A treatment caused a greater separation between WT and *TP53*$^{-/-}$ cell lines indicating more pronounced isoform expression differences. To set these alterations into relation, RPE-1 samples were compared to K562 samples (Figure 38D). This revealed that the cell type mainly influenced the differences in isoform expression. Splicing inhibition was only able to separate sample groups in the PC2. Nevertheless, GEX1A treatment caused induced pronounced changes in isoform expression, which were further explored by differential splicing analysis.

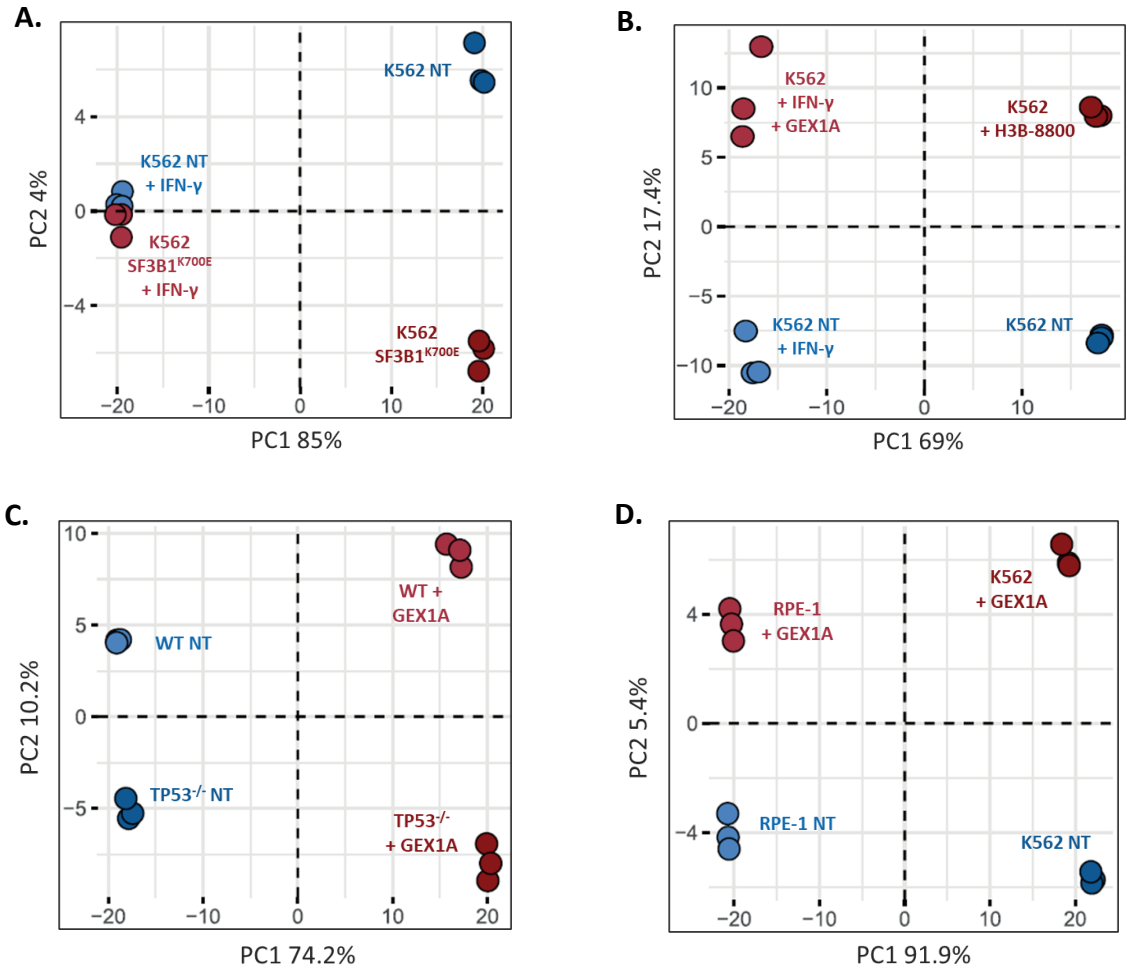


Figure 38 | Differences in isoform expression profiles between cell lines and treatment conditions. ***(A)*** *PCA based on isoform expression of non-treated and IFN-γ-stimulated K562 WT and SF3B1^{K700E} cell lines separated samples based on IFN-γ stimulation.* ***(B)*** *PCA based on isoform expression of non-treated and splicing inhibitor-treated K562 WT cells separated sample groups based on treatment conditions.* ***(C)*** *PCA based on isoform expression of non-treated and splicing inhibitor-treated RPE-1 WT and TP53$^{-/-}$ cell lines separated sample groups based on both GEX1A treatment and genotype.* ***(D)*** *PCA based on isoform expression of non-treated and splicing inhibitor-treated RPE-1 WT and K562 WT cell lines separated sample groups based on both GEX1A treatment and genotype. PCA was performed on normalized isoform expression data obtained from Salmon.*

3.8.2 *rMATS* differential splicing analysis

To explore the different types of splice events induced by perturbation of the splicing machinery, an *rMATS* analysis was performed. Analyzing alternative splicing differences between K562 WT GEX1A-treated vs. non-treated samples resulted in 12,221 significantly differential splice events (Figure 39A). Approximately two-thirds of all events were exon skipping (SE) events, followed by mutually exclusive exon (MXE) events. These two splice types showed a positive value for their average inclusion level difference, indicating that most of these exonic events were included in GEX1A-treated samples. In contrast, the differences in alternative splicing between K562 WT and K562 SF3B1^{K700E} cells were very subtle since only 60 differentially spliced events were detected by *rMATS* analysis (Figure 39B). This finding was in line with the previous observation that IFN-γ stimulation masked mutation-related changes in the alternative splicing landscape of K562 cells (Figure 37).

Accordingly, the splicing differences between different genotypes were much more pronounced between RPE-1 samples. rMATS analysis of RPE-1 *TP53*$^{-/-}$ vs. WT cells delivered 5,083 differentially spliced events (Figure 39D). Still, the most remarkable difference in the alternative splicing landscape was detected between RPE-1 WT GEX1A-treated and non-treated samples resulting in a total of 33,864 splice events (Figure 39C). As observed for K562 cells, approximately two-thirds of identified splice events were exon skipping (SE) events, followed by mutually exclusive exon (MXE) events.

Overall, rMATS analysis revealed that distinct changes in the alternative splicing landscape were achieved by treating cells with the splicing inhibitor GEX1A for a short time of 24 hours. RPE-1 cells were treated with 63 nM GEX1A corresponding to the GR_{75} value, while K562 cells were treated with only 10 nM GEX1A corresponding to the GR_{25} value. The higher number of events identified in RPE-1 samples than K562 samples (33,864 vs. 12,221 events) also demonstrated that a higher dose of GEX1A yielded more pronounced changes in alternative splicing patterns. In both cell lines, exon skipping events were the most frequently occurring splice event, which was detected in response to GEX1A treatment.

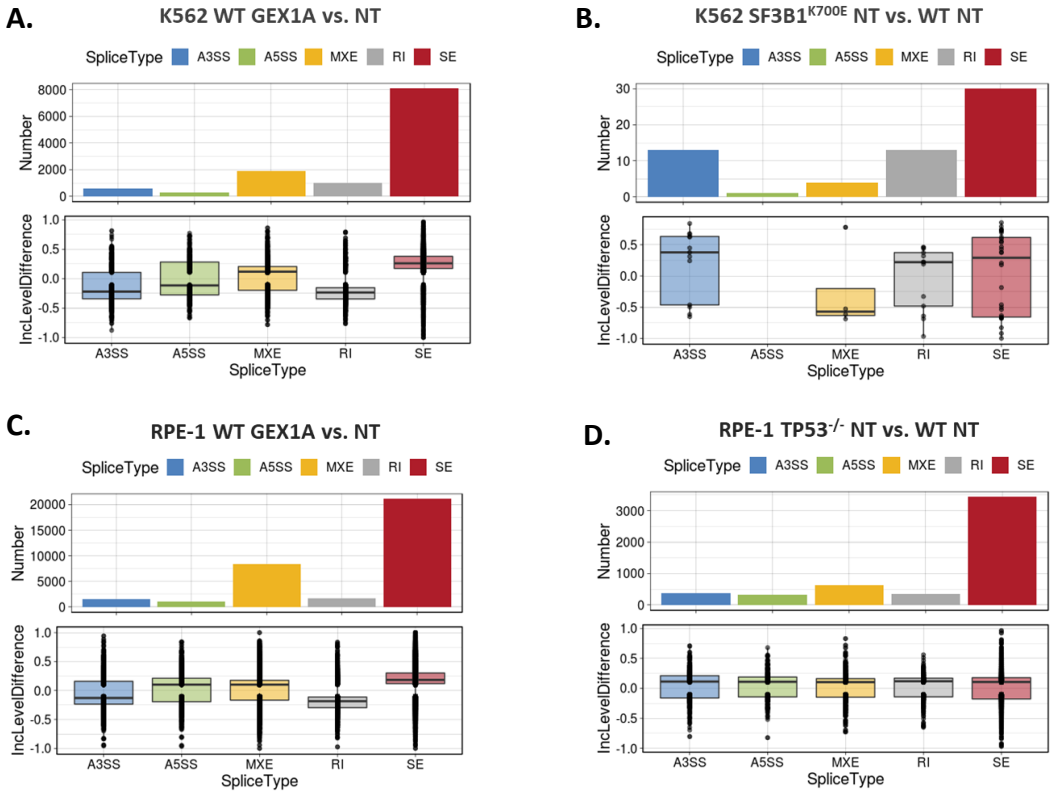


Figure 39 | rMATS analysis demonstrating the impact of splicing inhibition and different genotypes on the alternative splicing landscape. *Number of alternative splice events and inclusion level differences of* ***(A)*** *K562 WT non-treated vs. GEX1A-treated cells,* ***(B)*** *K562 WT vs. SF3B1^{K700E} cells,* ***(C)*** *RPE-1 WT non-treated vs. GEX1A-treated cells,* ***(D)*** *RPE-1 WT vs. TP53$^{-/-}$ cells. Cut-offs: FDR ≤ 0.05 & inclusion level difference ≥ 0.1. A3SS/A5SS = alternative 3'/5' splice site; MXE = mutually exclusive exon; RI = retained intron; SE = skipped exon.*

3.8.3 *JunctionSeq* differential splicing analysis

The *rMATS* analysis only provides insights into inclusion level differences of splice events. However, this strategy is limited to the detection of frequently occurring alternative splice events. Inclusion level differences do not allow for detecting lower abundant splice events if they do not appear in a particular proportion to the canonical event. However, especially transcripts of novel splice events do not occur very frequently. Although these splice events might not be very relevant for the biological function of a gene/protein in the cell, these aberrant transcripts can become relevant if they are translated into non-canonical peptide sequences. To be able to reveal all relevant splice events, *JunctionSeq* analysis was performed. Here, significant splice events were determined by fold change differences of splice junction expression levels between the sample groups instead of inclusion level differences.

In K562 cells, 14,792 significantly differential splice events were identified between non-treated and GEX1A-treated samples (Figure 40A). The majority of splice events (75.5%) were detected in the treated condition. Almost 60% of the events were connected to splice junctions and among them, 30% of all events were novel, non-annotated splice junctions (Figure 40B). Many of these novel splice

events exhibited an extremely high fold change (logFC ± 20) because they were specific to one condition. Similar to the results of the *rMATS* analysis, the alternative splicing phenotype was even more pronounced in RPE-1 cells, likely due to a higher dose of GEX1A treatment. Almost twice as many (33,227) differential splice events were found in RPE-1 cells (Figure 40C). Again, 3 out of 4 splice events were found to be upregulated in GEX1A-treated cells. Compared to K562 cells, a higher proportion of novel splice events (42.5%) were identified in RPE-1 cells in response to GEX1A treatment, suggesting that a higher splicing inhibitor concentration generated more aberrant non-annotated transcripts. The *JunctionSeq* analysis provided many differential splice events, including several low abundant novel splice junctions.

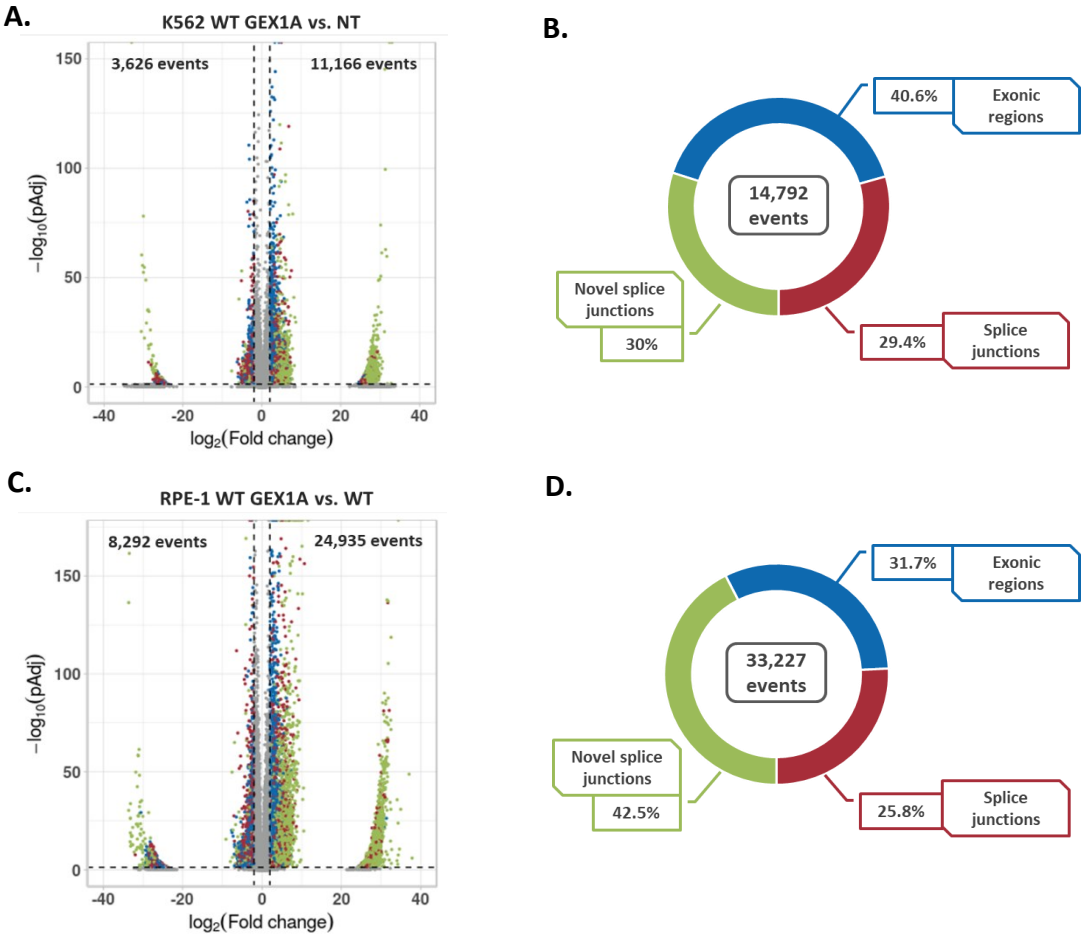


Figure 40 | JunctionSeq analysis identified thousands of differential splice events induced by splicing inhibition. ***(A)*** *Volcano plot of differentially regulated splice events in GEX1A-treated K562 cells.* ***(B)*** *Number of splice events of GEX1A-treated K562 cells classified by feature type.* ***(C)*** *Volcano plot of differentially regulated splice events in GEX1A-treated RPE-1 cells.* ***(D)*** *Number of splice events of GEX1A-treated RPE-1 cells classified by feature type. Cut-offs: pAdj ≤ 0.05 & logFC ≥ 2.*

3.8.4 *DJExpress* differential splicing analysis

A major point to consider for the differential splicing analysis was the altered gene expression between the conditions due to GEX1A treatment, as it has been shown above (Section 3.7.2). However, *JunctionSeq* does not detect changes in the overall gene expression (Hartley and Mullikin, 2016). This can become crucial since increased splice junction expression can not only result from differential splicing but also elevated gene expression. This was an important aspect for identifying splicing-derived neoepitopes considering most of the detected neoepitopes originated from altered gene expression levels (Figure 35). To include these gene expression changes between analyzed sample groups in the differential splicing analysis, the *Limma*-based *DJExpress* pipeline was developed (Gallego-Paez & Mauer, 2022).

The *DJExpress* analysis discovered 190,107 splice junctions in RPE-1 WT samples, from which 19,072 showed significant changes in junction usage between non-treated and GEX1A-treated samples (Figure 41A). As observed before, GEX1A treatment resulted in a higher proportion of upregulated splice events: 17,067 junctions showed a significantly higher differential usage. In comparison, only 2,005 junctions were significantly more often excluded in transcripts of GEX1A-treated cells. Additionally, the *DJExpress* analysis determined whether gene expression changes explained the altered frequency of a detected splice event. The correlation between junction expression of sample groups (absolute logFC) and relative junction expression within a specific gene (relative logFC) of RPE-1 WT samples was visualized in a 9-square scatter plot (Figure 41B). Here, grey points comprise junction without differential expression or differential usage. The red and blue highlighted events represent junctions with similar values of absolute and relative logFCs. This reflects changes in splicing patterns between experimental conditions without confounding alterations in the total expression of the gene (Gallego-Paez & Mauer, 2022). The junctions colored in green indicate that the differences in junction expression or usage between sample groups could be explained – at least partially – by altered gene expression levels or vice versa.

The correlation between altered gene expression and junction usage was particularly important for K562 samples because IFN-γ stimulation strongly affected gene expression levels, which can potentially mask alternative splicing events (Figure 37). However, the majority of significant splice events could not be explained by altered gene expression but resulted from differential junction usage (Figure 41D). In total, 143,806 splice junctions were detected, from which 2,604 junctions exhibited different junction usage between non-treated and GEX1A-treated K562 samples (Figure 41C). Most of the junctions (n = 2,239) showed a significantly higher differential usage in the treated condition, while only 365 junctions were significantly more often excluded in response to GEX1A treatment. These data

illustrate that *DJExpress* analysis was able to determine differential splice events independent of the overall expression of the corresponding gene.

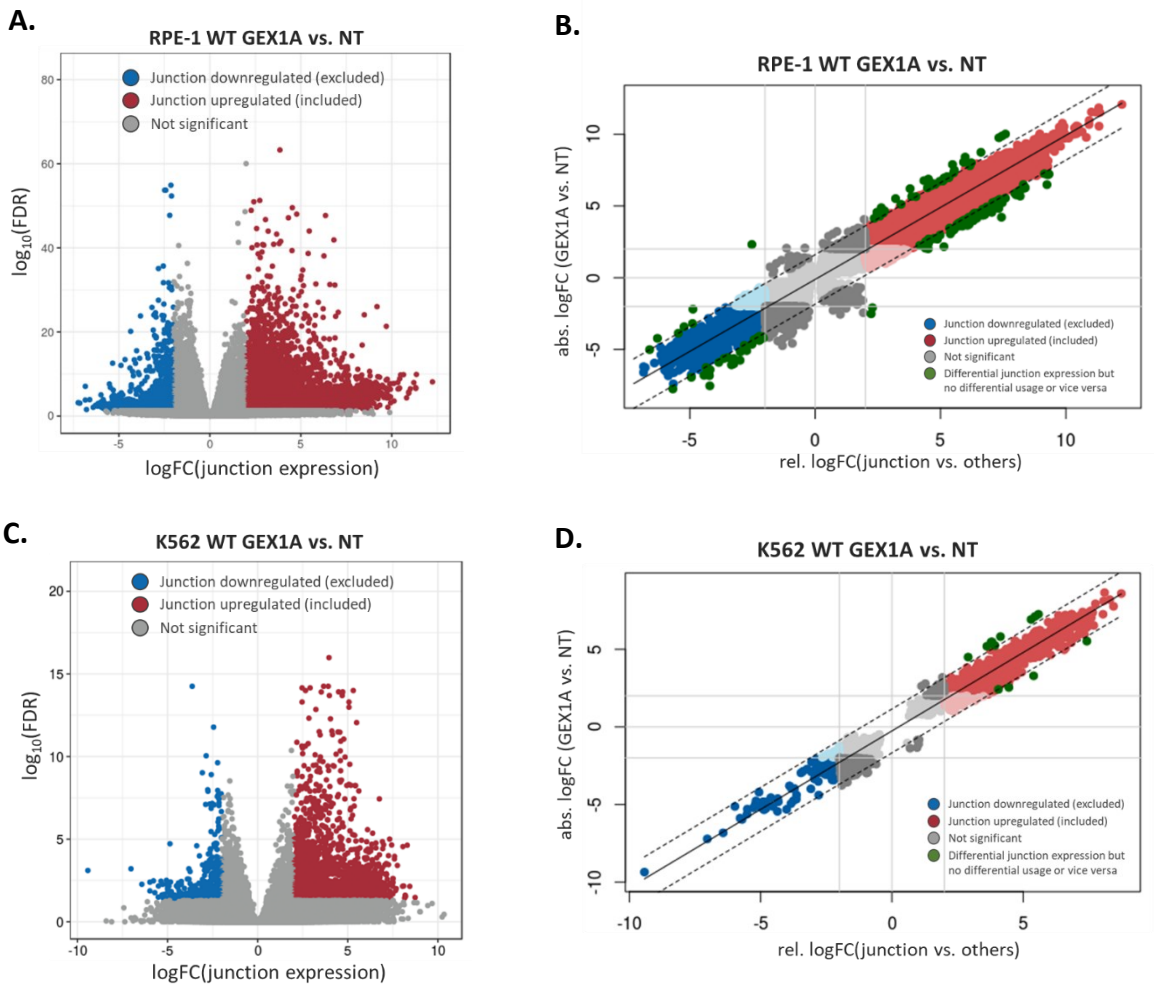


Figure 41 | DJExpress analysis reveals differential splice junction usage in response to GEX1A treatment. (A) *Volcano plot of differential junction expression between RPE-1 sample groups.* ***(B)*** *Quadrant scatter plot visualizing the categorization of junction events of RPE-1 samples.* ***(C)*** *Volcano plot of differential junction expression between K562 sample groups.* ***(D)*** *Quadrant scatter plot visualizing the categorization of junction events of K562 samples. Cut-offs: logFC ≥ 2; FDR ≤ 0.05; minimum junction read coverage = 5.*

In addition, the intensity of the changes in the splicing landscape between the sample groups was examined. As observed in the previous differential splicing analyses, a higher dose of GEX1A resulted in a more pronounced differential splicing phenotype reflected by the number of detected splice events in K562 and RPE-1 cells. K562 WT samples were treated with a GR_{25} dose of GEX1A, resulting in 2,279 differentially used junctions, while 19,072 junctions were obtained in RPE-1 WT samples treated with the GR_{75} dose of GEX1A (Table 16). Although treated with a similar dose of GEX1A, 7,403 additional differentially spliced junctions were detected in RPE-1 *TP53*$^{-/-}$ samples compared to WT

samples, indicating that *TP53*$^{-/-}$ cells were more sensitive to splicing inhibition. Overall, GEX1A treatment caused distinct changes in the splicing pattern of RPE-1 and K562 cells. In contrast, *DJExpress* analysis between RPE-1 WT and *TP53*$^{-/-}$ cells derived only 126 differentially used junctions. Similarly, only 1,007 significant splice events were detected between GEX1A-treated RPE-1 WT and *TP53*$^{-/-}$ cells. However, most splice junctions were found here due to GEX1A treatment in both conditions. Notably, the *DJExpress* analysis detected tens of thousands of novel splice junctions which were not annotated in the UniProt reference genome (Table 16). The analysis also identified junctions where read evidence was only found in the treated condition, here called neojunctions. Overall, GEX1A treatment induced the most significant changes in the splicing landscape. Thousands of differentially spliced events were detected by *DJExpress* analysis which could be the origin of previously identified neoepitopes.

Table 16 | Results statistics of DJExpress analysis between RPE-1 and K562 sample groups. *Non-annotated junctions were not found in UniProt annotation. Junctions with read evidence exclusively in the treated condition were called 'neojunctions'.*

***DJExpress* analysis**	**Identified junctions**	**Non-annotated junctions**	**Neojunctions**	**Significant junctions (down/upregulated)**
RPE-1 WT GEX1A vs. NT	190,107	41,020	10,168	2,005/17,067
RPE-1 TP53$^{-/-}$ GEX1A vs. NT	198,232	48,944	10,547	2,796/24,476
RPE-1 TP53$^{-/-}$ NT vs. RPE-1 WT NT	180,251	28,714	2,347	66/60
RPE-1 TP53$^{-/-}$ GEX1A vs. RPE-1 WT GEX1A	200,119	55,804	1,813	112/895
K562 WT GEX1A vs. NT	143,806	19,236	4,413	365/2,239

3.9 Identification of alternative splicing-derived neoepitopes

3.9.1 Generation of peptide reference libraries for aberrant splice events

To identify alternative splicing-derived neoepitopes, the previously identified candidate peptides were matched to differential splice events detected by *DJExpress*. The genomic information of the differential splice events was transformed into peptide sequences. Only junctions with significant differential expression between the conditions – independent of alterations in gene expression – were considered for the analysis (Table 17). Importantly, the approach was limited to upregulated (included) junctions. In the case of downregulated (excluded) junctions, the resulting sequence could not be determined undoubtedly. The junction coordinates of filtered *DJExpress* events were expanded by generating overhangs of 29 nucleotides resulting in junction-spanning sequences, which were then translated into all three reading frames (Figure 42). The length of 29 nucleotides was chosen since most HLA-I binders have a length between 8-10 amino acids/24-30 nucleotides (Gfeller et al., 2018b) and at least one nucleotide was required to be junction-spanning. Nevertheless, the absence of a junction would consequently increase the expression of neighboring junctions, and thus the corresponding splice event would also be represented in the final peptide sequence library.

These *Alternative Splicing Peptide Libraries* were individually generated for each sample group. Each library derived from differential junction expression events in response to GEX1A treatment contained tens of thousands of unique peptide sequences (Table 17). Except for the library generated from RPE-1 *TP53*$^{-/-}$ NT vs. WT NT, which was limited to 78 peptide sequences since only 35 splice events were obtained after filtering. In general, the size of the databases correlated with the number of differentially expressed junctions of each sample group. These peptide libraries contained all possible peptide sequences that could be generated from differential splice events and were used to match neoepitopes to differential splice events.

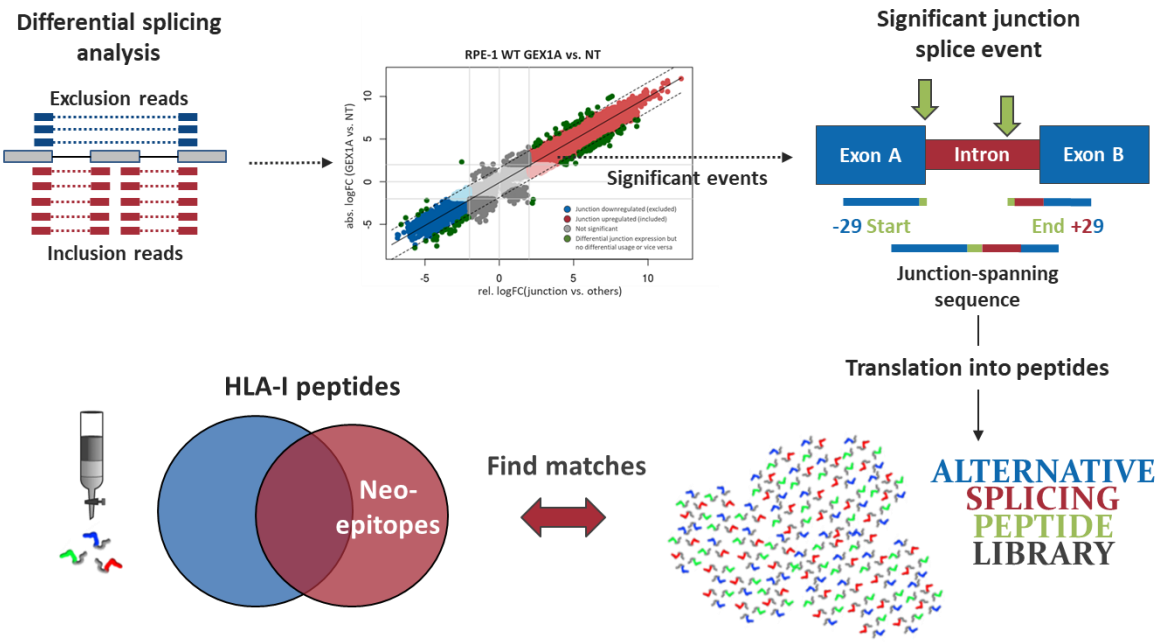


Figure 42 | Translation of significant splice events from alternative splicing analysis into peptide information. *Sample-specific peptide sequence libraries were generated from significantly upregulated splice events in the GEX1A-treated condition, which were not explained by altered gene expression (cut-offs: logFC ≥ 2; FDR ≤ 0.05). Only peptide sequences ≥ 8 amino acids were included in final libraries.*

Table 17 | Number of peptide sequences of sample-specific Alternative Splicing Peptide Libraries derived from DJExpress analysis.

DJExpress analysis	RPE-1 WT GEX1A vs. NT	RPE-1 TP53$^{-/-}$ GEX1A vs. NT	RPE-1 TP53$^{-/-}$ NT vs. RPE-1 WT NT	K562 WT GEX1A vs. NT
Filtered sig. splice events	15,034	21,324	35	992
No. peptide sequences	34,911	49,669	78	2,307

3.9.2 Identification of splicing-derived neoepitopes

Using the *Alternative Splicing Peptide Libraries*, it was possible to determine which of the previously identified neoepitope candidates originated from differentially regulated alternative splice events. The short peptide sequences of neoepitope candidates were searched in the peptide libraries derived from differential junction expression events. Following this strategy, several splicing-derived neoepitopes were identified across all sample groups (Table 18). The *MSH2*$^{-/-}$ T2 sample group provided the most neoepitope candidates and consequently delivered the highest number of splicing derived neoepitopes (n = 18) of all groups. In contrast, only two candidates were obtained from the WT sample group. Notably, the proportion of splicing-derived neoepitopes discovered from neoepitope

candidates was higher in cell lines carrying genetic deficiencies in the DNA damage repair pathway (Table 18)). Overall, neoepitope candidates were matched to 23 unique splice events. Here, one specific peptide matched two different splice events within the same gene. Also, two candidates were assigned to the same splice event. Twelve peptide candidates were identified from the canonical GENCODE annotation while the custom reference databases contributed 11 peptides to the overall 23 identified splicing-derived neoepitope candidates (Figure 43A).

Table 18 | Number of splicing-derived neoepitopes per sample group. *Neoepitope candidates found in ≥ 2/3 replicates treated with GEX1A were assigned to differential splice events (DJExpress results) by searching for matching peptide sequences in Alternative Splicing Peptide Libraries.*

Sample group	WT	TP53$^{-/-}$	MSH2$^{-/-}$ T2	MSH2$^{-/-}$ T6	POLE^{P286R} T16	POLE^{P286R} T17
No. of neoepitopes	366	728	751	382	217	402
Splicing-derived neoepitopes	2	7	18	5	4	10
Proportion	0.55%	0.96%	2.4%	1.31%	1.84%	2.49%

3.9.3 Filtering pipeline for neoepitope candidates

The 23 neoepitope candidates were then subjected to manual filtering (Figure 43B). At peptide level, this included a *BLASTp* search against the human proteome to confirm that a neoepitope candidate could only arise from one specific splice event. *BLASTp* search assigned 18 peptides to their protein accessions. The remaining five peptides did not share 100% sequence similarity with any database entry, making them ideal neoepitope candidates.

Moreover, it was crucial to determine the novelty of candidate peptide sequences. By looking at the differential splice events, it was evaluated whether the sequence indeed originated from an aberrant splice event in the GEX1A-treated condition. Here, 13 peptide sequences were found to be derived from the canonical splice event and were thus no longer considered neoepitopes. Notably, all twelve peptides identified from the UniProt reference database were among them. Consequently, the number of candidates was reduced to ten peptides which all originated from custom reference libraries (Figure 43C). The sequences of remaining candidates contained one to eight novel nucleotides, distinguishing them from the canonical isoform (Table 19).

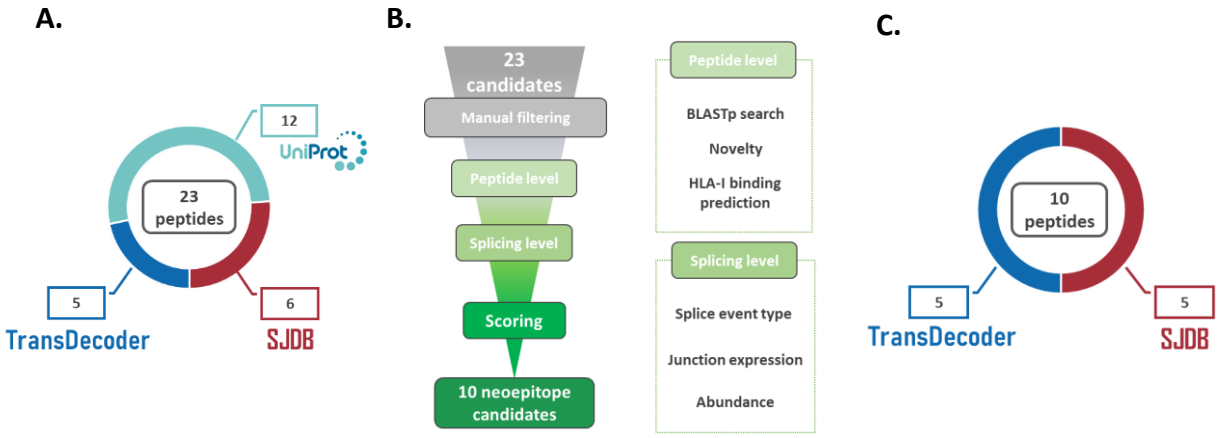


Figure 43 | Filtering approach for neoepitope candidates of RPE-1 sample groups identified from DJExpress results. (A) *Number of candidate peptides identified from different reference libraries.* ***(B)*** *Neoepitope candidates were evaluated on peptide and splicing level to select the most promising candidates* ***(C)*** *Number of selected candidate peptides identified from each custom reference library.*

Table 19 | Peptide information of filtered neoepitope candidates of RPE-1 sample groups identified from DJExpress analysis. *Amino acids varying from canonical isoform are labeled red.*

Peptide	Novelty	Reference database	BLASTp result	Assigned gene
SLAKALYEA	novel 5 nt	TransDecoder	no match	ANXA1
YLLSREENAF	novel 8 nt	SJDB	no match	ATP6V0E1
ILDNGEWTV	novel 1 nt	TransDecoder	CCNB1	CCNB1
KILDNGEWTV	novel 1 nt	TransDecoder	CCNB1	CCNB1
YIGDIHPLL	novel 2 nt	TransDecoder	no match	CYB5B
YIDIEARTRY	novel 1 nt	SJDB	DIS3L	DIS3L
KLNEYLLQY	novel 5 nt	SJDB	no match	PSMA4
ELFQRPNAL	novel 1 nt	TransDecoder	PUM1	PUM1
GMIIGPPRV	novel 1 nt	SJDB	UBE2V2/UBE2V1	UBE2V2
SYSDIVRIY	novel 8 nt	SJDB	no match	UBE4B

HLA-I binding prediction revealed binding affinities of candidate peptides to specific HLA-I subtypes of RPE-1 cells (Table 20). All candidates matched the binding pattern of at least one HLA-I allele. Five out of ten candidates were predicted binders of HLA-A*02:01, followed by HLA-A*01:01 and HLA-C*07:02. Three other peptides were also predicted binders of HLA-C alleles but exhibited a better affinity towards HLA-A or HLA-B alleles. Notably, there was no predicted binder of HLA-B*07:02 among the candidate peptides, although HLA-B*07:02 binders made up the second biggest peptide cluster identified by GibbsClustering (Figure 25).

Table 20 | HLA-I binding prediction results of filtered neoepitope candidates of RPE-1 sample groups identified from DJExpress analysis. *Numbers represent K_D values in nM. Coloring: dark green = best predicted binding allele; light green = predicted binding allele*

Peptide	HLA-A*01:01	HLA-A*02:01	HLA-B*07:02	HLA-B*08:01	HLA-C*07:01	HLA-C*07:02
SLAKALYEA		8.5				
YLLSREENAF				596.1		
ILDNGEWTV	4904.8	5				
KILDNGEWTV		19				
YIGDIHPLL		19.5			2130	738.1
YIDIEARTRY	87.4					
KLNEYLLQY	1446.5					3077.9
ELFQRPNAL				40.6	2377.4	1505.4
GMIIGPPRV		41				
SYSDIVRIY					605.3	166.1

Moreover, the associated splice events of neoepitope candidates were examined. In line with the previous results from the *rMATS* analysis (Figure 39), eight of ten splice events were exon skipping events (Table 21). In addition, one peptide derived from a mutually exclusive exon (MXE) event and another from an alternative 3' splice site were identified. Six splice events were annotated in GENCODE, while four novel, non-annotated splice events were discovered. All splice events showed a strong differential expression between treatment conditions independent of altered expression of the related gene (Table 21), passing the previous selection criteria. Very low median expression was detected for nine out of ten splice events in the non-treated control group. Three splice events were even labeled as 'neojunction' since there was no read evidence for the splice event detected in the non-treated condition. In contrast, the splice event of *ATP6V0E1* was also robustly detected in non-

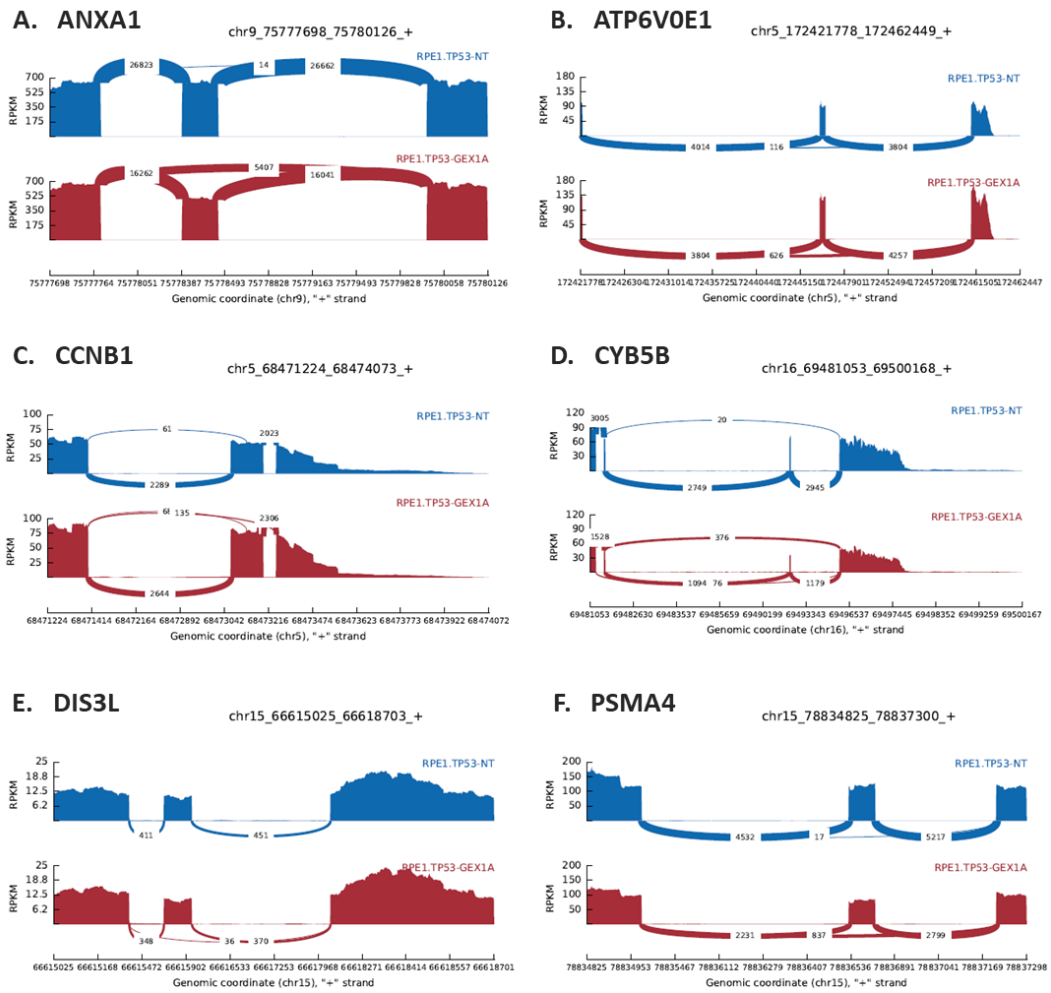
treated samples. Thus, this candidate needed to be considered with care since the peptide could potentially also be produced in the control condition. In general, there was a difference observed in the abundance of differentially expressed splice events. The exon skipping event in *ANXA1* was extremely evident which was reflected by the high median expression in the treated condition (Table 21). In contrast, neojunction events such as exon skipping in *DIS3L* were lowly abundant.

Table 21 | Splice event information of filtered neoepitope candidates of RPE-1 sample groups. *The values for sample median junction expression and logFC were obtained from DJExpress analysis. Amino acids varying from canonical isoform are labeled red. MXE = mutually exclusive exon; SS = splice site*

Peptide	Gene	Splice event type	Annotation	median expr. NT	median expr. GEX1A	logFC *DJExpress*
SLAKALYEA	ANXA1	exon 8 skipping	non-annotated	14	5956	10.14
YLLSREENAF	ATP6V0E1	MXE exon 3 (last exon)	GENCODE	109	731	2.58
ILDNGEWTV	CCNB1	exon 8 skipping	GENCODE	4	124	5.24
KILDNGEWTV	CCNB1	exon 8 skipping	GENCODE	4	124	5.24
YIGDIHPLL	CYB5B	exon 3+4 skipping	GENCODE	0	79	8.87
YIDIEARTRY	DIS3L	exon 10 skipping	non-annotated	0	31	5.20
KLNEYLLQY	PSMA4	exon 5 skipping	GENCODE	13	499	6.10
ELFQRPNAL	PUM1	exon 3-7 skipping	GENCODE	14	114	2.95
GMIIGPPRV	UBE2V2	exon 3 skipping	non-annotated	9	86	3.07
SYSDIVRIY	UBE4B	alt. 3' SS exon 4	non-annotated	0	29	6.42

The differential splice events and their abundance was further evaluated by Sashimi plots of raw read counts to visualize junction read coverage (Figure 44). This visualization highlighted the strong read evidence for the novel *ANXA1* exon skipping event (Figure 44A). Also read coverage was found for the

occurrence of the other splice events. For the exon skipping events in *CCNB1* and *DIS3L,* read evidence was exclusively found in GEX1A-treated samples (Figure 44C+E). The splice events in *ATP6V0E1, CYB5B, PSMA4, PUM1,* and *UBE2V2* were also detected in the non-treated condition but occurred to a much lower extend (Figure 44B,D,F,G,H). Except for *ATP6V0E1,* where the Sashimi plot illustrated the relatively high read coverage of the MXE event in non-treated samples (Figure 44B). Notably, the novel alternative 3' splice site in the *UBE4B* gene was not detected although read coverage for the specific region were detected as indicated by an elevated read count in the treated samples (Figure 44I). Overall, Sashimi plots confirmed the occurrence of all detected differential splice events and highlighted their different abundance in GEX1A-treated and non-treated samples.



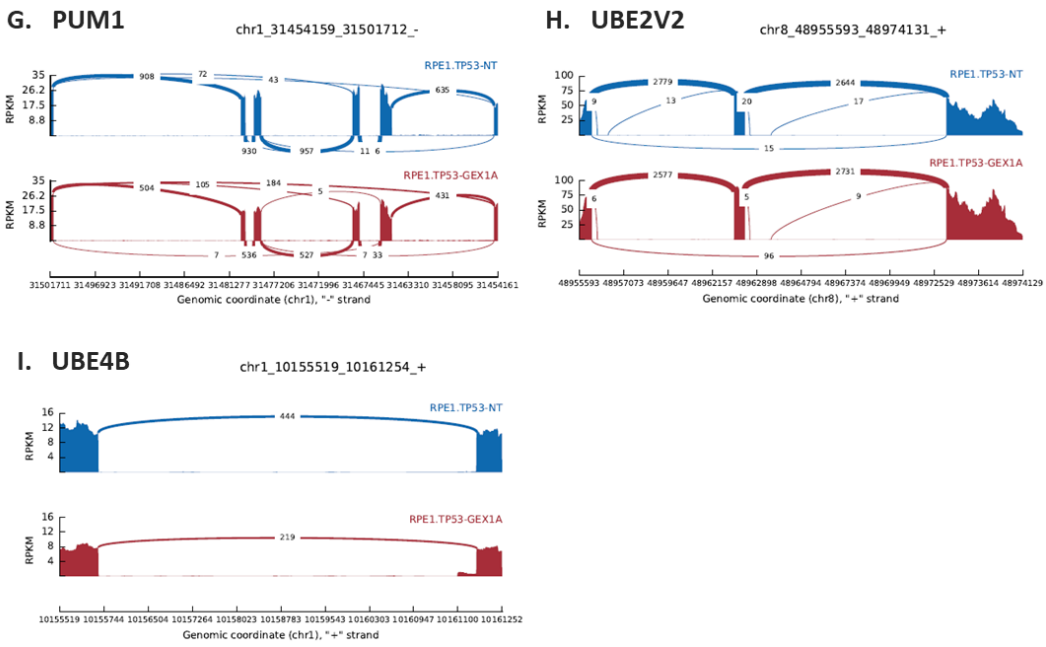


Figure 44 | Sashimi plots of differential splice events with matching neoepitopes. (A)-(G) *Read coverage of detected splice event in the non-treated (BLUE) and GEX1A-treated (RED) condition. Numbers represent read counts. Read files of three replicates per condition were grouped and averaged. RPKM: Reads per kilobase of transcript per Million mapped reads.*

3.10 *In vitro* validation of neoepitope candidate splice events

3.10.1 Validation of differential splice events by RT-PCR

The computational detected splice events of neoepitope candidates were reproduced in an independent experiment under similar conditions and experimentally validated by RT-PCR. The regions surrounding splice events were amplified by RT-PCR to detect different isoforms resulting from differential splicing (Figure 45A). Following this strategy, two additional isoforms were detected, derived from aberrant splicing of the ANXA1 exon 8 region upon GEX1A treatment of RPE-1 cells (Figure 45C). In the non-treated samples, only the canonical sequence was detected, whereas two additional isoforms appeared in GEX1A-treated samples in a dose-dependent manner. The lower band with a size of 159 bp belonged to the skipped exon 8 isoform. Besides confirming the presence of the novel exon 8 skipping event, the top band (~300 bp) gave evidence for a predicted intron retention event in the surrounding of ANXA1 exon 8. However, this intron retention event was not significantly differentially expressed, while three exon skipping events were detected across the whole ANXA1 gene (Figure 45B). These three events resulted in several transcript isoforms, which were successfully detected by RT-PCR (Figure 45D).

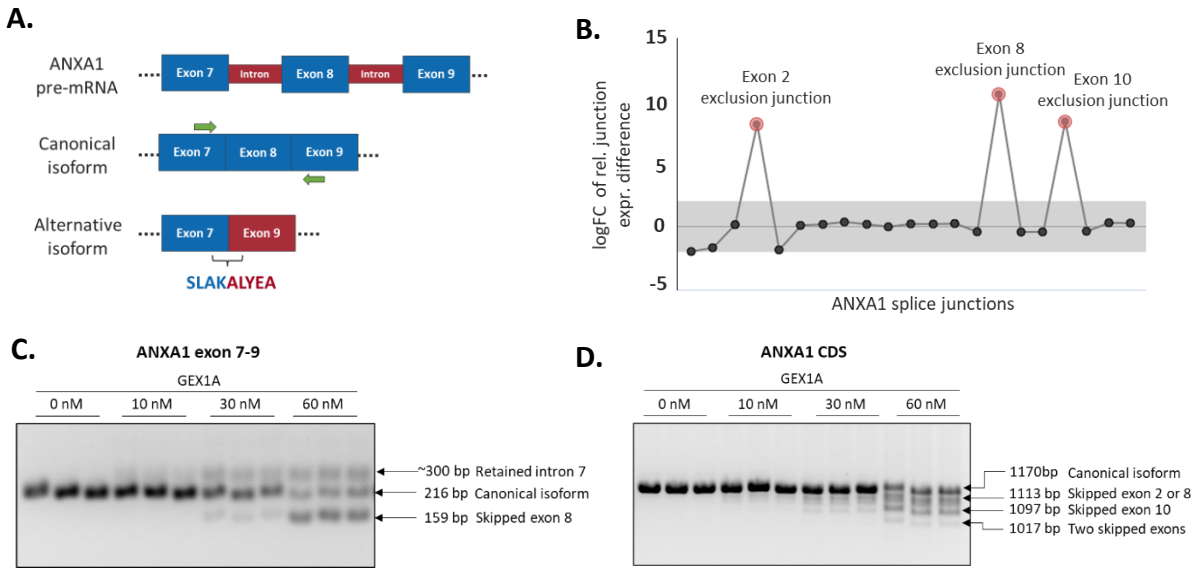


Figure 45 | Validation of a novel differential splicing event in the ANXA1 gene by RT-PCR. (A) *Illustration of the non-annotated differential splice event of exon 8 skipping in ANXA1.* ***(B)*** *Splice plot of ANXA1 illustrating differential junction expression across the gene between GEX1A-treated and non-treated RPE-1 cells. Red dots indicate significant different junction usage independent of gene expression.* ***(C)*** *RT-PCR products resulting from amplification of the region surrounding ANXA1 exon 8 in response to GEX1A treatment.* ***(D)*** *RT-PCR products resulting from amplification of the coding sequence of ANXA1 in response to GEX1A treatment. RT-PCR was performed on amplified cDNA from RPE-1 cells treated with 0, 10, 30, or 60 nM GEX1A for 24 h. PCR products were separated by electrophoresis to identify different splice isoforms based on their length.*

Following the same strategy, I validated splice events associated with other neoepitope candidates. For example, differential exon skipping events of *CCNB1*, *PSMA4*, and *UBE2V2* were confirmed by RT-PCR (Figure 46). In addition to the canonical transcript isoform, evidence for the skipped exon isoform was detected for these three genes in response to GEX1A treatment. All three events were previously predicted by *DJExpress* analysis to be significantly differentially expressed, as highlighted in the splice plots of Figure 46. Besides, an annotated intron retention event in *TUBA1A* was investigated. This event was predicted to be significantly differentially expressed by *JunctionSeq* analysis but not by *DJExpress* analysis. RT-PCR revealed no evidence of this intron retention event in *TUBA1A* in response to GEX1A treatment in RPE-1 cells. In general, read evidence of this splicing event was low and likely elevated due to an enhanced *TUBA1A* gene expression in GEX1A-treated samples. These data support the notion that *DJExpress* analysis can efficiently prevent the detection of false-positive differential splice events resulting from gene expression changes between sample groups.

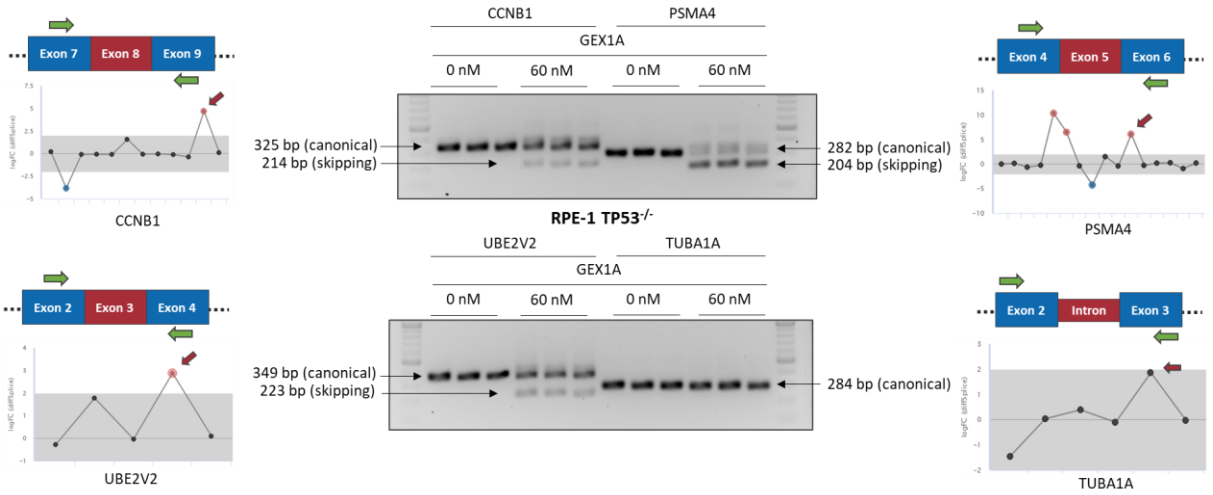


Figure 46 | Validation of candidate splice events by RT-PCR. *Differential splice events were highlighted in splice plots. Regions surrounding differential splice events of CCNB1, PSMA4, UBE2V2, and TUBA1A were amplified from cDNA of RPE-1 TP53-/- samples treated with 0 nM or 60 nM GEX1A for 24 h. PCR products were separated by electrophoresis to identify different splice isoforms based on their length. Red arrows mark junctions of interest.*

3.10.2 Identification of differential splice events in TCGA patient data

Since alternative splice events were discovered in cell models, the presence of novel splice events was explored in TCGA patient data to investigate their clinical potential. Indeed, novel splice events such as exon 8 skipping of *ANXA1* were detected in patient data and found to be differentially expressed across several cohorts (Figure 47A). Here, isoform expression separated patients into two groups, with the smaller subgroup of patients exhibiting high expression of the skipped exon 8 variant. Several head and neck cancer (HNSC) patients showed increased expression levels of this novel isoform, indicated by high expression variance inside the cohort. To examine the impact of the expression of the alternative *ANXA1* exon 8 isoform, a survival analysis was performed on patients expressing the HLA*02:01 allele (Figure 47B). Patients possessing high expression levels of the *ANXA1* exon 8 variant exhibited a significantly better prognosis ($p = 0.0068$) which was reflected in median survival of 2200 days. In contrast, patients with low expression levels had a poorer outcome reflected in median survival of only 1100 days.

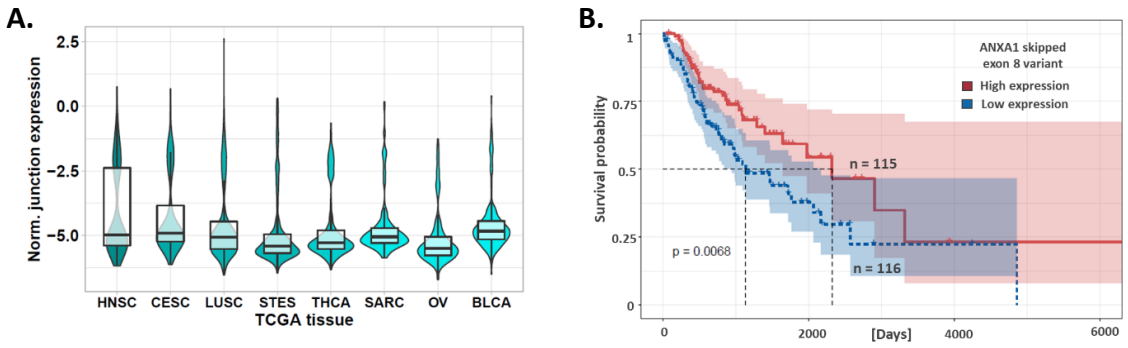
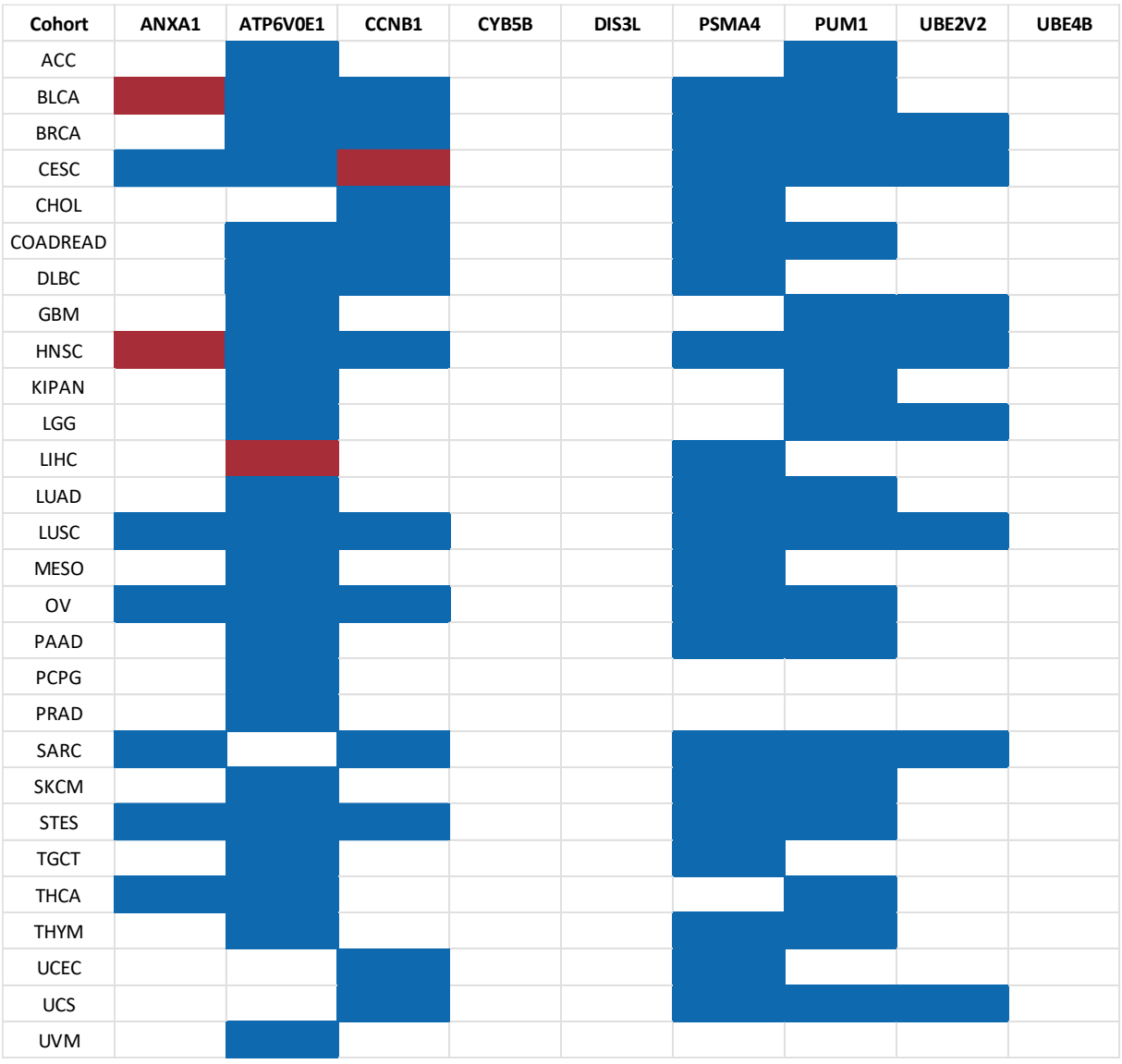


Figure 47 | Detection of the non-annotated ANXA1 exon 8 skipping event in TCGA patient data. (A) *Variance in the expression of novel ANXA1 exon 8 skipped isoform across TCGA cohorts. Normalized (norm.) junction expression describes the expression of the junction of interest in relation to the average expression levels of all junctions within the gene.* ***(B)*** *Survival of HLA-A*02:01-expressing HNSC patients (n = 231) based on skipped ANXA1 exon 8 isoform expression. Groups for survival analysis were separated by median isoform expression. Survival probability describes the ratio between the number of participants surviving and the number of participants at risk.*

Junction expression of differential splice events of all candidates was checked in TCGA patients expressing the required HLA-I allele for peptide presentation (Table 22). This revealed that six candidate splice events were differentially expressed between patients of multiple cohorts and normal tissue. In four cases, this differential junction expression had a significant impact on the survival of the patients, similar to the ANXA1 event displayed in Figure 47B. In contrast, splice events of *CYB5B*, *DIS3L*, and *UBEE4B* were not found differentially expressed between patients and normal tissue. Thus, the respective neoepitope candidates were not considered for subsequent immunogenicity assays. The detection of candidate splice events in TCGA data demonstrated that those events also occur in patients, highlighting their clinical potential.

Table 22 | Detection of differentially expressed splice events in TCGA data. *Patient data comprises junctions differentially expressed in cancer patients compared to a normal tissue control (CCLE). For each event, only patients were selected that expressed the matching HLA-I subtype for the respective candidate peptide. BLUE = junction event differentially expressed; RED = differential junction event expression detected with significant impact on patient survival (p < 0.05); WHITE = junction not found / not differentially expressed between patients and normal tissue.*



3.11 Immunogenicity assays of candidate neoepitopes

The immunopeptidomics workflow combined with the differential splicing analysis provided ten neoepitope candidates. As final step of the pipeline, I performed immunogenicity assays in collaboration with Alejandro Hernandez-Sanchez to evaluate the immunogenic potential of predicted differential splicing-derived candidates. The T cell response towards neoepitopes was determined by *in vivo* immunization assays using the HLA-A*02:01-transgenic mouse model (Pajot et al., 2004). These A2.DR1 mice are deficient for mouse MHC molecules and instead carry human HLA-A*02:01 and HLA-DR1 alleles. This enables A2.DR1 mice to present peptides identified from human cell models on the cell surface. Thus, this model system is suitable to evaluate the immunogenicity of human neoepitope candidates.

However, the approach is limited to evaluating peptides binding to the HLA-A*02:01 subtype. According to HLA-I binding analysis, five candidates were predicted binders of this subtype and were thus suitable for the *in vivo* immunization approach (Table 20). However, A2.DR1 mice were not immunized with the *YIGDIHPLL* peptide from *CYB5B* since the matching splice event was not detected in patient data (Table 22). Instead, the *KLNEYLLQY* peptide from *PMSA4* was included in the vaccination rounds (Table 23). Although this was not a predicted binder of HLA-A*02:01, this peptide originated from an interesting differential splice event and was a predicted binder of the sister allele HLA-A*01:01 with a similar binding motif. In addition, mice were vaccinated with two control peptides. As negative control, the *GMLPDPKNV* peptide from *ANT1* was included. This peptide was not only found in both treatment conditions of RPE-1 cells but also shares the same sequence in mouse and human. As positive control, the *YMLDLQPET* peptide from *HPV16-E7* induced an immune response in mice. For immunogenicity assays, we immunized the mice with two peptides per group to evaluate the T cell response in mice (Table 23).

Table 23 | Peptides used for immunization experiments of A2.DR1 mice.

Peptide	HLA-A*02:01 K_D [nM]	Function	Origin	Group (n = 6 mice)
SLAKALYEA	8.5	candidate	ANXA1	2
KLNEYLLQY	4432	candidate	PSMA4	2
ILDNGEWTV	4.98	candidate	CCNB1	3
GMIIGPPRV	41	candidate	UBE2V2	3
KILDNGEWTV	19	candidate	CCNB1	4
GMLPDPKNV	165	neg. control	ANT1	4
YMLDLQPET	22.3	pos. control	HPV16-E7	1

3.11.1 Validation of candidate peptide sequences

Before we immunized the animals with candidate peptides, I validated all peptide sequences. Due to the limited sensitivity and specificity of the LC-MS/MS approach, the detection of short peptides with similar sequence patterns can still be error-prone. To verify the peptide sequences of neoepitope candidates, I compared their experimental mass spectra to respective synthetic references. Synthetic peptide standards were measured by mass spectrometry using the exact same parameters and conditions as applied for the experimental data. Eventually, I determined the correlation of matched fragment ions between experimental and synthetic spectra (Figure 48). All candidates showed a tight correlation between the individual ion fragments, which was reflected by high Pearson's correlation values ($R^2 > 0.95$). In addition, differences in retention times of experimental data and synthetic peptides were limited to just a few minutes. This indicated that predicted peptides and synthetic standards shared physiochemical properties. Following this validation strategy, experimental peptide sequences of all six peptides used for immunization experiments were confirmed.

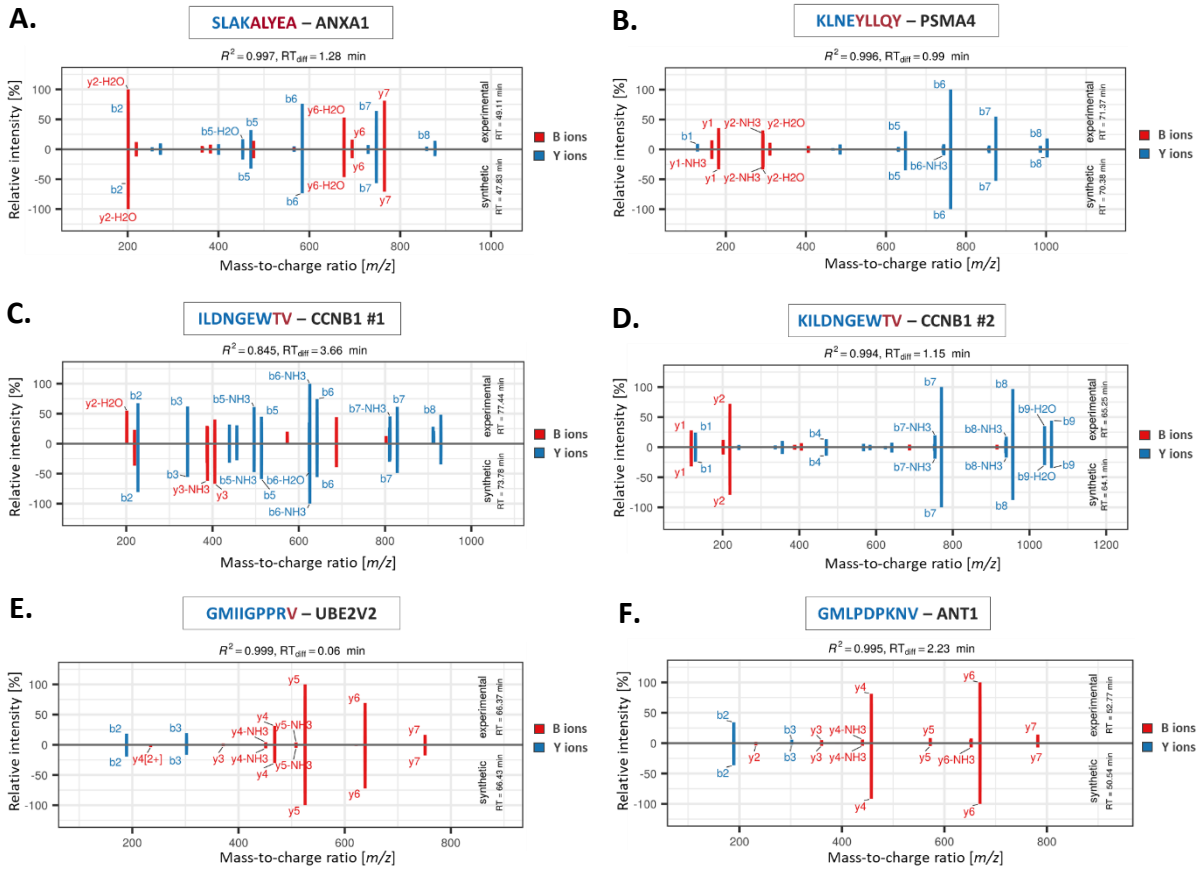
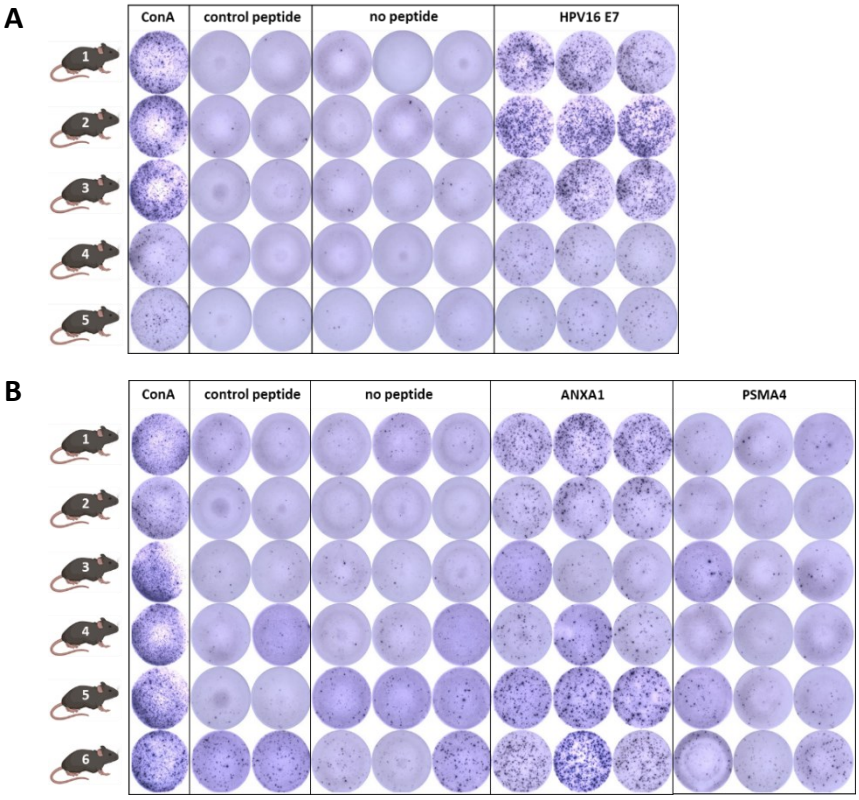


Figure 48 | Validation of peptide sequences of identified neoepitope candidates and control peptide used for immunogenicity assays. (A-F) *Synthetic standards based on experimentally predicted peptides were identified from mass spectrometry data. Ion fragments detected from experimental spectra (top) were compared to ion fragments of synthetic peptide spectra (bottom).*

3.11.2 Evaluation of T cell response by ELISpot

To determine the potential of candidate peptides to induce T cell activity, we performed ELISpot assays to capture IFN-γ production by activated T cells. We challenged isolated splenocytes from immunized mice with the T cell activator Concanavalin A (ConA) to determine the general capacity of T cells. ELISpot analysis revealed that this successfully induced T cell activation while no T cell response was observed for samples incubated with the negative control peptide or without peptide (Figure 49A). However, the magnitude of response varied between animals in Group 1. Mice #1-3 showed a more pronounced response than mice #4-5. Consistently, mice #4-5 also showed a weaker response towards the positive *HPV16-E7* control peptide than mice #1-3. Mice of Group 2 vaccinated with *ANXA1* and *PSM4A* peptides delivered a robust response to T cell activation by ConA (Figure 49B). In addition, a strong T cell response was observed when splenocytes were stimulated with the *ANXA1* peptide. In contrast, the immune response was feeble in samples incubated with the *PSMA4* peptide. All animals of Group 3 showed very pronounced T cell activation in response to stimulation with peptides of both *CCNB1 # 1* and *UBE2V2* (Figure 49C). Likewise, the second peptide of *CCNB1* was able to induce a strong T cell response (Figure 49D). In contrast, T cells were not activated by the stimulation with the *ANT1* peptide, which was derived from a canonical sequence found in both mouse and human. The ELISpot assays demonstrated that four of the five candidate peptides were able to induce a T cell response in mice.



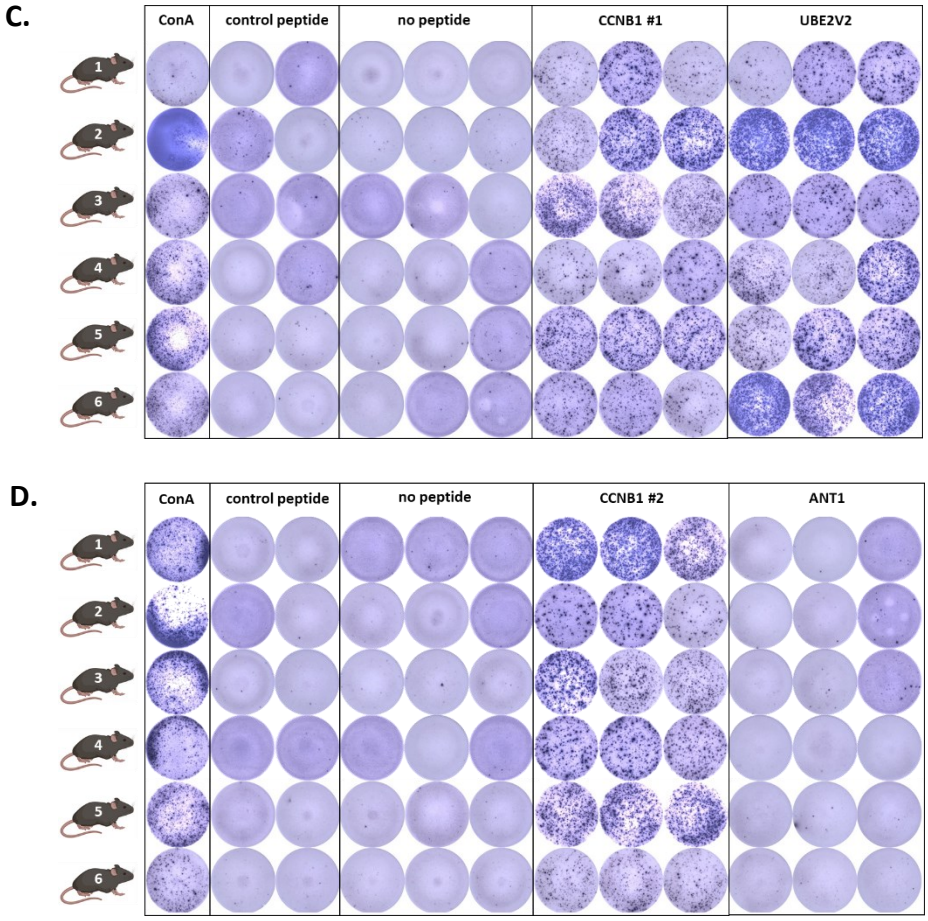


Figure 49 | T cell response to immunization of A2.DR1 mice with candidate neoepitopes. *Ex vivo IFN-γ ELISpot of isolated splenocytes (2x10^6/well) challenged with* ***(A)*** *HBV16-E7 positive control peptide;* ***(B)*** *neoepitope candidates ANXA1 + PSMA4;* ***(C)*** *neoepitope candidates CCNB1 #1 + UBE2V2;* ***(D)*** *neoepitope candidates CCNB1 #2 + ANT1. Stimulation with T cell activator Concanavalin A (ConA) was used as positive control. The SIINFEKL peptide, as known binder to H2-kb in mice, served as control peptide. A no peptide control was used as background reference.*

We quantified the results of the ELISpot assays to compare the T cell activation in response to the different candidate peptides (Figure 50). The *ANT1* peptide was derived from a canonical sequence and was thus not able to induce T cell response. In contrast, the positive control peptide derived from *HPV16-E7* induced a strong immune response reflected by approximately 200 IFN-γ spots per 1x10^6 cells. The *ANXA1* peptide induced a T cell response with a similar magnitude. In contrast, the PSMA4 peptide was not successful in elucidating a strong immune response. Both neoepitope candidates from *CCNB1* as well as the *UBE2V2*-derived candidate showed a very pronounced T cell response with up to twice as many spots as observed in the positive control. These data provide evidence that splicing-derived neoepitopes indeed possess a high immunogenic potential.

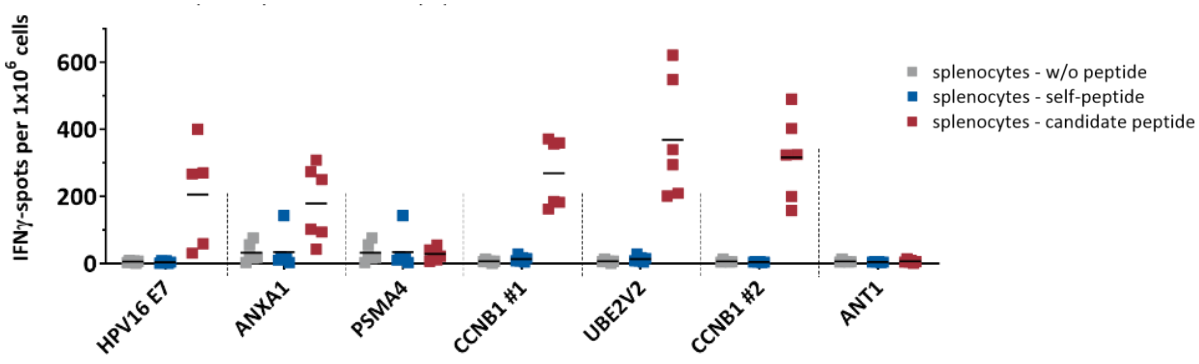


Figure 50 | Quantification of ex vivo IFN-γ ELISpot results. *Spots on membranes were counted for each replicate and normalized to 1x10^6 cells.*

3.11.3 Functional T cell killing assays

We further investigated whether the neoepitope candidates had not only the potential to activate T cell response but were also capable of inducing specific T cell-mediated killing of the neoepitope-presenting cells. These T cell killing assays were performed by Alejandro Hernandez-Sanchez, who isolated splenocytes and cultured them for a week in the presence of a candidate peptide. Afterwards, CD8$^+$ T cells primed for a specific peptide were isolated from splenocytes and co-cultured with target cells derived from the A2.DR1 cell line. Half of the target cells were loaded with a candidate peptide and labeled with CFSE. The other half was labeled with FarRed and loaded with Survivin, which T cells recognized as self-peptide. Target and effector cells were co-cultured for 48 h in different cell ratios. The following flow cytometry analysis revealed that the CD8$^+$ T cells primed against the *HBV16-E7* peptide specifically eliminated target cells presenting the very same peptide (Figure 51). The higher the concentration of CD8$^+$ T cells in the sample, the more peptide-presenting target cells were eliminated.

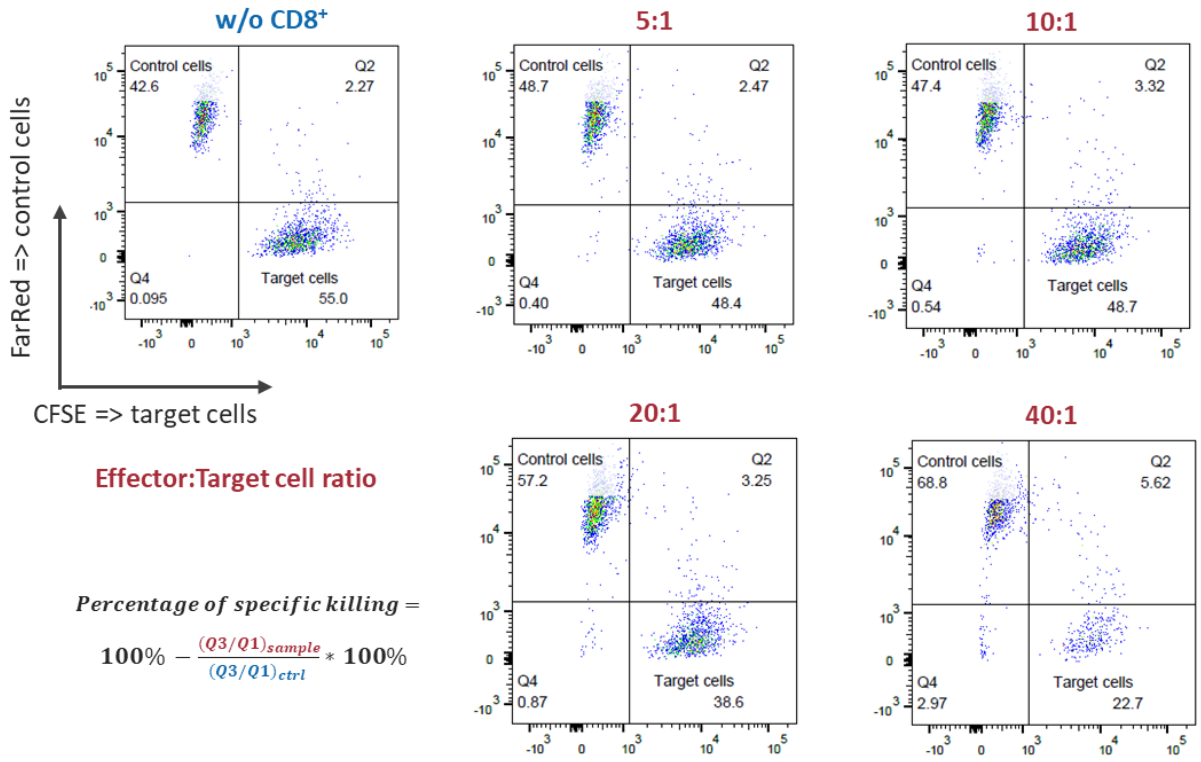


Figure 51 | Flow cytometry analysis revealed specific elimination of peptide-loaded target cells by primed CD8⁺ T cells. *A2.DR1 target cells were labeled with CFSE and loaded with peptides of interest or labeled with FarRed and loaded with Survivin (control). After 48 h of co-culturing, the ratio between the two populations was calculated for each sample to determine the percentage of target cells eliminated by primed CD8⁺ T cells.*

We determined the percentage of T cell-specific killing by calculating the ratio of target and control cells in relation to the same ratio measured in the control sample in the absence of CD8⁺ T cells. The results were plotted as T cell killing curves (Figure 52). In the positive control, up to 70% of target cells loaded with the *HBV16-E7* peptides were eliminated by CD8⁺ T cells (Figure 52A). Also, target cells presenting the ANXA1 peptide were eliminated, but only up to 40% in a ratio of 40:1 of effector to target cells (Figure 52B). The three other candidate peptides (UBE2V2, CCNB1 #1 & #2) achieved an even higher elimination rate, with more than 80% of eliminated target cells in the highest T cell concentration (Figure 52C-E). The highest percentage of killing was observed for target cells presenting the UBE2V2 peptide. Here, CD8⁺ T cells were able to eliminate close to 100% of target cells in the 40:1 condition presenting this candidate peptide (Figure 52D).

Overall, the T cell killing assays demonstrated the ability of splicing-derived neoepitopes to elicit an immune response and CD8⁺ T cell-mediated killing *in vivo*, thereby highlighting their potential to serve as functional targets for T cell-mediated elimination of aberrant tumor cells for cancer immunotherapy.

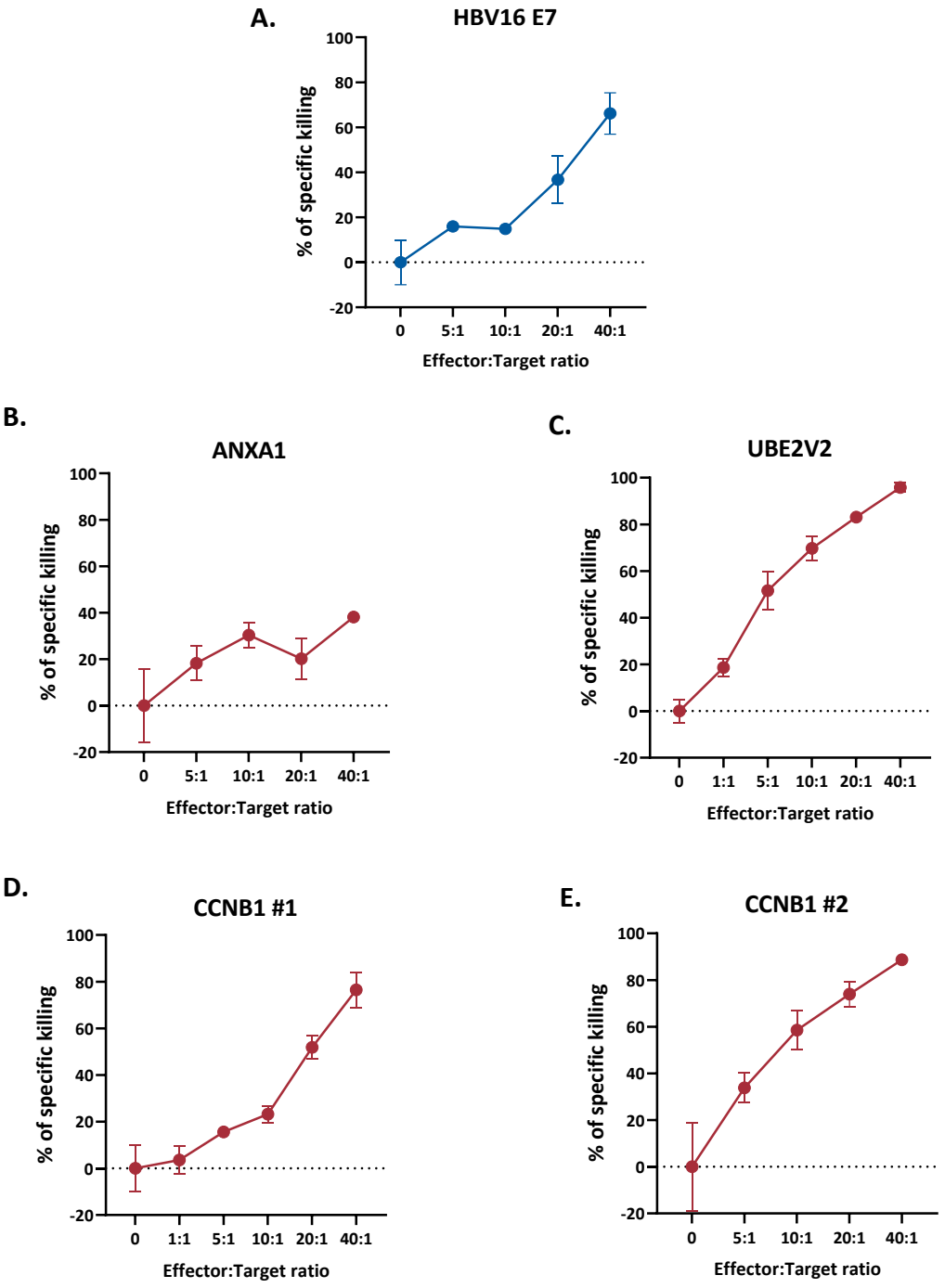


Figure 52 | Percentage of specific killing of peptide-loaded target cells by primed CD8⁺ T cells. *Percent of cells eliminated in different ratios of CD8⁺ T cells and target cells loaded with peptides derived from* ***(A)*** *HBV16-E7,* ***(B)*** *ANXA1,* ***(C)*** *CCNB1 #1 and UBE2V2,* ***(D)*** *CCNB1 #2. Data points represent the means with standard deviation of technical duplicates/triplicates.*

4. Discussion

With the work in my thesis, I aimed to investigate whether an increase in splicing burden generates aberrant splicing-derived peptides that can function as neoepitopes for cancer immunotherapy. To accomplish this, I established an immunopeptidomics pipeline which allowed me to explore the immunopeptidome of cell lines with different splicing burdens. This also included the generation of sample-specific reference libraries based on matching RNA-seq data. These custom databases enabled the identification of non-canonical, not-annotated peptides from raw mass spectrometry data.
In addition, I analyzed the transcriptome and identified all relevant significantly differentially spliced events that can give rise to potential neoepitopes. I then generated sample-specific peptide libraries by translating the genetic information of identified splice events into peptide sequences. Eventually, I combined these two efforts and matched potential neoepitopes obtained from the immunopeptidomics workflow with the peptide reference library to specifically identify neoepitope candidates originating from differential splice events.

Ultimately, I explored whether I could obtain neoepitopes with highly immunogenic properties from aberrant splice events. In addition, I wanted to evaluate how many potential new targets are generated by an increase in the splicing burden and whether this could be a predictive biomarker for response to immune checkpoint inhibitor therapy.

4.1 The immunopeptidomics approach captures splicing-derived peptides

Current strategies to identify neoepitopes derived from differential splicing changes rely on computational prediction methods. However, these computational strategies are limited in their ability to discover splicing-derived neoepitopes. These rely on error-prone HLA binding prediction tools and use mass spectrometry datasets in which splice junction peptides are underrepresented (reviewed in chapter 4.1.4). Thus, I used mass spectrometry-driven immunopeptidomics, the only analytical methodology that allows the direct identification of HLA-I peptides naturally presented *in vivo* (Bassani-Sternberg and Coukos, 2016; Chong et al., 2020). Eventually, I combined high throughput HLA-I IP, high-resolution mass spectrometry, and custom computational pipelines to explore the immunopeptidome of cell lines with alternating splicing burdens.

4.1.1 RPE-1 cells as a model to explore splicing-induced changes in the immunopeptidome

For proof-of-concept experiments, I recovered the immunopeptidome of K562 cell lines. I chose this cell model of chronic myeloid leukemia (CML) since spliceosomal mutations are commonly found in hematopoietic malignancies (Dvinge et al., 2016). Two of the most frequently reported gain of functions mutations are found in spliceosomal core components *SF3B1* and *U2AF1,* for which we had K562 cell models available at our lab. In addition, K562 are suspension cells and thus easy to expand to large numbers, which were needed as input for HLA-I IP experiments.

Despite these advantages, the K562 cell model suffered from one major shortcoming. Since K562 cells do not express HLA-I at baseline, HLA-I expression must be induced by IFN-γ stimulation. I demonstrated that HLA-I expression can be induced and maintained in a concentration and time-dependent manner at protein and cell surface level. Accordingly, I successfully recovered several thousands of peptides using HLA-I IP.

However, when I explored the alternative splicing landscape of K562 cells by performing an isoform expression analysis, I observed that K562 WT and *SF3B1*^{K700E} cells showed no difference in isoform expression patterns. I investigated this further by analyzing published RNA-seq data of unstimulated K562 cells (Seiler et al., 2018b). Here, I was able to detect isoform expression changes, which led me to conclude that IFN-γ treatment masked the largely subtle mutation-related splicing changes in my K562 dataset. Indeed, I discovered that the changes in gene expression level between the two K562 sample groups primarily resulted from IFN-γ stimulation, indicated by activated interferon response gene transcription. Moreover, it has been reported that IFN-γ stimulation induces splicing of interferon-responsive transcripts (Condino-Neto and Newburger, 2000) and alternative splicing is connected to IFN-γ response during viral infections (Liao and Garcia-Blanco, 2021).

To explore an alternative cell model where splicing is not affected by IFN-γ stimulation, I transitioned to RPE-1 cells. This also had the advantage that RPE-1 cells express HLA-I subtypes that frequently occur in the human population. For example, the HLA-A*02:01 allele occurs in 26.7% of registered German bone marrow donors and is used in common *in vivo* immunization models such as the transgenic mouse model we used in this study to measure immunogenicity of the identified neoepitope candidates. Another advantage was that I could compare an actual WT cell line with a ‘cancer-like’ cell line carrying a knockout in *TP53*. Additionally, I could explore the immunopeptidome of RPE-1 cell lines with genetic deficiencies in the DNA damage repair pathway. The only disadvantage

of RPE-1 cells was the challenging aspect that these cells grew adherent and needed to be detached. As described above, this leads to a potential loss of splicing-derived peptides. I solved this issue by using an enzyme-free detachment buffer to obtain sufficient input material for HLA-I IP while avoiding potential damage to cell surface proteins. These combined advantages of the RPE-1 cell lines made them a superior cell model for discovering splicing-derived neoepitopes.

4.1.2 HLA-I IP specifically recovers HLA-I bound peptides

The protocol for HLA-I high-throughput immunopurification was based on previous studies (Bassani-Sternberg, 2018; Chong et al., 2018) and adapted by a similar approach aiming at identifying frame-shift neoepitopes (Becker et al., 2021). Following this strategy, I successfully recovered thousands of HLA-I-bound peptides from cell lines with splicing perturbations. By comparing the K562 and RPE-1 datasets, I demonstrated haplotype-specificity of recovered peptides since only 1% of identified peptides overlapped between these two cell lines with completely different HLA-I haplotypes. In general, quality control procedures confirmed the accuracy of peptides identified from HLA-I IP samples. Thus, HLA-I IP has proven to be a suitable approach to explore the immunopeptidome of splicing-impaired cell lines.

The bottleneck of the immunopeptidomics approach is the limited sensitivity of mass spectrometry (Frankiw et al., 2019). This is further exacerbated by the fact that HLA-I peptides all possess a similar sequence and length, making their identification from mass spectra very challenging. Indeed, when I compared the replicate overlap of sample groups, I noticed that – dependent on the sample group – only 28-47% of peptides were shared by all replicates. I explored this in more detail and clustered samples based on their similarity, which revealed that the sample size of identified peptides was critical for sample similarity. However, the number of identified peptides strongly varied between replicates. Due to the use of enzyme-free detachment buffer, which led to substantial cell clumping, it was difficult to determine exact cell numbers, leading to a different amount of input material for HLA-I IP. Additionally, the efficacy of HLA-I IP can vary based on technical difficulties during the HLA-I IP workflow. Notably, although the original publication of the HLA-I IP workflow reports a sample overlap of 84% (Chong et al., 2018), recent studies either do not show sample overlap (Becker et al., 2021) or present barely overlapping replicates (Lu et al., 2021). These discrepancies potentially result from differences in the mass spectrometry data acquisition strategy, where more sensitive approaches lead to a greater sample diversity. Furthermore, the low reproducibility across samples can be caused by the data-driven acquisition method used for mass spectrometry, as explained later in Section 4.1.5.

The limited mass spectrometry sensitivity highlights the challenge of detecting low abundant peptides robustly across all replicates. Thus, neoepitope identification started by comparing all peptides that passed quality control procedures. Given the limited sensitivity of mass spectrometry, in principle, all identified peptides could be considered as neoepitopes for further analysis. Indeed, a recent similar study followed this approach to broaden the candidate search space (Lu et al., 2021). Nevertheless, I decided to limit the candidate pool to neoepitopes detected in at least two out of three replicates per sample group. This allowed me to identify strong peptide candidates without ambiguities which were thus suitable for *in vivo* immunization experiments.

4.1.3 Custom reference databases broaden the search space for peptide identification

During peptide identification by *PEAKS* search, I determined peptides from mass spectra by *de novo* peptide sequencing. These *de novo* peptide sequences were then confirmed by matching them against a reference database. First, I used the UniProt database as reference because it already contains manually curated protein information from SwissProt as well as computationally annotated entries from RefSeq (Bateman et al., 2021). By applying a stringent cut-off of 1% FDR for identified peptides, I obtained thousands of HLA-I-presented peptides as confirmed by quality control steps.

However, this did not include sample-specific peptides potentially generated by aberrant splicing induced by GEX1A treatment. Thus, I designed custom reference libraries that facilitated the identification of splicing-derived peptides by following two different approaches. First, I generated the *Splice Junction-spanning DataBase* (SJDB), consisting of short peptides representing all potential peptide sequences that could be generated from detected splice junctions in the samples. However, this database only covered splice junctions and there was no reading frame control when the genomic coordinates were translated into peptide sequences. Given the limited database size and the high proportion of potential false-positive entries, the SJDB was not suitable as primary reference database. I addressed the shortcomings of the SJDB by establishing a workflow that generated a sample-specific *de novo* proteome. By adapting the *TransDecoder* part of the *ProteomeGenerator* framework (Cifani et al., 2018), I was able to generate a robust sample-specific proteomic reference database, including peptide sequences derived from novel splice junctions. This database showed a similar performance as UniProt during *PEAKS* search, which was reflected by similar FDR values. Notably, a high proportion (> 98%) of *TransDecoder* peptides were already detected in UniProt, which was in line with published data (Cifani et al., 2018).

Eventually, I decided to perform a concatenated, multi-round *PEAKS* search against canonical and custom databases to allow peptide identification from the largest possible search space. The custom peptide reference libraries allowed me to identify hundreds of sample-specific peptides additionally. The presence of these peptides was confirmed by quality control steps in which peptides exhibited similar quality features as peptides identified by UniProt. Importantly, although these peptides accounted only for 1% of total peptides found, they represented 11 of 23 neoepitope candidates. Notably, after manual filtering, all remaining candidates were peptides identified from custom databases. These findings are supported by a previously published strategy, which also generated sample-specific reference databases to facilitate the identification of non-canonical peptides from mass spectrometry datasets (Chong et al., 2020). Hence, a sample-specific proteomic reference is essential for a comprehensive identification of neoepitopes from aberrant splice events.

4.1.4 Peptide identification bias towards highly expressed genes

As discussed above, the immunopeptidomics approach is limited by the sensitivity of mass spectrometry data acquisition methods (Frankiw et al., 2019). At the same time, most aberrant splice events, such as intron retention events, are found in low abundant transcripts (Pickrell et al., 2010). By matching the peptide information to gene expression data, I discovered that most peptides were indeed derived from highly expressed genes. Accordingly, genes matched to GEX1A-associated neoepitopes were differentially higher expressed in the GEX1A-treated condition. These data suggest that most potential neoepitopes result from gene expression changes rather than differential alternative splicing. This explains why I could only identify 23 splicing-derived peptide candidates from hundreds of detected neoepitopes in the GEX1A-treated condition. The majority of splicing-derived neoepitope candidates I identified were in-frame exon skipping events causing no disruption of the canonical reading frame. Since these transcripts are not targeted by nonsense-mediated decay, they occur more abundantly than transcripts derived from intron retention events or other splice events, causing reading frame errors and PTCs.

The bias of peptide detection towards highly expressed gene products highlights the limitation of the mass spectrometry-driven approach. However, mass spectrometry is the only analytical methodology that allows the direct identification of HLA-I peptides naturally presented *in vivo* (Bassani-Sternberg and Coukos, 2016; Chong et al., 2020). Due to the advantages of immunopeptidomics for identifying neoepitopes, researchers in the field aim to improve the methodology for mass spectrometry data acquisition, as discussed in Section 4.1.5.

4.1.5 Data-independent acquisition during mass spectrometry enables immunopeptidomics for translational research

The limited sensitivity of mass spectrometry prevents the identification of splicing-derived neoepitopes (Frankiw et al., 2019). Using immunopeptidomics, I could overcome technical burdens like the cleavage bias of Trypsin towards splice junctions. Still, discovery-driven peptide detection by mass spectrometry is biased towards abundant peptides of highly expressed transcripts, as discussed in Section 4.1.4. For explorative searches, this and other studies use data-dependent tandem acquisition (DDA) during mass spectrometry measurements. On the one hand, this method focuses on isolating peptide *m/z* before tandem mass spectrometry and thereby provides high-quality references of peptide fingerprints (Gatlin et al., 2000). On the other hand, the DDA method stochastically selects abundant ions and thus has a bias towards the detection of abundant peptides leading to an impaired sensitivity and reproducibility (Pak et al., 2021). This could be solved by using data-independent acquisition (DIA), which provides a sample-specific comprehensive digital map of fragment ions within given isolation *m/z* windows (Bruderer et al., 2017). These spectra are then compared to spectral libraries obtained from DDA datasets. Given the limited amount of input material retrieved from patients, the generation of DDA-based libraries is impractical, preventing the implementation of DIA for immunopeptidomics in translational research (Pak et al., 2021). In a recent study, Pak and colleagues developed a DIA immunopeptidomics workflow combining several data sources into a so-called multi-HLA spectral library. Implementing this, they significantly improved peptide identification from DIA data and were able to detect clinically relevant HLA ligands (Pak et al., 2021). These data demonstrate the ability of DIA to increase the sensitivity and reproducibility of immunopeptidomics while requiring less input material, thereby also enabling the use of patient biopsies samples. Moreover, a more sensitive mass spectrometry approach will also facilitate to overcome the detection bias towards highly abundant peptides, which was observed in this study. Taken together, the progress in mass spectrometry methodology will enable the application of immunopeptidomics for translational purposes in the future.

The technological advances in the mass spectrometry instruments and methodology will allow a more comprehensive and effective detection of HLA-I peptides, making immunopeptidomics a highly attractive approach for future studies.

4.2 Differential splicing analysis

4.2.1 GEX1A induces pronounced changes in the alternative splicing landscape

Recent studies reported altered alternative splicing patterns in tumor cells resulting from mutations and dysregulated expression of spliceosomal components and regulators (Baeza-Centurion et al., 2019; Jayasinghe et al., 2018). This overburdens the splicing machinery in cancer cells and causes alternative junction usage in patients (Kahles et al., 2018). In addition, tumors frequently express 'neojunctions' which are not typically found in normal tissue and are particularly attractive targets for neoepitope discovery (Kahles et al., 2018). By treating cells with splicing inhibitors, I aimed to recapitulate the aberrant alternative splicing phenotype observed in cancer cells.

Indeed, GEX1A treatment effectively induced splicing aberrancies reflected by several thousands of significant differential splice events, from which approximately two-thirds of them were caused by exon skipping. This was contradictory to findings of a previous study where researchers reported that treatment with the splicing inhibitor H3B-8800 mainly caused intron retention events, followed by exon skipping events (Seiler et al., 2018b). Like GEX1A, H3B-8800 broadly inhibits splicing by binding *SF3B1* to prevent the assembly of the spliceosome. The reason for the discrepancy between the numbers of detected intron retention events potentially lies in the detection criteria. For example, Seiler et al. did not introduce log fold change cut-offs to their data. When I re-analyzed their dataset using my parameters, I obtained mostly exon skipping events but notably almost no intron retention events (data not shown). Generally, intron retention events are not expected to be detected this frequently because they often cause PTCs and are thus degraded by NMD (Pickrell et al., 2010). Hence, in-frame exon skipping events are most promising for delivering a sufficient amount of protein that can be detected by mass spectrometry.

Initially, I used the GR_{25} dose of GEX1A to recapitulate the aberrant alternative splicing phenotype in K562 cells for proof-of-concept experiments. Since I observed no lethality during the short time of drug treatment, I decided to use the GR_{75} dose in RPE-1 cells to obtain an even more pronounced phenotype. Moreover, the RPE-1 cells with genetic deficiencies were more susceptible to GEX1A treatment. Consequently, the same GEX1A dose caused more prevalent differential splicing in RPE-1 *TP53*$^{-/-}$ cells.

Notably, the treatment-induced response resulted in a more pronounced aberrant splicing phenotype than mutation-induced changes, which was in line with reported findings (Seiler et al., 2018b). IFN-γ stimulation was able to mask the subtle splicing changes in K562 cells carrying *SF3B1* mutations. Consequently, I decided to focus on splicing inhibitor GEX1A, which induced substantial differential splicing changes in all cell models. The clinical implications of splicing inhibitors are further discussed in Section 4.5.

4.2.2 *DJExpress* is most suitable for detecting neoepitope-generating differential splice events

To assess changes in the alternative splicing landscape, I performed a Salmon analysis based on isoform abundance. However, this approach did not include the detection of novel splice junctions. These were of particular interest since they can give rise to novel, (cancer-)specific peptide sequences and frequently occur in cancer patients (Kahles et al., 2018). Thus, I switched to count-based differential splicing analysis methods, which incorporate information on novel splice junctions obtained from *STAR* alignment. The first computational tool I used to detect differential splice events was *rMATS*. This tool provided inclusion level differences of splice events between two sample groups based on percent spliced in (PSI) values, reflecting the percentage variant transcripts relative to all transcripts covering the junction of interest. This approach allowed me to detect highly abundant splice events that appear upregulated compared to the canonical event. However, especially novel, non-annotated or cryptic splice events do not occur very frequently within transcripts' structure. Although these splice events might not be highly relevant for the biological function of a gene/protein in the cell, the aberrant transcripts can be of high interest if they are translated into non-canonical peptide sequences. While the *rMATS* approach was helpful to determine different splice event types, such as exon skipping or intron retention events, it was limited in the ability to identify all relevant differential splice events in my datasets.

To be able to include all relevant splice events for further evaluations, I performed *JunctionSeq* analysis to detect changes in splicing patterns involving lowly expressed exon-exon junctions. Here, significant splice events were determined by fold change differences of splice junction expression levels between the sample groups instead of inclusion level differences. I obtained ten thousand differentially expressed splice events, including low abundant and novel, non-annotated splice events. The *JunctionSeq* analysis provided information about all splice events found differentially expressed between two sample groups. However, it was not designed to take into account differential gene expression (Hartley and Mullikin, 2016). This was an important point to consider since GEX1A

treatment resulted in altered gene expression between sample groups and most of the detected neoepitopes originated from those differentially expressed genes (also see Section 4.1.4). Thus, it was crucial to determine whether the differences in splice junction expression were genuinely caused by junction usage or simply resulted from changes in total gene expression in order to identify *bona fide* splicing-derived neoepitopes.

To include information about gene expression changes in the differential splicing analysis, our research group developed the *limma*-based *DJExpress* pipeline (Gallego-Paez & Mauer, 2022). Using this tool, I was able to determine whether the altered frequency of a detected splice event was explained by gene expression changes or by junction usage between the sample groups because the analysis allowed me to evaluate the expression of every splice event in the context of junction expression across a whole gene. Eventually, this approach allowed me to specifically select only events with differential junction usage.

4.2.3 Limitations of count-based identification of splice events

One challenging aspect of the count-based differential splicing analysis was that I was only able to detect a specific genomic region of a gene but could not unambiguously determine the respective transcript isoform (Figure 53). This introduced ambiguities when I translated the splice events into peptide sequences before matching them to identified neoepitopes. Since no reading frame control was available, these peptide libraries contained more than twice as many peptides than splice events. Thus, I had to consider that potentially non-existing peptides were also matched with neoepitopes. Moreover, the peptide reference libraries lacked coverage because these were limited to splice junction-spanning peptides originating from upregulated junctions.

These shortcomings could be solved by isoform-based approaches for splicing analysis, where full-length gene isoforms are accurately reconstructed. However, these tools are generally not suitable for detecting non-annotated splice events. One strategy that should be considered for future studies is the *IsoformSwitchAnalyzeR* workflow which determines isoform usage, for instance from *Salmon* analysis based on a sample-specific assembled transcriptome reference including novel splice junctions (Vitting-Seerup and Sandelin, 2018).

Alternatively, full-length transcripts can be obtained from long-read RNA-seq. However, this is a cost-intensive method and was therefore not applicable to this large dataset. A cost-saving alternative would be *Nanopore MinION* sequencing which has become more established over the past years. For example, a recent study demonstrated that this method can provide robust information on isoform expression to identify differential splice events (Oka et al., 2021). However, until very recently, *MinION* sequencing achieves approximately only 90% sequencing accuracy (Seki et al., 2019), which is sufficient for the precise extraction of splice patterns but not for the identification of neoepitopes where accurate base calling of genomic sequences is critical. Consequently, Oka *et al.* combined short-read and long-read RNA-seq in their study to identify splicing-derived neoepitopes. Nevertheless, with the continuing progress in long-read RNA-seq technologies, these will become the method of choice to detect differences in alternative splicing patterns for neoepitope identification.

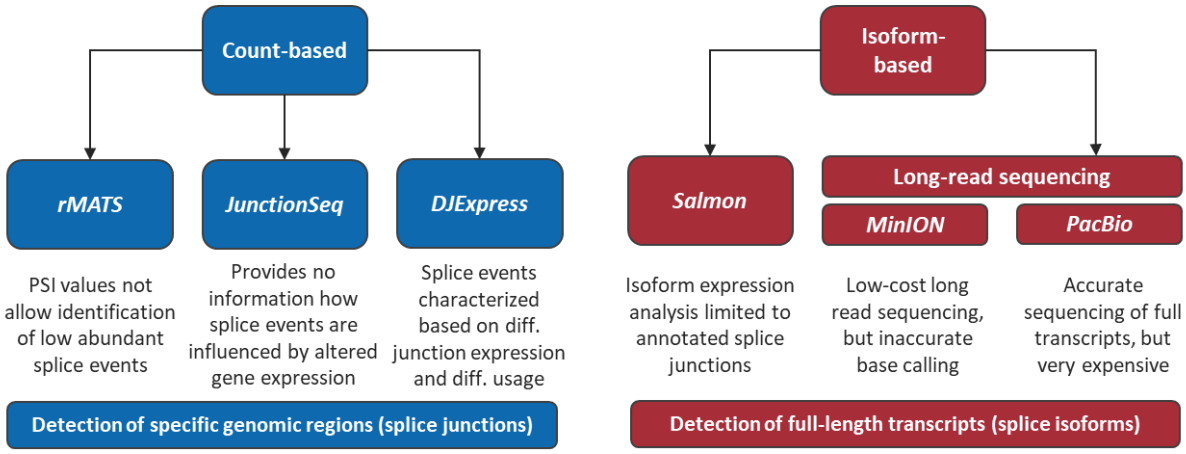


Figure 53 | Overview on count-based and isoform-based approaches for differential alternative splicing analysis. *Count-based tools were chosen to enable the identification of non-annotated alternative splice events. PSI = percent spliced in; LFC = log fold change.*

4.3 Identification of splicing-derived neoepitopes

4.3.1 Cell lines with genetic deficiencies in the DNA damage repair pathway produce more splicing-derived neoepitopes

By merging information from immunopeptidomics and differential splicing analysis, I could identify 23 individual splicing-derived neoepitope candidates across all RPE-1 cell lines. Here, the proportion of splicing-derived neoepitopes discovered from neoepitope candidates was higher in cell lines carrying genetic deficiencies in the DNA damage repair pathway. This is likely caused by a more pronounced aberrant splicing phenotype due to the increased susceptibility of these cell lines to GEX1A splicing inhibition. However, I did not analyze this at transcriptome level since the RNA-seq analysis was limited to RPE-1 WT and *TP53*$^{-/-}$ cell lines. Nevertheless, differential splicing analyses revealed that already loss of *TP53*$^{-/-}$ was able to cause a more prevalent aberrant splicing phenotype in response to splicing inhibition.

In addition, the genetic deficiencies in components of the mismatch repair (MMR) pathway affect the genomic stability and facilitate the generation of aberrant transcripts (Bokhari et al., 2018). This affects the efficiency of the NMD pathway because it cannot cope with the number of error-prone transcripts. Consequently, more transcripts will escape this quality control mechanism explaining the higher proportion of splicing-derived neoepitopes in MMR-deficient cell lines. Accordingly, escape from NMD is connected to anti-tumor immunogenicity in clinical datasets (Litch et al., 2020) and modulating the NMD pathway to enhance encoding of immunogenic peptides is of great interest for cancer immunotherapy, particularly for microsatellite-instable colorectal cancer (Becker et al., 2021; Bokhari et al., 2018; Bongiorno et al., 2021). Given the interplay between mRNA transcription and the splicing machinery, modulating spliceosomal activity would be an equally attractive target to increase the number of neoepitope-producing transcripts in tumor cells.

Manual filtering was still a critical step during candidate selection. Although all peptide candidates passed BLASTp search and HLA-I binding predicting, only 10 out of 23 peptides were derived from a novel splicing isoform. The other 13 peptide sequences matched the canonical isoform of the protein. This happened due to technical limitations during the generation of the *Alternative Splicing Peptide Libraries*. Since the genomic coordinates were extended by 29 nt, nine aa could match the canonical protein sequence. Therefore, the database also identified HLA-I peptides which were not derived from an alternative splice isoform. These peptides were most likely not detected in control samples due to gene expression changes between the sample groups, as discussed earlier (Section 4.1.4). This

shortcoming would be addressed by implementing an isoform-based approach for the generation of the *Alternative Splicing Peptide Libraries* to obtain full-length transcripts directly. Until then, manual curation of peptides candidates remains a critical step in this workflow to eliminate peptides matching canonical sequences.

4.3.2 The workflow allows robust detection of splicing-derived neoepitope candidates

The final neoepitope candidates were successfully validated on splicing and peptide level. Due to the limited sensitivity and specificity of the LC-MS/MS approach, the detection of short peptides with similar sequence patterns can still be error-prone. I was able to successfully confirm all peptide sequences of neoepitope candidates by comparing their experimental mass spectra to synthetic references. This demonstrated that the mass spectrometry data acquisition and the peptide identification from mass spectra using *PEAKS* both worked accurately.

Moreover, I explored the abundance of differential splice events at transcript level by generating Sashimi plots. This revealed that the *ANXA1* exon skipping event was the most promising candidate given its high expression in response to GEX1A treatment while it was (almost) absent in the control condition. Although all splice events showed expression differences between the conditions, a relatively high read coverage in the control condition was detected for the MXE event of *ATP6V0E1*. Detection of such events could be avoided by only choosing 'neojunctions' from the *DJExpress* results, which exclusively appear in the GEX1A-treated condition. However, these junctions were found to be expressed at low levels. Thus, it was unlikely to find matching peptide evidence in the dataset considering the limited sensitivity of mass spectrometry. In contrast, even a few mis-spliced transcripts in the control condition will be detected by a highly sensitive method like RNA-seq. However, it is questionable whether these few transcripts will be translated into peptide sequences (apart from the peptides from pioneer translation rounds, which are considerably rare events). Therefore, I decided to keep splice events with very low read evidence in the control condition if there was a distinct expression difference between the sample groups. The same rationale was followed by a study similar to this one (Lu et al., 2021).

To further address whether the splice events have relevance in the control condition, I performed an independent experiment to validate detected splice events *in vitro* by RT-PCR. Here, I did not detect any candidate splice events in the untreated condition. Instead, I could observe that the splice events occurred in a concentration-dependent manner in response to GEX1A treatment. Importantly, also the

presence of novel, non-annotated splice events was confirmed. These validation experiments demonstrated the accuracy of *DJExpress* analysis to determine differential splice events. Overall, the combination of immunopeptidomics and differential splicing analysis accurately delivered promising splicing-derived neoepitope candidates.

4.4 Clinical implications of splicing-derived neoepitopes

Neoepitopes originating from dysregulated splice events are particularly attractive targets for immunotherapeutic approaches because they were described to occur highly frequently in tumors (Kahles et al., 2018) and are predicted to be shared between patients (Pellagatti et al., 2018). In addition, previous studies which explored dysregulated alternative splicing across tumors of TCGA patient data suggested a high immunogenic potential of splicing-derived neoepitopes (Jayasinghe et al., 2018; Kahles et al., 2018). Using a splice variant expression system, it was shown that pharmacological splicing modulation results in an altered immunopeptidome and thereby enhances antitumor immune response (Darrigrand et al., 2021). Other recent studies evaluated specific targets and showed the ability of splicing-derived neoepitopes to activate T cell response in mice (Lu et al., 2021; Oka et al., 2021). Thus, I wanted to evaluate whether my approach also successfully identified immunogenic neoepitope candidates that could be considered for future clinical approaches.

4.4.1 Identified splice events demonstrate clinical potential

Since I investigated spliceosomal perturbations in cell models, I was curious to explore the clinical potential of differential splice events by detecting their expression in TCGA patient data. Indeed, I discovered that six out of nine detected splice events were significantly higher expressed in cancer patients across multiple cohorts. Notably, the other three events could also be present in patients but showed no differential expression between cancer and normal tissue. In a few cohorts, high expression of the neoepitope-producing isoform was associated with better disease outcome. This could potentially be explained by a higher immunogenicity of the tumors due to the generation of a splicing-derived neoepitope. A previous study suggested that alternative splice isoforms are highly immunogenic because they were found to correlate with high T cell immune response and elevated PD-L1 expression (Jayasinghe et al., 2018). However, I did not observe a correlation between the expression of the neoepitope-producing variant of *ANXA1* and immune infiltration in TCGA cancer

patients. Alternatively, the increased survival of patients expressing the skipped *ANXA1* variant might be explained by an altered biological function of the transcript isoform.
This was particularly interesting for *ANXA1*, as the alternative transcripts were highly abundant. ANXA1 is involved in a variety of different cellular functions and can play an ambivalent role in cancer (Foo et al., 2019). At structural level, skipping of exon 8 shortens the AXNA1 repeat 3 (Rosengarth et al., 2003), which is involved in the activation of ANXA1 for signal transduction (Han et al., 2020). Further studies will be necessary to unravel the biological role of this novel *ANXA1* isoform.

Apart from *ANXA1*, the expression of the other splice variants might not have a strong biological relevance because these are expressed at a much lower level compared to the canonical isoform. In any case, due to their differential expression in cancer patients, the identified splice events are promising targets for future immunotherapeutic approaches.

4.4.2 Splicing-derived neoepitopes show high immunogenic potential

Ultimately, peptide immunogenicity will determine the suitability of the identified neoepitope candidates as potential targets for immunotherapeutic approaches. I followed the strategy of other publications in the field (Becker et al., 2021; Lu et al., 2021; Oka et al., 2021) to evaluate the T cell response towards the neoepitope candidates by *in vivo* immunization assays using the HLA-A*02:01-transgenic mouse model (Pajot et al., 2004). This model was restricted to one HLA-I subtype, preventing me from investigating the immunogenic potential of all candidates. Still, HLA-A*02:01 is the most frequently occurring HLA-I subtype in humans and thus has the highest clinical relevance. Using HLA-I binding prediction, I identified five out of ten candidates as predicted HLA-A*02:01 binders. However, the splice event of *ATP6V0E1* was also detected in non-treated samples at a relatively high level and thus, the corresponding neoepitope candidate was not considered for immunization experiments. Instead, I included the peptide candidate of *PSMA4* because it was predicted to bind the sister allele HLA-A*01:01 with a similar binding motif. However, the *PSMA4* candidate was the only of the five peptides that could not activate a T cell response. Apparently, the binding affinity of *PSMA4* towards HLA-A*02:01 was too weak and thus, it was not presented on the cell surface. The four predicted HLA-A*02:01 binders all induced a strong T cell response captured by ELISpot measurements. This highlights not only the binding specificity of peptides for their respective HLA-I subtypes but also that peptides originating from differential splice events are able to induce an immune response.

In collaboration with Alejandro Hernandez-Sanchez, I further investigated the immunogenic potential of the splicing-derived neoepitope candidates. We observed that the candidate peptides caused specific T cell-mediated elimination of peptide-presenting target cells. This demonstrated that the identified splicing-derived peptides have indeed a high immunogenic potential. Of note, this mouse model provided evidence that peptides can interact with T cells and induce an immune response. However, it was not suitable to confirm the immunogenicity of the peptides in an environment with matured, haplotype-specific human T cells. This would be important to demonstrate considering the severe off-target effects of cell-based immunotherapies (Frankiw et al., 2019; Yarchoan et al., 2017b). Ultimately, a more sophisticated model needs to evaluate whether the neoepitope candidates show similar tumor-specific immunogenic potential in patients. Suitable protocols have been established to identify immunogenic frame-shift peptides (Leoni et al., 2020), but these were out of scope for this study. Nevertheless, the results from the *in vivo* immunization experiments strongly suggest that splicing perturbation broadens the neoepitope target space by generating highly immunogenic peptides derived from differential splice events.

4.4.3 The potential role of splicing-derived neoepitopes for cancer immunotherapy

The identified immunogenic neoepitope candidates could serve, for instance as targets for personalized adoptive cell therapies. Alternatively, it was suggested that splicing-derived neoepitopes increase the immunogenicity of a tumor cell and thereby facilitate the response to immune checkpoint inhibition therapies. A recent study demonstrated that combination treatment with anti-PD1 antibody and splicing inhibitor indisulam enhances anti-tumor response in mice and even has the potential to effectively target PD1-resistant tumors (Lu et al., 2021). Here, the researchers followed a strategy that combined immunopeptidomics with peptide identification from sample-specific reference databases, similar to the workflow presented in this thesis. However, they used different computational tools to generate sample-specific proteomes and predict differential splice events in mice. Nevertheless, the published data support the rationale of this thesis since the authors demonstrated that pharmacological modulation of splicing increases the anti-tumor immunogenicity (Lu et al., 2021).

In the future, it will be decided whether pharmacological splicing modulation can be utilized to improve the response of existing immunotherapeutic approaches. Several compounds, such as GEX1A and H3B-8800, were found to inhibit the assembly of the core spliceosome by binding *SF3B1* (Seiler et al., 2018b; Sellin et al., 2019). Although several splicing inhibitor candidates were in clinical trials, none of them has been approved yet due to severe side effects. For example, patients treated with splicing inhibitor

E7107 unexpectedly exhibited toxicity of bilateral optic neuritis (Hong et al., 2013). Notably, in a more recent clinical trial of H3B-8800, no such side effects were observed (Steensma et al., 2021). Still, drugs that unspecifically inhibit splicing can affect not only tumors but also immune and hematopoietic compartments. Indeed, it was reported that these inhibitors were immunosuppressive, while more specific splicing modulators like indisulam and MS-023 were nontoxic at therapeutic concentrations (Lu et al., 2021). Future studies will show whether therapeutic targeting of RNA splicing can facilitate the clinical response to cancer immunotherapies.

Moreover, the tumor splicing burden could become a predictive biomarker for response to immune checkpoint inhibition therapy. The TMB already serves as an important biomarker for clinical outcome, likely because the frequency of somatic mutation correlates with the number of neoepitopes that can be targeted by the immune system. However, there was no correlation between the tumor splicing burden and the response to immune checkpoint inhibitors detected so far (Jayasinghe et al., 2018; Oka et al., 2021). More comprehensive studies, including data from clinical trials of different solid and hematological tumor types, will be necessary for further evaluation.

4.5 Conclusions and perspectives on splicing-derived neoepitopes

In this study, I established a workflow to specifically identify and validate neoepitope candidates derived from alternative splice variants. This was motivated by findings reporting that the process of alternative splicing is widely dysregulated across several cancer types resulting in cryptic splice sites that can give rise to neoepitopes (Jayasinghe et al., 2018). Accordingly, a comprehensive analysis of patient tumors revealed that the presence of novel splice junctions can generate peptides which increase the target space for immunotherapeutic approaches (Kahles et al., 2018). However, these studies solely on computational approaches that broadly explored the potential contribution of splicing-derived peptides to the neoepitope repertoire and suggested a high immunological potential of such candidates.

In 2018, a proof-of-concept study focusing on cancer-specific intron retention events suggested that splicing-derived peptides can be indeed presented by HLA-I on the cell surface (Smart et al., 2018). Accordingly, another study identified hundreds of tumor-specific, non-canonical peptides presented by HLA-I (Chong et al., 2020). However, these studies prioritized specific targets to evaluate the immunogenicity of splicing-derived neoepitopes.

To investigate the splicing-derived neoepitope repertoire, I increased the cellular splicing burden by inhibiting the spliceosomal assembly using the small molecule inhibitor GEX1A. With this, I aimed to increase splicing-derived neoepitope presentation since it was reported that pharmacological splicing modulation using a splice variant expression system results in an altered immunopeptidome and thereby enhances antitumor immune response (Darrigrand et al., 2021).

The changes in the alternative splicing landscape induced by splicing modulation were determined by alternative splicing analysis. Here, I used the *DJExpress* pipeline that allowed me to robustly identify differentially expressed splice events independently of altered gene expression (Gallego-Paez & Mauer, 2022). However, this count-based analysis introduced ambiguities when translating genomic information into peptide sequences. This shortcoming could be addressed by using long-read sequencing data that facilitates the identification of transcript variants. A recent study used a combination of *MinION* sequencing and short-read sequencing to detect aberrant splice isoforms, which generate potential neoepitopes in small cell lung cancer (Oka et al., 2021). Future efforts will make long-read sequencing more accessible and robust for the identification of neoepitopes derived from splice variants.

Following the immunopeptidomics workflow, I was able to recover sample-specific HLA-I ligandomes. Here, it was particularly helpful to generate sample-specific reference libraries that enabled the identification of non-annotated splicing-derived peptides from mass spectra. Notably, all final neoepitope candidates were obtained from such custom databases highlighting the importance of sample-specific references for detecting peptides originating from non-annotated splice events. However, the immunopeptidomics approach was limited by its bias towards peptides derived from highly expressed genes. Since most alternative splicing events are found in low abundant transcripts (Pickrell et al., 2010), the contribution of splicing-derived neoepitopes to the immunopeptidome might be even more significant than it could be accessed in this study. Conventionally, it was thought that most aberrant splice events result in non-functional transcripts since these are targeted by NMD. Notably, it was reported that degradation occurs after a pioneer round of translation, making it the major source of antigenic peptides presented on HLA-I molecules (Apcher et al., 2011). Improved mass spectrometry methodology will benefit from an enhanced sensitivity and facilitate unraveling the overall contribution of splicing-derived peptides to the total neoepitope repertoire (Pak et al., 2021). Moreover, due to its limited sensitivity, the current immunopeptidomics approach requires extensive biological input material, making it incompatible with tumor biopsies material of patients (Bassani-Sternberg and Coukos, 2016). The technological advances in the mass spectrometry instruments and methodology will increase the sensitivity and reproducibility of immunopeptidomics and make it accessible for translational research.

Following the combined strategy of mass-spectrometry-based immunopeptidomics and differential splicing analysis, I was able to identify several neoepitope candidates that originated from aberrant splice events. I further demonstrated their high immunogenic potential by performing *in vivo* immunization assays. Here, the neoepitope candidates were able to induce T cell response as well as T cell-mediated killing if presented by HLA-I on the surface of target cells.

Hence, my findings suggest that splicing-derived neoepitopes are promising targets for immunotherapeutic approaches. Beyond targeting these candidates for personalized adoptive cell therapies, they could enhance the intrinsic T cell response in combination with established immune checkpoint blockade therapies. Indeed, a very recent study investigating the benefits of such a combination therapy observed an increased anti-tumor immune response in mice and could also effectively target PD1-resistant tumors (Lu et al., 2021). This study indicates that the enhanced immune response results from an increased splicing burden, leading to aberrant splice events giving rise to neoepitopes. Here, Lu et al. followed a strategy that combined immunopeptidomics with

peptide identification from sample-specific reference databases, thereby supporting the rationale of the approach I developed in this thesis.

Despite these promising results of this combination therapy, one needs to keep in mind that no splicing inhibitor has been approved for therapeutic use in patients yet. Although several splicing inhibitor candidates were in clinical trials, none of them succeeded due to severe side effects (Wang and Aifantis, 2020). While these off-target effects resulted from broad inhibition of the splicing machinery, the new generation of splicing inhibitors is more selective, thereby avoiding unwanted side effects. Future studies will evaluate the clinical potential of targeting the RNA splicing machinery for cancer (immuno)therapy.

Taken together, the presented work provides a workflow to specifically detect splicing-derived neoepitopes with high immunogenic potential. The data demonstrate that pharmacological splicing modulation increases the anti-tumor immunogenicity by promoting the presentation of neoepitopes derived from alternative splice variants. These findings have potential implications for immunotherapy of cancer types with low tumor mutational burden, where exploration of the splicing-derived neoepitope repertoire could reveal novel therapeutic targets. In addition, the tumor splicing burden could become a predictive biomarker for response to immune checkpoint inhibitor therapy.

5. Literature references

Adams, S., Robbins, F.M., Chen, D., Wagage, D., Holbeck, S.L., Morse, H.C., Stroncek, D., and Marincola, F.M. (2005). HLA class I and II genotype of the NCI-60 cell lines. J. Transl. Med. *3*, 1–8.

Aggarwal, C., Cohen, R.B., Morrow, M.P., Kraynyak, K.A., Sylvester, A.J., Knoblock, D.M., Bauml, J.M., Weinstein, G.S., Lin, A., Boyer, J., et al. (2019). Immunotherapy targeting HPV16/18 generates potent immune responses in HPV-associated head and neck cancer. Clin. Cancer Res. *25*, 110–124.

Andreatta, M., Alvarez, B., and Nielsen, M. (2017). GibbsCluster: Unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res. *45*, W458–W463.

Apcher, S., Daskalogianni, C., Lejeune, F., Manoury, B., Imhoos, G., and Heslop, L. (2011). Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *108*.

Baeza-Centurion, P., Miñana, B., Schmiedel, J.M., Valcárcel, J., and Lehner, B. (2019). Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. Cell *176*, 549-563.e23.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. *18*, 437–451.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Barnstable, C.J., Bodmer, W.F., and Brown, G. (1978). Production of Monoclonal Antibodies to Group A Erythrocytes , HLA and Other Human Cell Surface Tools for Genetic Analysis. *14*, 9–20.

Bassani-Sternberg, M. (2018). Mass spectrometry based immunopeptidomics for the discovery of cancer neoantigens. Methods Mol. Biol. *1719*, 209–221.

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

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H.S., Gannon, P.O., Kandalaft, L.E., Coukos, G., and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput. Biol. *13*, e1005725.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., et al. (2021). UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.

Becker, J.P., Helm, D., Rettel, M., Stein, F., Hernandez-Sanchez, A., Urban, K., Gebert, J., Kloor, M.,

Neu-Yilik, G., von Knebel Doeberitz, M., et al. (2021). NMD inhibition by 5-azacytidine augments presentation of immunogenic frameshift-derived neoepitopes. IScience *24*, 102389.

Boegel, S., Löwer, M., Bukur, T., Sahin, U., and Castle, J.C. (2014). A catalog of HLA type, HLA expression, and neoepitope candidates in human cancer cell lines. Oncoimmunology *3*, 1–12.

Bokhari, A., Jonchere, V., Lagrange, A., Bertrand, R., Svrcek, M., Marisa, L., Buhard, O., Greene, M., Demidova, A., Jia, J., et al. (2018). Targeting nonsense-mediated mRNA decay in colorectal cancers with microsatellite instability. Oncogenesis *7*.

Bongiorno, R., Colombo, M.P., and Lecis, D. (2021). Deciphering the nonsense-mediated mRNA decay pathway to identify cancer cell vulnerabilities for effective cancer therapy. J. Exp. Clin. Cancer Res. *40*.

Bonnal, S.C., López-Oreja, I., and Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer — implications for care. Nat. Rev. Clin. Oncol. *17*, 457–474.

Bruderer, R., Bernhardt, O.M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-varela, D., and Reiter, L. (2017). Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results *. Mol. Cell. Proteomics *16*, 2296–2309.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics *10*, 1–9.

Cameron, B.J., Gerry, A.B., Dukes, J., Harper, J. V., Kannan, V., Bianchi, F.C., Grand, F., Brewer, J.E., Gupta, M., Plesa, G., et al. (2013). Identification of a titin-derived HLA-A1-presented peptide as a cross-reactive target for engineered MAGE A3-directed T cells. Sci. Transl. Med.

Caron, E., Kowalewski, D.J., Koh, C.C., Sturm, T., Schuster, H., and Aebersold, R. (2015). Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. Mol. Cell. Proteomics.

Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. 207–232.

Chong, C., Marino, F., Pak, H., Racle, J., Daniel, R.T., Mü Ller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2018). High-throughput and sensitive immunopeptidomics platform reveals profound interferon γ-mediated remodeling of the human leukocyte antigen (HLA) ligandome. Mol. Cell. Proteomics *17*, 533–548.

Chong, C., Müller, M., Pak, H.S., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B.J., et al. (2020). Integrated proteogenomic deep sequencing and analytics accurately

identify non-canonical peptides in tumor immunopeptidomes. Nat. Commun. *11*.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell *12*, 1–8.

Cifani, P., Dhabaria, A., Chen, Z., Yoshimi, A., Kawaler, E., Abdel-Wahab, O., Poirier, J.T., and Kentsis, A. (2018). ProteomeGenerator: A Framework for Comprehensive Proteomics Based on de Novo Transcriptome Assembly and High-Accuracy Peptide Mass Spectral Matching. J. Proteome Res. *17*, 3681–3692.

Condino-Neto, A., and Newburger, P.E. (2000). Interferon-gamma improves splicing efficiency of CYBB gene transcripts in an interferon-responsive variant of chronic granulomatous disease due to a splice site consensus region mutation. Blood *95*, 3548–3554.

Coulie, P.G., Van Den Eynde, B.J., Van Der Bruggen, P., and Boon, T. (2014). Tumour antigens recognized by T lymphocytes: At the core of cancer immunotherapy. Nat. Rev. Cancer *14*, 135–146.

Couzin-Frankel, J. (2013). Breakthrough of the year 2013. Cancer Immunotherapy. Science (80-.). *342*, 1432–1433.

Danial, N.N. (2007). BCL-2 family proteins: Critical checkpoints of apoptotic cell death. Clin. Cancer Res. *13*, 7254–7263.

Darrigrand, R., Pierson, A., Rouillon, M., Renko, D., Boulpicante, M., Bouyssié, D., Mouton-barbosa, E., Marcoux, J., Garcia, C., Ghosh, M., et al. (2021). immune response against tumor antigens. Commun. Biol. 1–14.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R.K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. Nat. Rev. Cancer *16*, 413–430.

Farina, A.R., Cappabianca, L., Sebastiano, M., Zelli, V., Guadagni, S., and Mackay, A.R. (2020). Hypoxia-induced alternative splicing: The 11th Hallmark of Cancer. J. Exp. Clin. Cancer Res. *39*, 1–30.

Fedoriw, A., Rajapurkar, S.R., O'Brien, S., Gerhart, S. V., Mitchell, L.H., Adams, N.D., Rioux, N., Lingaraj, T., Ribich, S.A., Pappalardi, M.B., et al. (2019). Anti-tumor Activity of the Type I PRMT Inhibitor, GSK3368715, Synergizes with PRMT5 Inhibition through MTAP Loss. Cancer Cell *36*, 100-114.e25.

Foo, S.L., Yap, G., Cui, J., and Lim, L.H.K. (2019). Annexin-A1 – A Blessing or a Curse in Cancer ? Trends Mol. Med. *xx*, 1–13.

Fox, R.G., Lytle, N.K., Jaquish, D. V., Park, F.D., Ito, T., Bajaj, J., Koechlein, C.S., Zimdahl, B., Yano, M., Kopp, J.L., et al. (2016). Image-based detection and targeting of therapy resistance in pancreatic adenocarcinoma. Nature *534*, 407–411.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Frankiw, L., Baltimore, D., and Li, G. (2019). Alternative mRNA splicing in cancer immunotherapy. Nat. Rev. Immunol. *19*, 675–687.

Fu, X.D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Gallego-Paez, L.M., Bordone, M.C., Leote, A.C., Saraiva-Agostinho, N., Ascensão-Ferreira, M., and Barbosa-Morais, N.L. (2017). Alternative splicing: the pledge, the turn, and the prestige: The key role of alternative splicing in human biological systems. Hum. Genet. *136*, 1015–1042.

Gatlin, C.L., Eng, J.K., Cross, S.T., Detter, J.C., Iii, J.R.Y., Hb, C., Hb, E., Angeles, H.D., G-, H., Hope, H., et al. (2000). Automated Identification of Amino Acid Sequence Variations in Proteins by HPLC / Microspray Tandem Mass Spectrometry identify amino acid changes in six variants of human. *72*, 757–763.

Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R.T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018a). The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. J. Immunol.

Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R.T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018b). The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. J. Immunol. *201*, 3705–3716.

Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M. Dos, Jones, J., Takeshita, L., Ortega-Rivera, N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., et al. (2020). Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. *48*, D783–D788.

Gros, A., Parkhurst, M.R., Tran, E., Pasetto, A., Robbins, P.F., Ilyas, S., Prickett, T.D., Gartner, J.J., Crystal, J.S., Roberts, I.M., et al. (2016). Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. Nat. Med. *22*, 433–438.

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

Haas, B. (2018). TransDecoder.

Han, P., Che, X., Li, H., Gao, Y., Wei, X., and Li, P. (2020). Annexin A1 involved in the regulation of in fl ammation and cell signaling pathways. Chinese J. Traumatol. *23*, 96–101.

Han, T., Goralski, M., Gaskill, N., Capota, E., Kim, J., Ting, T.C., Xie, Y., Williams, N.S., and Nijhawan, D. (2017). Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. Science (80-.). *356*.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674.

Hartley, S.W., and Mullikin, J.C. (2015). QoRTs: A comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics *16*, 1–7.

Hartley, S.W., and Mullikin, J.C. (2016). Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. Nucleic Acids Res.

Havens, M.A., and Hastings, M.L. (2016). Splice-switching antisense oligonucleotides as therapeutic drugs. Nucleic Acids Res. *44*, 6549–6563.

Hong, D.S., Kurzrock, R., Naing, A., and Wheler, J.J. (2013). A phase I , open-label , single-arm , dose-escalation study of E7107 , a precursor messenger ribonucleic acid (pre-mRNA) splicesome inhibitor administered intravenously on days 1 and 8 every 21 days to patients with solid tumors.

Ilagan, J.O., Ramakrishnan, A., Hayes, B., Murphy, M.E., Zebari, A.S., Bradley, P., and Bradley, R.K. (2015). U2AF1 mutations alter splice site recognition in hematological malignancies. Genome Res. *25*, 14–26.

Inoue, D., Chew, G.L., Liu, B., Michel, B.C., Pangallo, J., D'Avino, A.R., Hitchman, T., North, K., Lee, S.C.W., Bitner, L., et al. (2019). Spliceosomal disruption of the non-canonical BAF complex in cancer. Nature *574*, 432–436.

Jayasinghe, R.G., Cao, S., Gao, Q., Wendl, M.C., Vo, N.S., Reynolds, S.M., Zhao, Y., Climente-González, H., Chai, S., Wang, F., et al. (2018). Systematic Analysis of Splice-Site-Creating Mutations in Cancer. Cell Rep. *23*, 270-281.e3.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. *47*, 1242–1248.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0:

Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J. Immunol. *199*, 3360–3368.

Kahles, A., Lehmann, K. Van, Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Caesar-Johnson, S.J., et al. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211-224.e6.

Kassambara, A., and Mundt, F. (2020). Package 'factoextra': Extract and visualize the results of multivariate data analyses. CRAN- R Packag. 84.

Kim, E., Ilagan, J.O., Liang, Y., Daubner, G.M., Lee, S.C.-W., Ramakrishnan, A., Li, Y., Chung, Y.R., Micol, J., Murphy, M.E., et al. (2015). SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. Cancer Cell *27*, 617–630.

Koh, C.M., Bezzi, M., Low, D.H.P., Ang, W.X., Teo, S.X., Gay, F.P.H., Al-Haddawi, M., Tan, S.Y., Osato, M., Sabò, A., et al. (2015). MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. Nature *523*, 96–100.

Krokhin, O. V., Ying, S., Cortens, J.P., Ghosh, D., Spicer, V., Ens, W., Standing, K.G., Beavis, R.C., and Wilkins, J.A. (2006). Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS. Anal. Chem.

Ladomery, M. (2013). Aberrant Alternative Splicing Is Another Hallmark of Cancer. Int. J. Cell Biol. *2013*, 1–6.

Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu. Rev. Biochem. *84*, 291–323.

Lennerz, V., Fatho, M., Gentilini, C., Frye, R.A., Lifke, A., Ferel, D., Wölfel, C., Huber, C., and Wölfel, T. (2005). The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. Proc. Natl. Acad. Sci. U. S. A. *102*, 16013–16018.

Leoni, G., Alise, A.M.D., Cotugno, G., Langone, F., Garzia, I., Lucia, M. De, Fichera, I., Vitale, R., Bignone, V., Tucci, F.G., et al. (2020). CANCER RESEARCH | TUMOR BIOLOGY AND IMMUNOLOGY A Genetic Vaccine Encoding Shared Cancer Neoantigens to Treat Tumors with Microsatellite Instability.

Liao, K.C., and Garcia-Blanco, M.A. (2021). Role of alternative splicing in regulating host response to viral infection. Cells *10*, 1–12.

Linette, G.P., Stadtmauer, E.A., Maus, M. V., Rapoport, A.P., Levine, B.L., Emery, L., Litzky, L., Bagg, A., Carreno, B.M., Cimino, P.J., et al. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. Blood.

Litch, K., Reading, J.L., Lim, E.L., Xu, H., Liu, P., Al-bakir, M., Ning, Y., Wong, S., Rowan, A., Funt, S.A., et al. (2020). Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. 1–11.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell *165*, 535–550.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 1–21.

Lu, S.X., De Neef, E., Thomas, J.D., Sabio, E., Rousseau, B., Gigoux, M., Knorr, D.A., Greenbaum, B., Elhanati, Y., Hogg, S.J., et al. (2021). Pharmacologic modulation of RNA splicing enhances anti-tumor immunity. Cell *184*, 4032-4047.e31.

Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., Sanada, M., Grossmann, V., Sundaresan, J., Shiraishi, Y., et al. (2015). Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. Nat. Commun. *6*.

Marsh, S.G.E., Albert, E.D., Bodmer, W.F., Bontrop, R.E., Dupont, B., Erlich, H.A., Fernández-Viña, M., Geraghty, D.E., Holdsworth, R., Hurley, C.K., et al. (2010). Nomenclature for factors of the HLA system, 2010. Tissue Antigens *75*, 291–455.

Morgan, R.A., Yang, J.C., Kitano, M., Dudley, M.E., Laurencot, C.M., and Rosenberg, S.A. (2010). Case report of a serious adverse event following the administration of t cells transduced with a chimeric antigen receptor recognizing ERBB2. Mol. Ther. *18*, 843–851.

Murphy, K., and Weaver, C. (2016). Janeway's Immunobiology. In Janeway's Immunobiology, (Garland Science), p.

Ohe, K., and Hagiwara, M. (2015). Modulation of alternative splicing with chemical compounds in new therapeutics for human diseases. ACS Chem. Biol. *10*, 914–924.

Oka, M., Xu, L., Suzuki, T., Yoshikawa, T., Sakamoto, H., and Uemura, H. (2021). Aberrant splicing isoforms detected by full- length transcriptome sequencing as transcripts of potential neoantigens in non- small cell lung cancer. 1–30.

Oltean, S., and Bates, D.O. (2014). Hallmarks of alternative splicing in cancer. Oncogene *33*, 5311–5318.

Pajot, A., Michel, M.L., Fazilleau, N., Pancré, V., Auriault, C., Ojcius, D.M., Lemonnier, F.A., and Lone, Y.C. (2004). A mouse model of human adaptive immune functions: HLA-A2.1-/HLA-DR1-transgenic H-2 class I-/class II-knockout mice. Eur. J. Immunol. *34*, 3060–3069.

Pak, H., Michaux, J., Huber, F., Chong, C., Stevenson, B.J., Müller, M., Coukos, G., Bassani-sternberg, M., Pak, H., Michaux, J., et al. (2021). Special Issue : Immunopeptidomics Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries , Data-Independent Acquisition , and MS / MS Prediction Authors Special Issue : Immunopeptidomics Sensitive Immunopeptidomics by. Mol Cell Proteomics *20*, 100080.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Pellagatti, A., Armstrong, R.N., Steeples, V., Sharma, E., Repapi, E., Singh, S., Sanchi, A., Radujkovic, A., Horn, P., Dolatshad, H., et al. (2018). Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood *132*, 1225–1240.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet.

Poulikakos, P.I., Zhang, C., Bollag, G., Shokat, K.M., and Rosen, N. (2010). RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. Nature *464*, 427–430.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Core Team (2018). A Language and Environment for Statistical Computing.

Rigo, F., Hua, Y., Krainer, A.R., and Frank Bennett, C. (2012). Antisense-based therapy for the treatment of spinal muscular atrophy. J. Cell Biol. *199*, 21–25.

Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G.E. (2015). The IPD and IMGT/HLA database: Allele variant databases. Nucleic Acids Res. *43*, D423–D431.

Rolfs, Z., Müller, M., Shortreed, M.R., Smith, L.M., and Bassani-Sternberg, M. (2019). Comment on “a subset of HLA-I peptides are not genomically templated: Evidence for cis- And trans-spliced peptide ligands.” Sci. Immunol.

Rosengarth, A., Luecke, H., and Hall, M. (2003). A Calcium-driven Conformational Switch of the N-terminal and Core Domains of Annexin A1. J. Mol. Biol. *2836*, 1317–1325.

van Rossum, G. (2015). Python.

Saltzman, A.L., Pan, Q., and Blencowe, B.J. (2011). Regulation of alternative splicing by the core spliceosomal machinery. Genes Dev. *25*, 373–384.

Schmidt, J., Guillaume, P., Dojcinovic, D., Karbach, J., Coukos, G., and Luescher, I. (2017). In silico and cell-based analyses reveal strong divergence between prediction and observation of T-cell-recognized tumor antigen T-cell epitopes. J. Biol. Chem. *292*, 11840–11849.

Schneider-Poetsch, T., Chhipi-Shrestha, J.K., and Yoshida, M. (2021). Splicing modulators: on the way from nature to clinic. J. Antibiot. (Tokyo). *74*, 603–616.

Schoch, K.M., and Miller, T.M. (2017). Antisense Oligonucleotides: Translation from Mouse Models to Human Neurodegenerative Diseases. Neuron *94*, 1056–1070.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. *26*, 732–744.

Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Buonamici, S., Yu, L., Caesar-Johnson, S.J., et al. (2018a). Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell Rep. *23*, 282-296.e4.

Seiler, M., Yoshimi, A., Darman, R., Chan, B., Keaney, G., Thomas, M., Agrawal, A.A., Caleb, B., Csibi, A., Sean, E., et al. (2018b). H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat. Med.

Seki, M., Katsumata, E., Suzuki, A., Sereewattanawoot, S., Sakamoto, Y., Mizushima-Sugano, J., Sugano, S., Kohno, T., Frith, M.C., Tsuchihara, K., et al. (2019). Evaluation and application of RNA-Seq by MinION. DNA Res. *26*, 55–65.

Sellin, M., Mack, R., Rhodes, M.C., Zhang, L., Wei, W., Breslin, P., Berg, S., Huang, S., Taylor, R., and Zhang, J. (2019). The Splicing Modulator GEX1A Exhibits Potent Anti-Leukemic Activity Both in Vitro and In Vivo through Inducing an MCL1 Splice-Switch in Pre-Clinical Models of Acute Myeloid Leukemia. Blood *134*, 2666.

Shah, M.Y., Ferrajoli, A., Sood, A.K., Lopez-Berestein, G., and Calin, G.A. (2016). microRNA Therapeutics in Cancer — An Emerging Concept. EBioMedicine *12*, 34–42.

Shen, H., Zheng, X., Luecke, S., and Green, M.R. (2010). The U2AF35-related protein Urp contacts the 3′ splice site to promote U12-type intron splicing and the second step of U2-type intron splicing. Genes Dev. *24*, 2389–2394.

Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593–E5601.

Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K., and Allen, E.M. Van (2018). Intron retention is a source of neoepitopes in cancer. Nat. Publ. Gr.

Smith, C.C., Selitsky, S.R., Chai, S., Armistead, P.M., Vincent, B.G., and Serody, J.S. (2019). Alternative tumour-specific antigens. Nat. Rev. Cancer *19*, 465–478.

Soneson, C., Love, M.I., and Robinson, M.D. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Research *4*, 1–19.

Steensma, D.P., Wermke, M., Klimek, V.M., Greenberg, P.L., Font, P., Komrokji, R.S., Yang, J., Brunner, A.M., Carraway, H.E., Ades, L., et al. (2021). Phase I First-in-Human Dose Escalation Study of the oral SF3B1 modulator H3B-8800 in myeloid neoplasms Enrolled patients.

Stone, J.D., Harris, D.T., and Kranz, D.M. (2015). TCR affinity for p/MHC formed by tumor antigens that are self-proteins: impact on efficacy and toxicity. Curr. Opin. Immunol. *33*, 16–22.

Sutherland, J., Mannoni, P., Rosa, F., Huyat, D., Turner, A.R., and Fellous, M. (1985). Induction of the expression of HLA class I antigens on K562 by interferons and sodium butyrate. Hum. Immunol. *12*, 65–73.

Tang, Y., Horikawa, I., Ajiro, M., Robles, A.I., Fujita, K., Mondal, A.M., Stauffer, J.K., Zheng, Z.M., and Harris, C.C. (2013). Downregulation of splicing factor SRSF3 induces p53β, an alternatively spliced isoform of p53 that promotes cellular senescence. Oncogene *32*, 2792–2798.

The Nobel Assembly at Karolinska Institutet (2018). Press release: The Nobel Prize in Physiology or Medicine 2018. 1–5.

Thomsen, M.C.F., 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*, 281–287.

Tian, S., Maile, R., Collins, E.J., and Frelinger, J.A. (2007). CD8 + T Cell Activation Is Governed by TCR-Peptide/MHC Affinity, Not Dissociation Rate . J. Immunol. *179*, 2952–2960.

Tran, E., Robbins, P.F., Lu, Y.-C., Prickett, T.D., Gartner, J.J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E.M., et al. (2016). T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. N. Engl. J. Med. *375*, 2255–2262.

Tran, N.H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. (2019). Deep

learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nat. Methods.

Ulgen, E., Ozisik, O., and Sezerman, O.U. (2019). PathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front. Genet. *10*, 1–33.

Vieler, M., and Sanyal, S. (2018). P53 isoforms and their implications in cancer. Cancers (Basel). *10*.

Vitting-Seerup, K., and Sandelin, A. (2018). IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. IsoformSwitchAnalyzeR Anal. Chang. Genome-Wide Patterns Altern. Splicing Its Funct. Consequences. 399642.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718.

Waldman, A.D., Fritz, J.M., and Lenardo, M.J. (2020). A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nat. Rev. Immunol. *20*, 651–668.

Wang, E., and Aifantis, I. (2020). RNA Splicing and Cancer. Trends in Cancer *6*, 631–644.

Wang, E., Lu, S.X., Pastore, A., Chen, X., Imig, J., Chun-Wei Lee, S., Hockemeyer, K., Ghebrechristos, Y.E., Yoshimi, A., Inoue, D., et al. (2019). Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. Cancer Cell *35*, 369-384.e7.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, L., Brooks, A.N., Fan, J., Wan, Y., Gambe, R., Li, S., Hergert, S., Yin, S., Freeman, S.S., Levin, J.Z., et al. (2016). Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. Cancer Cell *30*, 750–763.

Wang, X., Codreanu, S.G., Wen, B., Li, K., Chambers, M.C., Liebler, D.C., and Zhang, B. (2018). Detection of proteome diversity resulted from alternative splicing is limited by Trypsin cleavage specificity. Mol. Cell. Proteomics.

Ward, J.P., Gubin, M.M., and Schreiber, R.D. (2016). The Role of Neoantigens in Naturally Occurring and Therapeutically Induced Immune Responses to Cancer. In Adv Immunol., pp. 25–74.

Warnes, G., Bolker, B., Bonebakker, L., and Venables, B. (2020). gplots: Various R Programming Tools for Plotting Data.

Westergren Jakobsson, A., Segerman, B., Wallerman, O., Bergström Lind, S., Zhao, H., Rubin, C.-J.,

Pettersson, U., and Akusjärvi, G. (2021). The Human Adenovirus 2 Transcriptome: an Amazing Complexity of Alternatively Spliced mRNAs. J. Virol. *95*, 1–20.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wiesch, D.G., and Meyers, D.A. (2001). The epitope recognized by pan-HLA class I-reactive monoclonal antibody W6/32 and its relationship to unusual stability of the HLA-B27/β2-microglobulin complex. Immunogenetics *53*, 440–446.

Xie, Z., Tseng, Y.-T., and Xing, Y. (2015). rmats2sashimiplot.

Yarchoan, M., Johnson, B.A., Lutz, E.R., Laheru, D.A., and Jaffee, E.M. (2017a). Targeting neoantigens to augment antitumour immunity. Nat. Rev. Cancer *17*, 209–222.

Yarchoan, M., Hopkins, A., and Jaffee, E.M. (2017b). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. N. Engl. J. Med. *377*, 2500–2501.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69.

Yoshimi, A., Lin, K., Wiseman, D.H., Rahman, M.A., Pastore, A., Wang, B., Lee, S.C., Micol, J., Zhang, X.J., de Botton, S., et al. (2019). Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. Nature *574*, 273–277.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi. A J. Integr. Biol. *16*, 284–287.

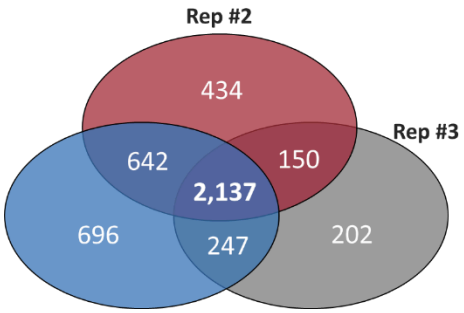
Zhang, J., and Manley, J.L. (2013). Misregulation of pre-mRNA alternative splicing in cancer. Cancer Discov. *3*, 1228–1237.

Zhang, B., Shan, H., Li, D., Li, Z., Zhu, K., Jiang, Z., and Huang, M. (2012). Different methods of detaching adherent cells significantly affect the detection of TRAIL receptors. Tumori J. *98*, 800–803.

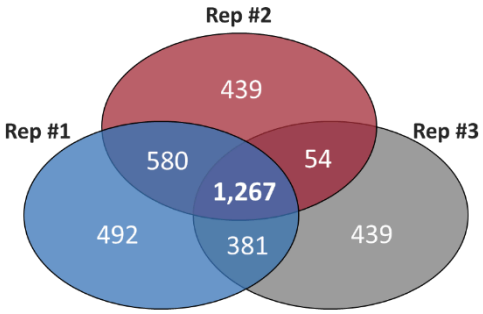
Zhou, J., Dudley, M.E., Rosenberg, S.A., and Robbins, P.F. (2005). Persistence of multiple tumor-specific T-cell clones is associated with complete tumor regression in a melanoma patient receiving adoptive cell transfer therapy. J. Immunother. *28*, 53–62.

Zolg, D.P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D.J., Gessulat, S., Ehrlich, H.C., Weininger, M., et al. (2017). Building ProteomeTools based on a complete synthetic human proteome. Nat. Methods *14*, 259–262.

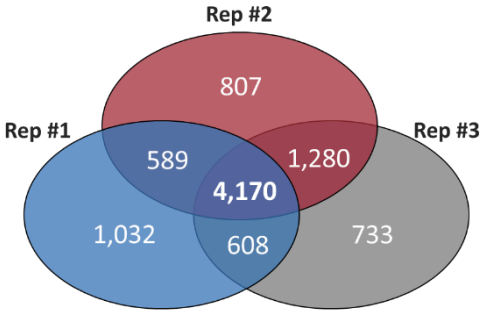
6. Appendix



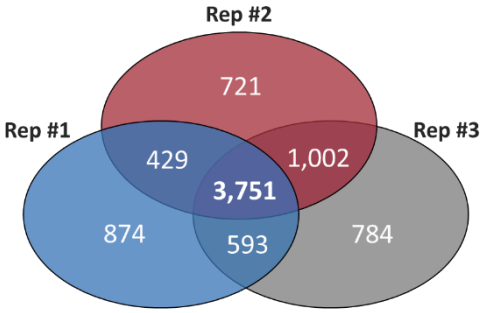
K562 NT unique peptides = 4,508



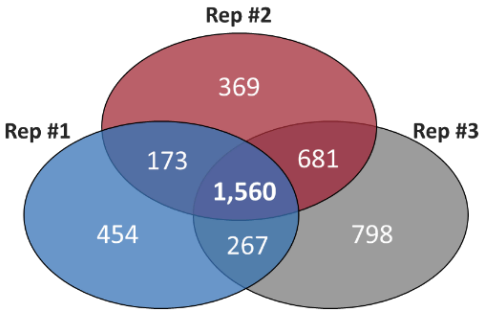
K562 GEX1A unique peptides = 3,577



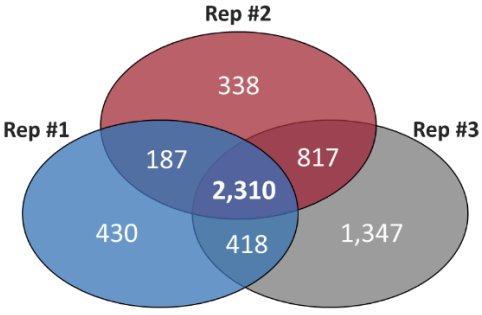
RPE-1 WT NT unique peptides = 9,219



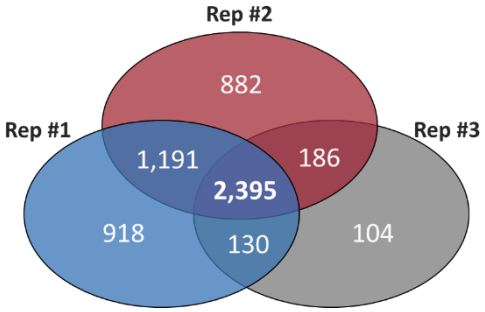
RPE-1 WT GEX1A unique peptides = 8,154



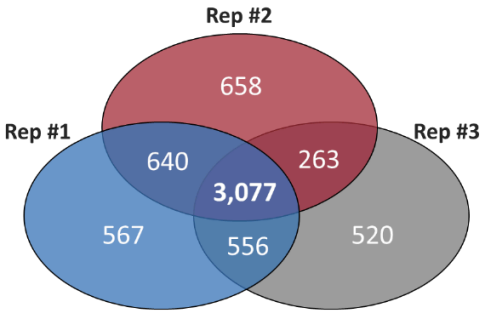
RPE-1 *TP53*$^{-/-}$ NT unique peptides = 4,302



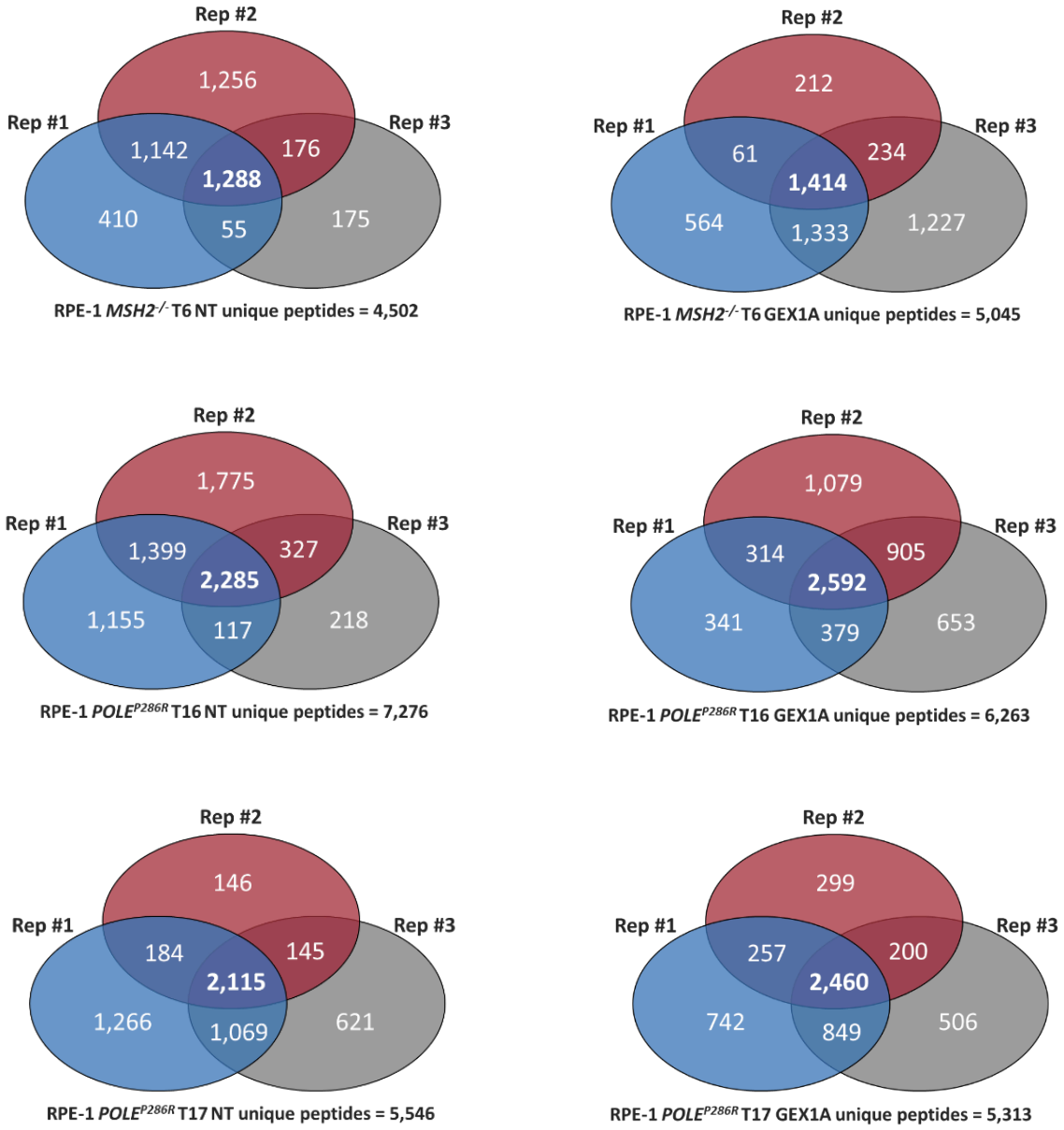
RPE-1 *TP53*$^{-/-}$ GEX1A unique peptides = 5,847



RPE-1 *MSH2*$^{-/-}$ T2 NT unique peptides = 5,806



RPE-1 *MSH2*$^{-/-}$ T2 GEX1A unique peptides = 6,281



Supplement figure 1: Replicate overlap of identified peptides recovered by immunopeptidomics for all analyzed conditions and cell lines.