Université de Montréal

Development of bio-informatic tools to unveil the hybrid MHC-I immunopeptidome

Par

Frederic Saab

Département de biochimie et de médecine moléculaire,

Faculté de médecine

Mémoire présenté en vue de l'obtention du grade de Maîtrise ès sciences (M.Sc.)

en bio-informatique

15 Juillet 2022

© Frederic Saab, 2022

Université de Montréal

Unité académique : Département de biochimie et de médecine moléculaire, Faculté de médecine

Ce mémoire intitulé

Development of bio-informatic tools to unveil the hybrid MHC-I immunopeptidome

Présenté par

Frederic Saab

A été évalué(e) par un jury composé des personnes suivantes

Daniel Sinnett
Président-rapporteur

Étienne Caron
Directeur de recherche

Marie-Claude Bourgeois-Daigneault
Membre du jury

Résumé

La chimiothérapie demeure jusqu'à présent l'approche thérapeutique la plus utilisée pour combattre le cancer. Cependant, la chiomiothérapie affecte les cellules normales et détériore ainsi de manière importante la santé des patients. En revanche, l'immunothérapie ciblée est une approche relativement innovante permettant au système immunitaire de combattre spécifiquement les cellules cancéreuses, réduisant ainsi les effets secondaires. Afin de cibler les tumeurs de manière efficace, ces approches immunothérapeutiques doivent cibler des peptides spécifiques aux tumeurs présentés à la surface des cellules par le complexe majeur d'histocompatibilité de classe I (CMH I). Ces peptides spécifiques aux tumeurs sont nommés 'néoantigènes 'et sont reconnus par les lymphocytes T pour l'élimination des cellules tumorales . Les néoantigènes 'hybrides' ont récemment démontré un potentiel thérapeutique important. Ces néoantigènes proviennent de différents fragments peptidiques ou encore de gènes de fusion, lesquels représentent de puissants oncogènes impliqués dans le développement du cancer.

Bien que ces néoantigènes hybrides soient très pertinents pour la création de stratégies immunothérapeutiques ciblées, leur découverte à grande échelle demeure un défi de taille. Jusqu'à ce jour, la spectrométrie de masse représente la méthode la plus directe pour la découverte de ces néoantigènes. Cependant, l'inaccessibilité d'outils bio-informatiques pour la prédiction et l'identification de ces néoantigènes hybrides identifiés par spectrométrie de masse est une problématique importante, ralentissant anisi leur découverte à l'échelle internationale.

Deux outils bio-informatiques facilement accessibles seront développés dans le contexte de mon projet de maîtrise: RHybridFinder (Saab et al. 2021) et FusionchoppeR (manuscript en préparation). Ces deux outils se basent sur le language informatique 'R', fréquemment utilisé en bio-informtique dans le domaine des sciences de la vie. Brièvement, ces outils intègrent l'analyse de données générées par spectrométrie de masse (PEAKS), des prédicteurs d'affinité de liaison de peptides aux molécules du CMH (NetMHCpan et MHCFlurry), ainsi que des analyses de modélisation 3D des complexes CMH- néoantigènes (RosettaMHC). Le code de ces outils est présentement ou sera déposé, dans les ressources bio-informatiques publiques CRAN et Zenodo.

Mots-clés : néoantigènes hybrides, bio-informatique, fusion de gènes, peptides épissés, RHybridFinder, FusionChoppeR, immunopeptidome hybride

Abstract

Chemotherapy is to this day the main therapeutic strategy for treating cancer. However, chemotherapy affects normal tissues as well as the overall health of patients. In contrast, targeted immunotherapies allow the immune system to selectively destroy cancer cells, thereby reducing secondary effects. To specifically and efficiently target tumour cells, immunotherapies rely on the identification of peptides presented on MHC-I (Major Histocompatibility Complex Class I) molecules. The peptides specific to tumour cells, also called "neoantigens", constitute ideal targets owing to their communication with the immune system. The recognition of these by the immune system leads to lysis of the cell. 'Hybrid' neoantigens have recently demonstrated their important therapeutic relevance. These are either derived from the ligation of two peptide fragments, or from oncogenic driver gene fusions.

Despite the relevance of these hybrid neoantigens for the development of novel targeted immunotherapeutic strategies, their discovery on a large scale remains a challenge. To date, mass spectrometry is the main direct method for the identification of these neoantigens. Nevertheless, the lack of tools for the prediction and identification of these hybrid neoantigens identified by mass spectrometry is an important issue that affects the rate of their discovery on a world-wide scale.

Two easily accessible bio-informatic tools are developed as part of my master's project: RHybridFinder (Saab et al., 2021) and FusionchoppeR (manuscript being prepared). Both tools have been developed in R language which is commonly used in the health sciences field. Briefly, these tools rely on the use of data generated by mass spectrometry (PEAKS), peptide-HLA binding predictors (netMHCpan and MHCflurry) as well as 3D structural modeling analyses of peptide-MHC-I complexes (RosettaMHC). The code for these tools is currently or will be available soon on public repositories CRAN and Zenodo.

Keywords : hybrid neoantigens, bio-informatic, gene fusions, spliced peptides, RHybridFinder, FusionChoppeR, hybrid immunopeptidome

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ALL : Acute Lymphoblastic leukemia

AML : Acute Myeloid Leukemia

ANN : Artificial Neural Network

ATP : Adenosine Triphosphate

ATPase : Adenosine Triphosphatase

BLAST : Basic Local Alignment Search Tool

BLCA : Bladder Urothelial Carcinoma

BRCA : Breast invasive carcinoma

CESC : Cervical squamous cell carcinoma and endocervical adenocarcinoma

CHU-SJ : Centre Hospitalier Universitaire Sainte Justine

COAD : Colon adenocarcinoma

COSMIC : Catalogue Of Somatic Mutations in Cancer

CTL: Cytotoxic T Lymphocyte

DLBC : Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

ERAP : Endoplasmic Reticulum Aminopeptidases

ESCA : Esophageal carcinoma

FCR : FusionChoppeR

FDA : Food and Drugs Association

FP : False Positive

HLA : Human Leukocyte Antigen

HNSC : Head and Neck squamous cell carcinoma

IFN : Interferon gamma

KIRC : Kidney renal clear cell carcinoma

KNN : K-Nearest Neighbors

LCLL : Chronic Lymphocytic Leukemia

LGG : Brain Lower Grade Glioma

LIHC : Liver hepatocellular carcinoma

LUSC : Lung squamous cell carcinoma

MHC : Major Histocompatibility Complex

MM : Multiple Myeloma Plasma cell leukemia

MS : Mass Spectrometry

NMDS : Non Metric Dimensional Scaling

OV : Ovarian serous cystadenocarcinoma

PAAD : Pancreatic adenocarcinoma

PBMC : Peripheral Blood Mononuclear Cells

PPV : Positive Predictive value

PRAD : Prostate adenocarcinoma

SARC : Sarcoma

SKCM : Skin Cutaneous Melanoma

STAD : Stomach adenocarcinoma

TAA : Tumour-Associated Antigens

TAP : Transporter associated with Antigen Processing

THCA : Thyroid carcinoma

TP : True Positive

TSA : Tumour-Specific Antigen

Remerciements

I would first and foremost like to thank my supervisor Dr. Étienne Caron, who was instrumental in helping me define and refine my ideas as well as all the stimulating science conversations. I would also like to thank Isabelle Sirois, Peter Kubiniok and my Master's thesis parrain Dr. Sebastien Lemieux for their great support and scientific advice. This work would not have been possible without their help and guidance.

I would also like to thank our collaborators Chen Li, Anthony Purcell, Pouya Faridi, John Sidney whose help was invaluable for my project.

Furthermore, I would like to thank Etienne and Isabelle for pushing me to present my project to the wider community all-the-while helping me improve my scientific communication skills. I have had the amazing opportunity of presenting at 5 conferences: 2 local (Montreal Immunology Meetings, 34e journée scientifique), 1 provincial (Congrès provincial sur la recherche mère-enfant), 1 national (Canadian National Proteomics Network) 1 international (HUman Proteome Organisation). I would also like to thank Étienne's encouragement and support for me applying for the Bourse de fin d'études en maîtrise which I was a recipient of.

I am also grateful to all the lab members of the Caronlab who have all been helpful and with whom I have enjoyed interacting.

Finally, I would also like to greatly thank my brother Patrick, as well as my family, my loved ones, my friends, especially Roger, for their support and belief in me.

Chapter 1 : Literature Review

1. MHC Class-I Antigen presentation

1.1. The Players

1.1.1. The Proteasome

The proteasome is a protein complex located in the cytoplasm and responsible for the degradation of intracellular proteins which are then presented on HLA-I alleles. Protein homeostasis (proteostatis) is maintained in large part to the proteasome. Nevertheless, proteolysis by the proteasome is essential to many other cellular processes (Fig. 1A) (Thibaudeau and Smith, 2019).

The proteasome is structurally composed of a Regulatory Particle (19S component, S being the sedimentation coefficient) and a Catalytic core Particle (20S component) (Fig. 1B) (Sahu and Glickman, 2021) The 19S component consists of 19 subunits, 13 **n**on-ATPase **R**egulatory **P**articles (RPn) and 6 **t**riple A-ATPase Regulatory Particles (RPt) which are responsible for the capture, de-ubiquitylation and unfolding of poly-ubiquitylated polypeptide substrates, following the opening of the gate. The 20S component on the other hand, is responsible for the processing of polypeptide substrates into shorter peptides with a length ranging up to 15 amino acids.

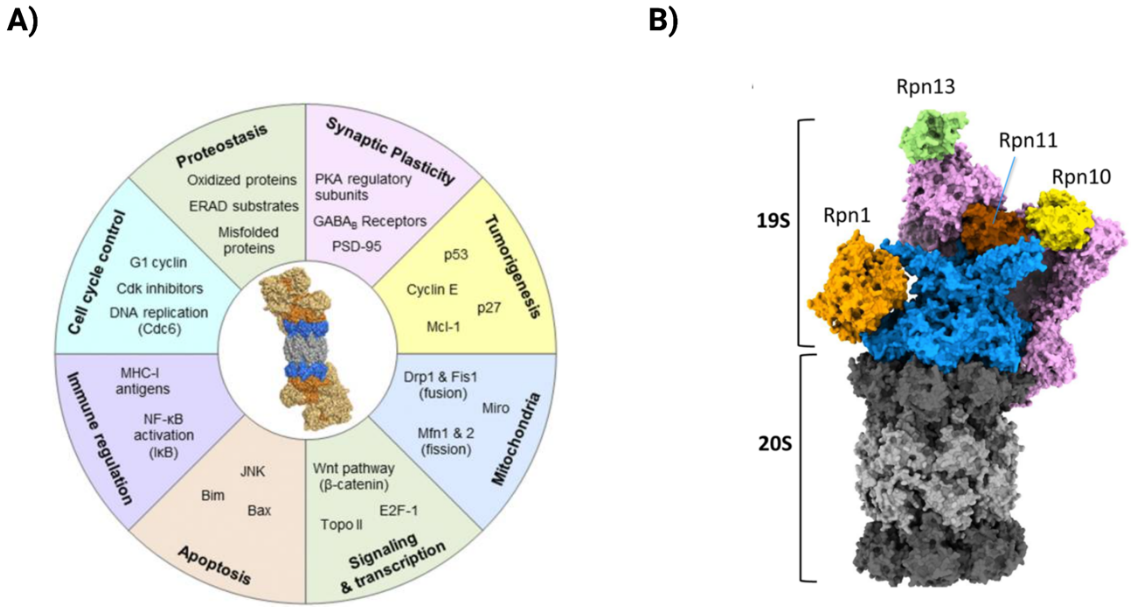


Figure 1. – The proteasome: its role in biological processes and its structure.

A) The proteasome's role is crucial for different cellular processes. Figure from (Thibaudeau and Smith, 2019). B) 3D structural representation showing the 19S and 20S components. Only 4 of the 19S subunits are annotated on the figure. The 20S component displays the α- and β-rings. The α-rings are on the top and bottom of the 20S and displayed in a darker color whereas the inner β-rings are clearer. Figure from (Sahu and Glickman, 2021).

The 20S component is formed by two outer α-rings and two inner β-rings. The α-rings act as gates. On the other hand, β-rings consist of 3 main subunits that possess catalytic activities. β1 has caspase-like function, β2 has trypsin-like function and β5 has chymotrypsin-like function (Sahu and Glickman, 2021; Tanaka, 2009)

Furthermore, two other proteasome subtypes are known. First, the immunoproteasome which is constitutively expressed in immune cells and following pro-inflammatory cytokine induction in non-immune cells and differs from the standard proteasome in its cooperative integration of three β-subunits: β1i, β2i, β5i. These have different size substrate binding pockets and therefore can accommodate different types of amino acids especially nonpolar and hydrophobic, consequently, the immunoproteasome constitutes a source of peptides suitable for the MHC-I presentation pathway (introduced in the next section). Second, the thymoproteasome

(proteasome mainly expressed in the thymus), has β1i, β2i and β5t subunits incorporated. In mice, thymoproteasome-deficient mice (proteasomes lacking the β5t) showed a reduced abundance of CD8+ T cells populations hinting at the thymoproteasome's role in generating an exclusive proportion of peptides (Eshof et al., 2021; P.-M. Kloetzel, 2001; Murata et al., 2018; Sahu and Glickman, 2021).

1.1.2. HLA-I alleles

The HLA (Human Leukocyte Antigen) class I region on chromosome 6 (6p21) is composed of genes encoding HLA-A, B & C molecules. These HLA-I molecules consist of a heavy chain (encoded by genes chromosome 6) that has three α domains and a β2-microglobulin domain (encoded by chromosome 15). HLA-I molecules are expressed on the surface of all nucleated cells. An individual can express up to 6 different HLA-I molecules, 2 from each gene: A, B, C (Fig. 2). HLA-I molecules are best known for their association with transplants whereby compatibility between the donor and the recipients' HLAs is important for the success of a transplant (Choo, 2007).

The history of their discovery dates to the 1950s (Juji, 1988; Park and Terasaki, 2000). Their nomenclature as well as their typing was established through a series of workshops and conferences (Park and Terasaki, 2000). The HLA genes are highly polymorphic and to date, 24308 HLA class I (HLA-I) alleles have been identified according to the IPD-IMGT/HLA database compared with 266 identified in 1996, and 9182 HLA class II (HLA-II) alleles (Parham and Ohta, 1996; Robinson et al., 2019).

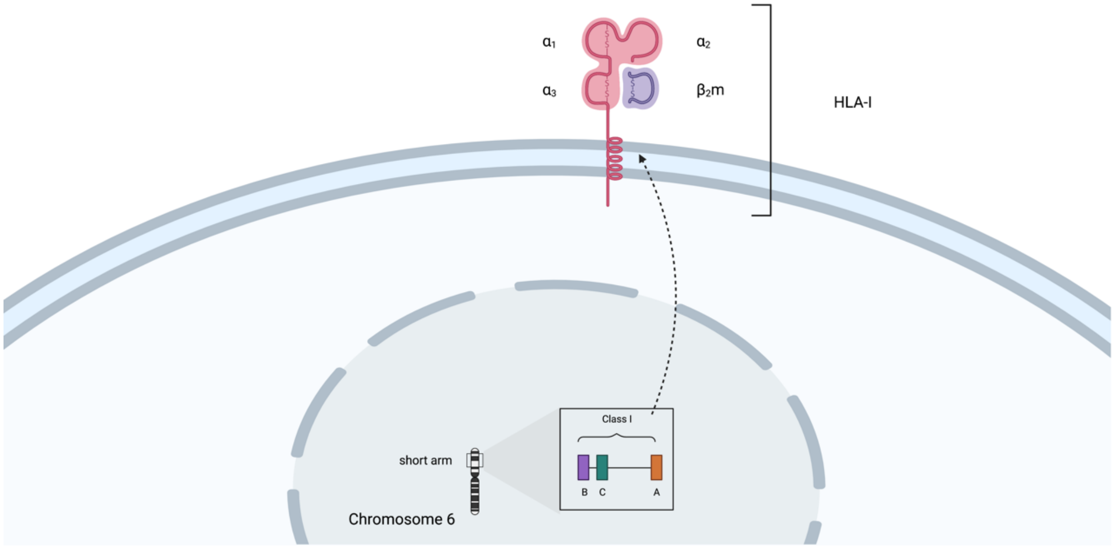


Figure 2. – The genes expressing HLA-I alleles.

The genes that code for the HLA-I alleles are located on the short arm of chromosome 6. 3 HLA-I genes code for 3 types of molecules presented on the surface of the cell: A, B & C. The HLA-I molecule is composed of one heavy chain which has three domains: $\alpha1$, $\alpha2$, $\alpha3$.

The HLA-I molecules are commonly represented by the gene coding the HLA allele (here, we focus on the HLA- A, B & C genes) followed by an asterisk, and then two digits corresponding to the allele family, then a colon and lastly, two digits which correspond to the order at which this specific allele was identified.

In this manuscript, I will refer to these molecules as HLA-I and MHC-I interchangeably. However, please note that MHC-I (Major Histocompatibility Class I) designates these molecules in general terms across species whereas HLA-I (Human Leukocyte Antigen class I) refers specifically to those in humans.

1.2. HLA diversity

1.2.1. Origins

The high polymorphic nature of HLA-I alleles allows them to have different peptide binding specificities. Populations around the world carry different sets of HLA-I alleles. In evolutionary terms, the HLA-I diversity was explained through a selection process that either (1) favors a strong immune response through the heterozygosity of HLA-I allele in order to enhance presentation and therefore protection of an individual, or (2) is the result of selection based on pathogens present within certain geographical areas and time periods or (3) through a combination of gene conversion, linkage disequilibrium and random shift (Parham and Ohta, 1996). Interestingly, statistical analyses of population peptide-binding coverage of HLA-I molecules showed that both the geography/time & the divergent allele advantage (favoring a strong immune response through allele diversity) along with gene conversion explained the diversity of HLA-I alleles (Buhler et al., 2016; Pierini and Lenz, 2018). Moreover, because allele homozygosity was not found to necessarily pose a disadvantage, optimal protection is thought to be associated with a complementarity of HLA-I alleles carried (Kaufman, 2018; Parham and Ohta, 1996). This balance is described between generalist and fastidious alleles. The former have a tendency to bind a large variety of peptides that confer general protection against a wide array of pathogens whereas the latter bind a narrower set of peptides but are more specific in their protection against certain pathogens (Kaufman, 2018). Recently, the heterozygosity of HLA-I alleles has also been linked with outcome of cancer immunotherapy with Immune checkpoint inhibitors (Chowell et al., 2018, 2019). In contrast, Manczinger et al. (2021) found no correlation between allele heterozygosity and outcome but rather suggested an effect by the size of the HLA-I repertoire, which, if large enough could lead to tolerance of tumors (Manczinger et al., 2021). To sum up, balance is key for HLA-I alleles. Whether their diversity could be a blessing or a curse in disguise, they are important in order to confer protection.

1.2.2. Effect on peptide binding

The diversity of HLA-I alleles impacts the type of peptides that they bind to. The conformation of the $\alpha1$ and $\alpha2$ domains of the HLA-I molecules creates a peptide binding cleft composed of six pockets that engage amino acids within a peptide molecule presented on it (Nguyen et al., 2021). These different pockets are named “A” to “F” and engage amino acids at different positions within a peptide (Fig. 3 and Table 1). Notably, HLA-I alleles have binding specificities that differ between them especially on the level of the B and F pockets. These bind amino acids at the 2nd and 9th positions of the peptide, respectively. However, this might be different for some HLA-I alleles such as HLA-B*08:01 which has binding particularities for peptide residue in fifth position binding its C pocket (Barber et al., 1995; Nguyen et al., 2021). Based on peptide binding specificities of 945 HLA-A & B alleles (with 4 digits (HLA-(A/B)*xx:xx)), 80% of these were clustered into 12 supertypes and the rest was considered as unclassified (I. A. Doytchinova et al., 2004; I. Doytchinova and Flower, 2003; Hertz and Yanover, 2007; A. Sette and Sidney, 1999; Sidney et al., 2008). In summary, the requirement of each HLA-I allele for amino acids in their binding pockets creates different sets of HLA-peptide repertoires that that could or not overlap. This in turn affects the peptides being presented on the surface of the cell to the immune system.

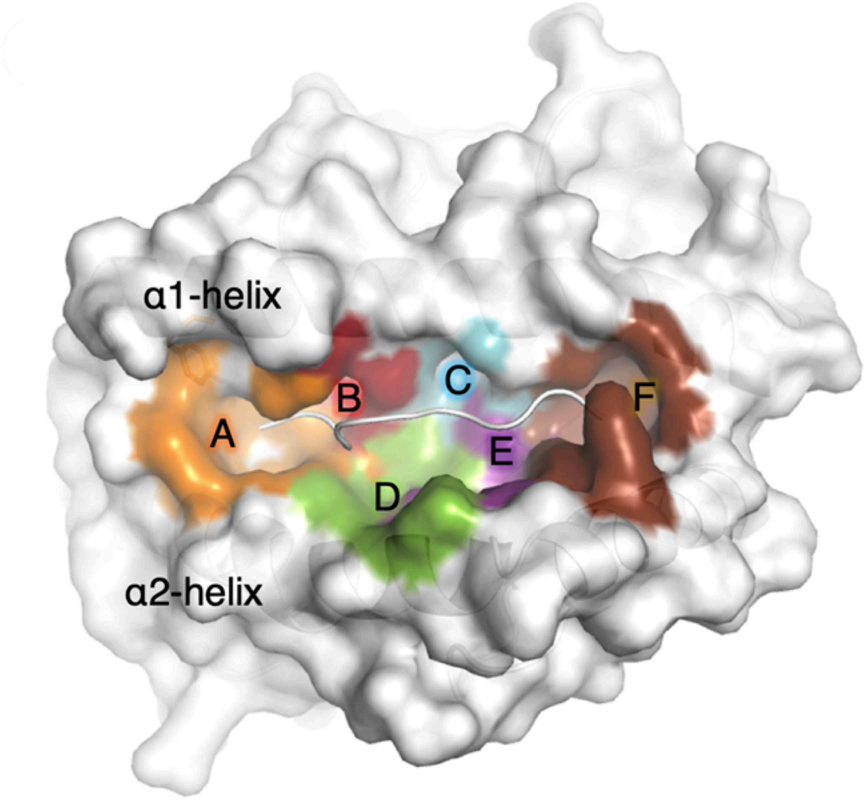


Figure 3. – 3D structure representation of HLA-I peptide binding pockets.

Residues within the α1 and α2 domains of the HLA-I molecule form peptide binding pockets which could have their own speicificities when it comes to the amino acids at different positions within a peptide. Figure from (Nguyen et al., 2021)

Pocket	Residues	Role of the pocket
A	5, 7, 59, 63, 66, 159, 163, 167, 171	Wall of the N-terminal part of the binding cleft, bind P1 residue
B	7, 9, 24, 34, 45, 63, 66, 67, 70, 99	Bind primary anchor residue P2
C	9, 70, 73, 74, 97	Bind secondary anchor residue at P3 and P5/P6 when presents, face pocket D
D	99, 114, 155, 156, 159, 160	Bind secondary anchor residue at P3 and P5/P6 when presents, face pocket C
E	97, 114, 147, 152, 156	Overlap with C/D pockets and contact secondary anchor residue at P5/P6 when presents and the C-terminal part of the peptide
F	77, 80, 81, 84, 95, 123, 143, 146, 147	Bind primary anchor residue PΩ, wall of the C-terminal part of the binding cleft

Tableau 1. – HLA-I structural peptide binding pockets. Adapted from (Nguyen et al., 2021).

1.3. The MHC-I presentation pathway

The roots of MHC-I and its involvement with the immune system could be traced back to transplant studies in the 1950s. The findings by Billingham, Brent & Medawar (1953) that the immune system of fetal mice is taught to 'tolerate' certain antigens presented by the MHC-I was phenomenal and has shaped much for our understanding of adaptive immunity (BILLINGHAM et al., 1953).

Moreover, since then, research has intensified to characterize these MHC-I alleles as well as the peptides they present: their origin, their specificities, the mechanism by which these are created. It is well-known now that the MHC-I presentation pathway is a way for the cell to channel samples of its interior state to the outside of the cell (Caron et al., 2011). In healthy cells, peptides on MHC-I are derived from normal proteins expressed in the cell, whereas in a pathogenic state such as in the case of intracellular pathogens or tumour the peptides presented on the MHC-I molecule are influenced by the state of the cell. The MHC Class I presentation pathway consists of the presentation of short peptides (8-12 amino acids) on MHC-I molecules at the surface of the cell. After protein digestion by the proteasome, peptides are then degraded further by aminopeptidases in order to 'recycle' the amino acids. However, few peptides are transported into the endoplasmic reticulum with the help of the Transporter associated with Antigen Processing (TAP) and loaded onto MHC-I molecule. Then, the peptide on the MHC-I molecule might undergo further trimming by ERAP (Endoplasmic Reticulum Aminopeptidases) before the complex is transported into the Golgi Apparatus and then presented on the surface of the cell to the immune system (CD8+ T cells) (Fig. 4 – not all these steps are shown in this figure)(Carluccio et al., 2018) . In other words, the immune system is rummaging through the cell's 'trash' that is displayed on the surface of the cell and it destroys what it recognizes as foreign (Hewitt, 2003). Rock et al. (2016) have also whimsically likened the process to a presented meal by 'waiters' (MHC-I molecules) on the surface of the cell which is judged by a patrol officer (representing Cytotoxic T Lymphocytes (CTL)) (Rock et al., 2016).

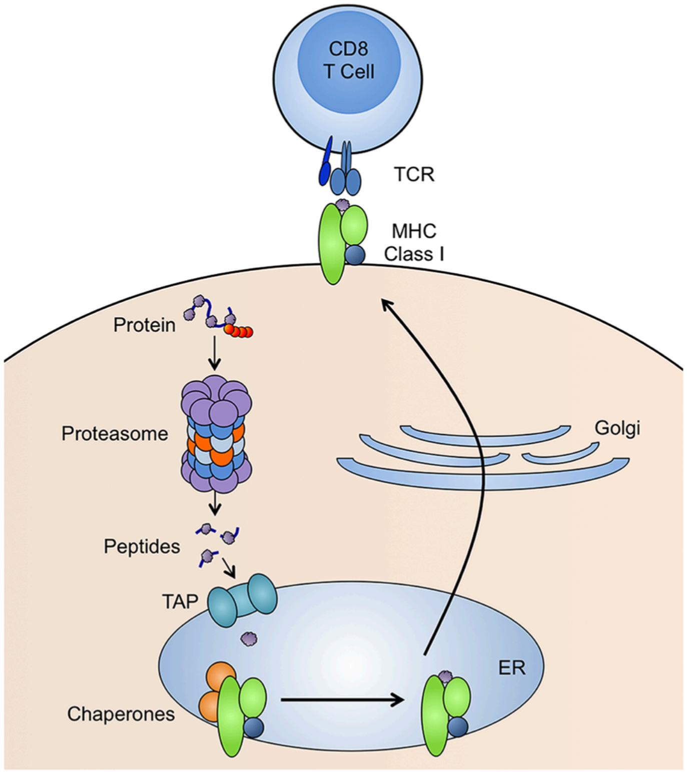


Figure 4. – MHC Class I Antigen Presentation Pathway.

After degradation of a protein by the proteasome into peptides, few of those are transported into the Endoplasmic reticulum through TAP (Transporter Associated with Antigen Processing) into the Endoplasmic reticulum whereby the binding of a peptide to a fully folded MHC-I molecule determines its presentation will lead to CD8+ T cells. Figure adapted from (McCarthy and Weinberg, 2015).

This pathway is essential for eliminating compromised cells. Therefore, the survival of such cells relies upon the disruption of the MHC-I antigen presentation pathway in order to hide from the immune system. Mechanisms involved in immune escape consist of direct alterations of MHC loci (Loss Of Heterozygosity, gene deletion), epigenetic dysregulation of genes associated with the antigen presentation pathway, indirectly - through the manipulation of signaling and transcription of these genes. Therefore, disrupting any of the steps of the presentation could be key for preventing cell recognition and lysis by lymphocytes which confirms the importance of this pathway for targeting tumor cells (Dersh et al., 2020; Pyke et al., 2022).

1.4. Immunopeptidome

"Immunopeptidome", or the collection of immune peptides was first coined by Johnathan Yewdell in 2004 and now constitutes an entire investigative field of the peptides bound to MHC-I (Istrail et al., 2004).

1.4.1. MHC-I peptides, neo-antigens for immunotherapy

Tumours have their own immunopeptidomic signatures and the hunt for tumor targets has classified tumor targets as Tumour-Associated Antigens (TAA) and Tumour-Specific Antigens (TSA). The former are antigens associated with aberrantly expressed normal proteins, these can also be expressed by normal tissues. On the other hand, TSAs are derived from "modified" proteins and therefore are specific to tumour cells - these are often also called "neoantigens". Because the immunopeptidome can act as a 'whistleblower' of neoantigens, targeting these in order to harness a tumour-specific immune response is the rationale for the development of various immunotherapies. Moreover, to develop these, target (neoantigen) discovery is key and relies on factors such as 1) its expression only in tumors, 2) MHC binding, this can be achieved in-silico in a first step using prediction algorithms. It should nevertheless be validated with peptide-HLA in-vitro binding affinity measurement and MHC-peptide elution using immunopeptidomic protocols. Finally, 3) the ability of the neoantigen to bind multiple HLA alleles. This confers a great advantage in order to circumvent HLA loss of heterozygosity. Also, considering the diversity of HLA-I alleles across populations, this helps in the application of cross-HLA-I allele immunotherapies (McGranahan and Swanton, 2019; Yarmarkovich et al., 2021).

To date, clinical trials involving neoantigens are ongoing and have already shown great promise in treating melanoma patients (Carreno et al., 2015; Ott et al., 2017; Sahin et al., 2017). Carreno et al. (2015) conducted a small phase 1 clinical trial on three patients with stage III resected cutaneous melanoma who had also received immune checkpoint inhibitor therapy. The researchers vaccinated all three patients using mature autologous dendritic cells pulsed with synthetic HLA-A*02:01-peptides derived from non-synonymous mutations, personalized for each patient. In their study, they showed that vaccination highly increased CD8+ neoantigen-specific T cell responses in PBMCs extracted pre- and post- vaccination and all patients survived (Carreno et al., 2015) In 2017, Sahin et al. used RNA-based vaccines containing 10 selected mutations per patients encoding 27-mer peptides (the peptides contained 13 amino acids from each side of the mutated residue) on 13 stage III and IV melanoma patients. While one patient died due to rapid disease progression, a second patient died because of rapid disease progression and a loss of β2-microglobulin allele, without which an HLA-I molecule is not complete and cannot therefore be transported onto the surface. One patient's vaccination was stopped (due to multiple relapses) and received immune checkpoint inhibitor therapy, and eventually experienced a complete response. Nevertheless, 9/13 patients were recurrence free after vaccination and exhibited strong immune response. Ott et al. (2017) vaccinated 6 patients in stage III and IV melanoma with pools of synthetic long peptides spanning 20 neoantigens per patient. Two patients with untreated lung metastases had disease recurrence after vaccination and then received checkpoint inhibitor therapy. They then exhibited a clinical response and remained recurrence-free for the remainder the trial phase period. The rest of the patients had no recurrence after 25 months post-vaccination(Ott et al., 2017). Together, the results of all three of these personalized vaccination trials on melanoma patients have demonstrated the safety of neoantigen vaccines and the efficacy of these whether alone or with checkpoint inhibitor therapy.

1.4.2. HLA polymorphisms & the immunopeptidome

The binding specificities of HLA-I molecules are often represented by logo plots showing the information content or the importance of amino acids at each of the positions of the peptide (Fig. 5). These specificities or "binding motifs" have been established for various alleles since 1992 (Bassani-Sternberg et al., 2017; Chelvanayagam, 1996; Kubo et al., 1994; Matsumura et al., 1992; Rammensee et al., 1995). Furthermore, mutations of immunogenic peptides at specific residues leading to immune escape through loss of binding have also contributed to validating these motifs (Cardinaud et al., 2011; Draenert et al., 2004; Murakoshi et al., 2018).

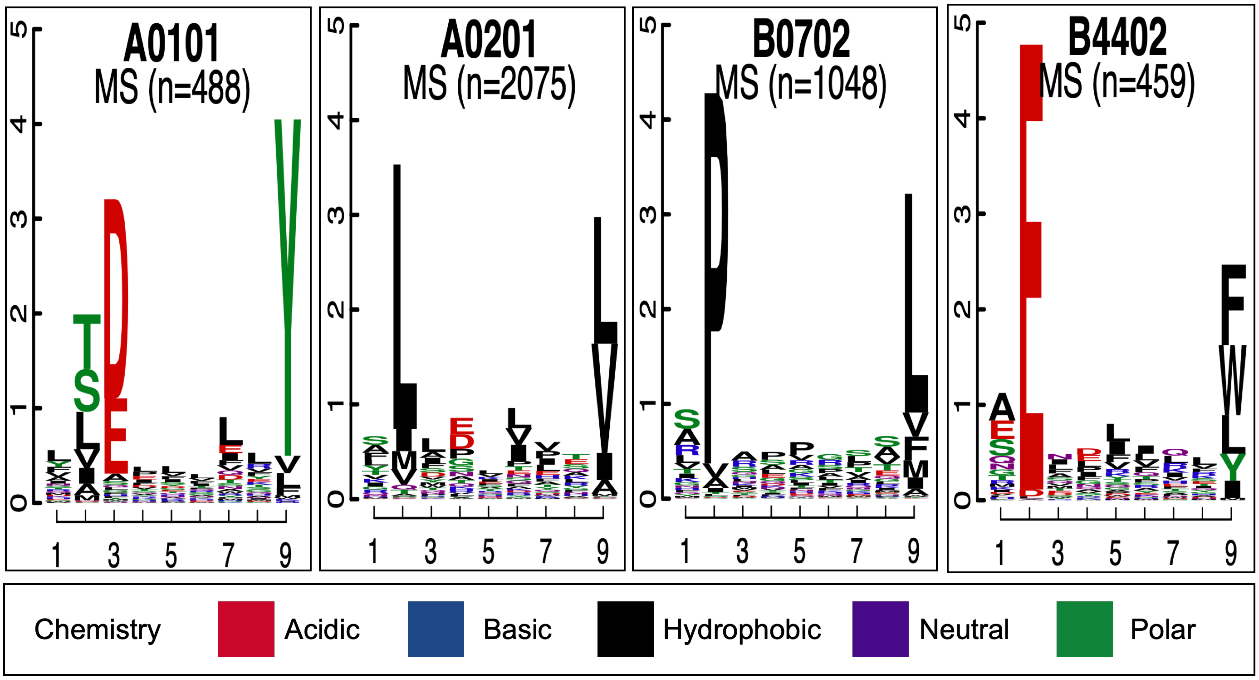


Figure 5. – Example of logo plots.

*Four logo plots for 9-mers for four HLA-I alleles: HLA-A*01:01, HLA-A*02:01, HLA-B*07:02 and HLA-B*44:02. Each of these plots represents the collection of peptides associated with these HLA-I molecules. The x-axis represents the positions within a peptide, the y-axis represents the bits associated with the different amino acids at different positions within a peptide sequence. The height of amino acids is representative of its importance at that position. The amino acids are color-coded based on their properties as shown in the legend below the plots. Figures adapted from (Sarkizova et al., 2020).*

Moreover, in recent years, MS sequencing of HLA-I ligands eluted from mono-allelic cell lines has further improved the characterization of the specificities of various HLA-I alleles (Abelin et al., 2017; Trolle et al., 2016). Notably, Sarkizova et al. (2020) have led a large-scale sequencing of the eluted ligands of 95 mono-allelic cell lines. This has allowed to elucidate the binding motifs of many alleles that were underrepresented (Sarkizova et al., 2020a).

1.4.3. Cancer Neoantigens: Personal vs. Public

Neoantigens can be classified into two categories based on the whether these are specific to a patient or not. Personal neoantigens have shown to be great targets for immunotherapies due to their specific to the individual. However, because of their specificity to each individual, personal neoantigens are expensive. Alternatively, for certain tumors and age ranges the protein driver sequence could be the same across individuals, and therefore neoantigens derived from these can be shared, these are considered "public" neoantigens. Neoantigens that are shared between individuals constitute ideal targets as these can pave the way for an "off-the-shelf" type of cancer immunotherapy (Biernacki and Bleakley, 2020; Chandran et al., 2022; Chen et al., 2020; Pearlman et al., 2021). The most documented sources of shared neoantigens include recurrent driver mutations of oncogenes as well as gene fusions (Biernacki and Bleakley, 2020; Chandran et al., 2022; Pearlman et al., 2021). Other modifications that show potential to be sources include post-translationally modified peptides, such as proteasome spliced peptides which have been shown to trigger immune responses in multiple patients (Ebstein et al., 2016; Faridi et al., 2020). Therefore, evidently, the identification of shared neoantigens is of great importance.

1.5. Hybrid immunopeptidome (HIP)

Since immunopeptidome designates the collection of peptides bound to MHC-I alleles, the "hybrid immunopeptidome" (a term coined in this thesis) is meant to represent the hybrid peptides bound to MHC-I molecules (Fig. 6). Hybrid refers to the peptide as being formed either from the same parent protein or from two parent proteins. The hybrid peptides I will focus on in my thesis are gene fusion-derived peptides and peptides derived from proteasome-catalyzed peptide splicing.

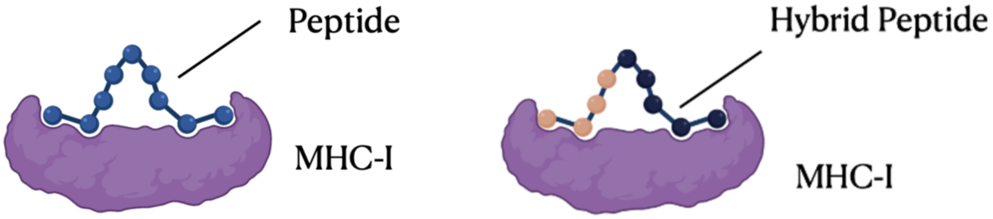


Figure 6. – Canonical vs. hybrid immunopeptidome.

(On the left) the canonical immunopeptidome represents the collection of peptides that bind MHC-I molecules and are derived from a protein. (On the right) the hybrid immunopeptidome on the other hand is the collection of peptides that bind MHC-I molecules but that are derived from two regions within this protein or from two different proteins.

1.6. Gene fusions

1.6.1. Gene fusions in pediatric cancers

Cancer remains the primary cause of death of pediatric patients and despite the improvements of survival rates, traditional therapies have been documented to cause various adverse long-term effects for these patients (Hamilton et al., 2017; Mulhern and Butler, 2004). Therefore, the development of therapies that specifically target the tumours is needed for pediatric patients.

Pediatric cancers are characterized by a low mutational burden (low number of non-inherited mutations per million bases) and an embryonic origin (Vogelstein et al., 2013). Chromosomal rearrangements such as gene fusions are frequently observed in pediatric cancers and they are hypothesized to be the drivers of these (Dupain et al., 2017; Loupe et al., 2017).

Moreover, gene fusion events are chromosomal structural rearrangements that involve through various mechanisms (translocations, inversion, deletion, etc.) the fusion of genes. A gene fusion is formed by the juxtaposition of two genes, a 5' and a 3' partner genes (also called head and tail genes), often represented by the name of the genes separated by a special character, i.e., ETV6 - RUNX1. These can either derive from the same chromosome or from different chromosomes. Generally, abnormalities in tumours are selected for their role in ensuring the survival of the cells, therefore, gene fusions should be non-random events especially these considered as being driver fusions (Latysheva and Babu, 2016; Mitelman et al., 2007; Roukos and Misteli, 2014). Gene fusion events can dysregulate protein expression, incur loss of function or generate a chimaeric protein

(Fig. 7A). Consequently, to understand the function of a chimaeric protein, it is important to know the functions of the parent genes as well as which protein domains are retained in the novel chimaeric protein (Fig. 7B) and lastly determine whether a frameshift is incurred or not. That is, whether the reading frame is preserved in the second protein (in-frame) or there has been a frame shift (also termed "out-of-frame") that alters the reading frame and the sequence of the resulting protein (Fig. 7C).

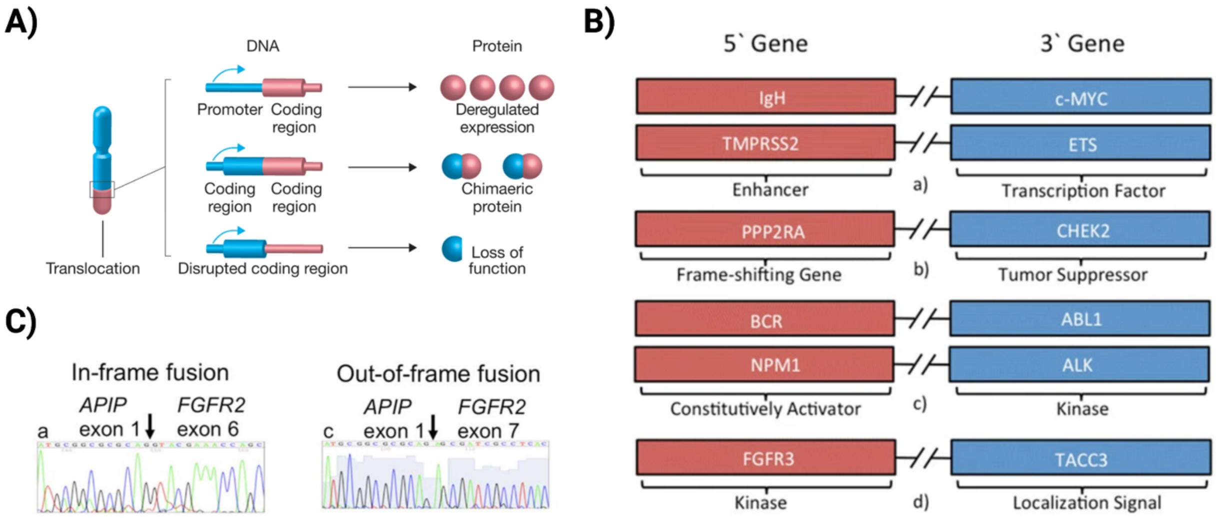


Figure 7. – General characteristics of gene fusion events.

A) Figure adapted from (Roukos and Misteli, 2014) showing the different effects that a translocation can have. Either activity of the juxtaposing coding region deregulated due to the fusion of the promoter, or a loss of function is incurred through a fusion that disrupts the coding region of the juxtaposing coding region. A last potential effect is through the creation of a chimaeric protein composed protein. Examples of the two deregulatory effects of gene fusions are depicted in Panel B) subpanels a) and b) whereas the subpanels c) and d) provide examples of chimaeric proteins. Figure adapted from (Abate et al., 2014). Panel C) depicts the possible frame-reading shifts caused by these translocations. Based on whether the fusion changes the reading frame of gene 2, the resulting protein can be in-frame (no shift) or out-of-frame (frameshifted). Figure from (Okuda et al., 2017).

Owing to the breakpoints having a tendency of being localized at certain genomic regions which are transcriptionally active, or that enhance the susceptibility of breakage (characterized by repeats, fragile sites and endonuclease misrecognition sites) the translocations are recurrent (Latysheva and Babu, 2016; Roukos and Misteli, 2014). The same rearrangement in several pediatric patients has been described in the literature (Biernacki and Bleakley, 2020; Chang et al., 2019a; LaHaye et al., 2021; Reshmi et al., 2017). Two pediatric gene fusions which represent ~20-

25% (ETV6-RUNX1) of pediatric ALL (Acute Lymphoblastic Leukemia) and ~12% (CBFB-MYH11) of pediatric AML (Acute myeloid Leukemia) were studied in the context of MHC-I immunogenicity. Both were shown to generate specific T cell responses against neoantigens derived from these gene fusions (Biernacki et al., 2020a; Zamora et al., 2019a). This begs the question, could certain gene fusion-derived neoantigens be shared across cohorts?

These two gene fusions are associated with a positive clinical outcome however, pediatric patients would nevertheless benefit from less harmful and more targeted treatment approaches. Also, there remains a large number of pediatric gene fusions that (1) do not involve kinase genes, (2) have an unfavorable prognosis and (3) are known to be exclusive to pediatric patients, such as those involving NUP98, CBFA2T3, KMT2A (Bolouri et al., 2018). Taken together, given the potential for therapeutic treatment of Gene Fusion-derived Neoantigens (GF-Neo) (Biernacki et al., 2020a; Yotnda, Garcia, et al., 1998; Zamora et al., 2019a), and the recurrence of gene fusions in pediatric cancers, exploring the landscape of GF-Neo of gene fusions in pediatric cancers is warranted.

1.6.2. Beyond the genomic level, towards an immunopeptidomic advantage

The degradation of intracellular proteins by the proteasome results in peptides that would bind the MHC-I alleles on the surface of the cell. A proof-of-concept study by Yang et al. (2019) on recovered cancer patients revealed that these patients harbored T-cells specifically reactive to a gene fusion peptide (Yang et al., 2019). Therefore, peptides from these chimeric proteins are key to the development of immunotherapies.

1.6.2.1. The junction region is important

The relevance of the fusion junction region as source of neoantigens was established since 1998 (Yotnda, Garcia, et al., 1998). The researchers sought to evaluate the immunogenic potential of one peptide derived from the junction region of the ETV6-RUNX1 fusion protein, which is the result of the t(12;21) translocation. This translocation is present in over 25% of childhood ALL. The peptide spanning the junction region RIAECILGM (Fig. 8A) tested in their study was found to trigger an immune CTL response in an HLA-A02 restricted manner. This has also been observed in a study 21 years later with patient-derived T-cells (Zamora et al., 2019a). Furthermore, a fusion junction neoantigen from CBFB-MYH11 has also been validated for its immunogenicity (Biernacki et al., 2020a) (Fig. 8B).

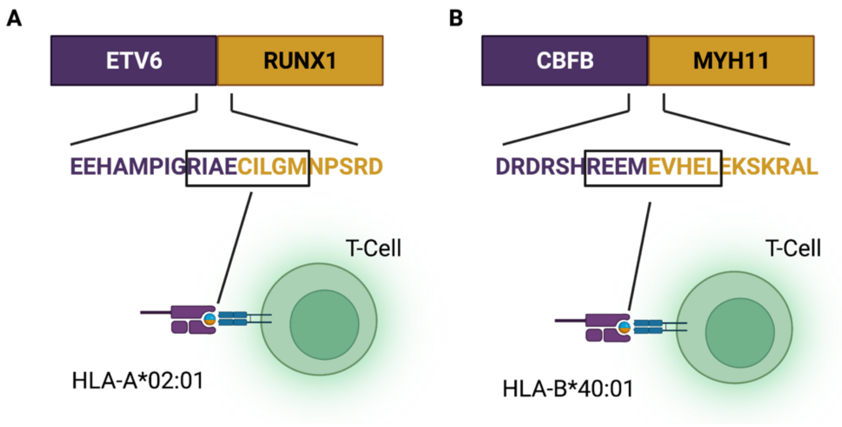
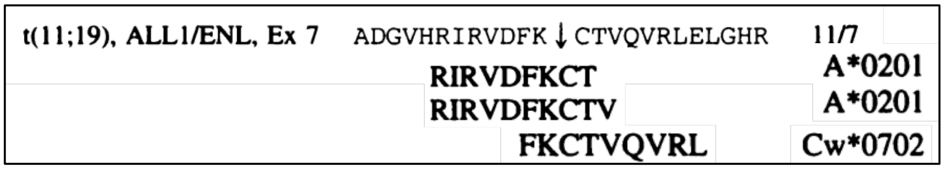


Figure 8. – The junction region is a source of neoantigens

The junction peptides from ETV6-RUNX1 and CBFB-MYH11 displayed above that have been shown in the literature to be immunogenic highlight the relevance of the junction in acting as a neoantigen source.

1.6.2.2. Gene fusion proteins are a source of neoantigens

In 1997, Gambacorti-Passerni et al. predicted gene fusion peptides via mapping HLA-I binding motifs to fusion proteins in order to extract those that fit certain motifs (Gambacorti-Passerini et al., 1997). They mapped junction peptides of 44 gene fusions to HLA-I alleles. Then, they tested the binding of their candidates to these in vitro through a stabilization assay and found that 13-40% were binders. However, they only considered 3 alleles: HLA-I alleles HLA-A*02:01, HLA-A*03:01 and HLA-C*07:02. One interesting finding was that some fusion peptides derived from rare gene fusions fit the binding motif of highly frequent HLA-I alleles. Also, they found that fusions involving the same genes but where the rearrangement could be different could generate peptides that could bind other alleles. Lastly, they observed that a fusion junction polypeptide, could generate more than one HLA-I-peptides (Fig. 9). Particularly, these were mapped to different HLA-I alleles. They have called such gene fusions as being heterogenous. One such example in their study was ALL1/ENL (Fig. 9), two peptides were predicted (based on motif mapping) to bind HLA-A*02:01 and one peptide to HLA-C*07:02. Furthermore, according to their in vitro results these were also deemed as HLA-I binders. To sum up, gene fusions can be a source of neoantigens, HLA-I-peptides and gene fusions and these can also generate peptides that bind to different HLA-I alleles.



t(11;19), ALL1/ENL, Ex 7	ADGVHRIRVDFK ↓ CTVQVRLELGHR	11/7
	RIRVDFKCT	A*0201
	RIRVDFKCTV	A*0201
	FKCTVQVRL	Cw*0702

Figure 9. – “Heterogeneous” gene fusion.

The figure clearly displays what the authors named as being a “heterogeneous gene fusion”, that is, a fusion protein that could generate peptides that are predicted to bind different sets of alleles. In the example is shown the fusion protein sequence of ALL1/ENL. Below, are shown three peptide sequences, 2 of which are predicted to bind the same HLA-I allele, and then second one, aa completely different HLA-I allele. The figure was created from screenshots of tables from (Gambacorti-Passerini et al., 1997).

1.6.2.3. Gene fusion peptides in clinical treatments

In 2005, Bocchia et al. aimed at improving the persistent residual disease (that is the persistence of cancer cells associated with the disease) seen in BCR-ABL1 patients treated with tyrosine kinase inhibitors (TKIs) or interferon alfa. TKIs are the main drugs used to treatment patients with gene fusions involving a kinase. To specifically target cells that express this fusion, they vaccinated 16 patients who had undergone either of these treatments and have not exhibited full removal of the disease. The vaccine included peptides of different lengths that would bind HLA-I and HLA-II molecules. Focusing on previously TKI-treated patients and that had stable residual diseases, while all 9 patients showed progressive improvements, 5 displayed complete cytogenetic remission. Interestingly, 4 of these 5 patients had more than one of HLA I or II appropriate for the presentation of the peptides used in the vaccine (M Bocchia et al., 2005). This same peptide vaccine was administered to a patient with BCR-ABL1 fusion whose interferon alpha treatment had been discontinued after 6 years of continuous treatment and was diagnosed with chronic phase Philadelphia positive chronic myeloid leukemia (Philadelphia/Philadelphia chromosome refers to the BCR-ABL1 gene fusion). The patient started responding to the vaccine after vaccine boosting. Assessment of vaccine response was done by 1) measuring peptide-specific CD4+ T cells in vitro isolated from patient PBMCs (pre- and during vaccination) and 2) by measuring level of the fusion transcript in blood or marrow of the patient. After 1 year after the start of vaccination, the patient had no residual leukemic Philadelphia positive cells and an undetectable level of BCR-

ABL1 fusion transcript (Monica Bocchia et al., 2010). Taken together, this has proven the efficacy of neoantigen-based vaccines to specifically target cells expressing the gene fusion and eliminate residual disease to prevent potential relapse.

1.7. Post-translationally spliced peptides

In 2004, Hanada and colleagues found through extensive experimental validation that the proteasome is not only lysing proteins but also ligating their peptide fragments. Throughout their experiments they were able to demonstrate that an immunogenic peptide was indeed generated by the proteasome through the ligation of two fragments from the same FGF-5 protein but separated by ~40 amino acids in the FGF-5 protein (Hanada et al., 2004). By 2016, six proteasomally-spliced peptides had already been described in the literature. (Dalet et al., 2011; Ebstein et al., 2016; Hanada et al., 2004; Michaux et al., 2014; Vigneron et al., 2004; Warren et al., 2006). These included the finding that peptides can also be spliced in a 'reverse-cis-spliced' manner. In other words, the fragments spliced are in a different order than in the parent protein (Fig. 10C). Lastly, while highly debated, it was also posited that the proteasome could splice fragments from different parent proteins (Fig. 10D) (Faridi et al., 2018; Liepe et al., 2016; Specht et al., 2020).

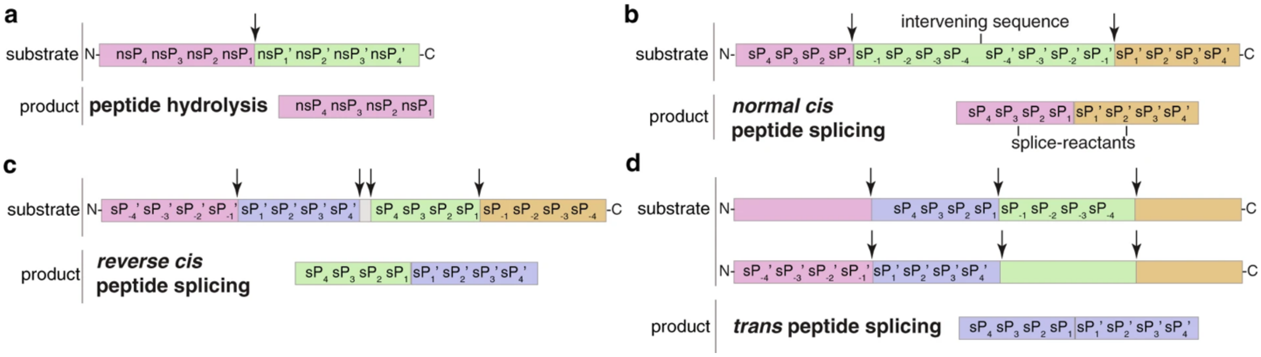

a
substrate
N- nsP_4 nsP_3 nsP_2 nsP_1 nsP_1' nsP_2' nsP_3' nsP_4' -C
product
peptide hydrolysis
nsP_4 nsP_3 nsP_2 nsP_1
b
intervening sequence
substrate
N- sP_4 sP_3 sP_2 sP_1 sP_{-1} sP_{-2} sP_{-3} sP_{-4} sP_{-4}' sP_{-3}' sP_{-2}' sP_{-1}' sP_1' sP_2' sP_3' sP_4' -C
product
normal cis peptide splicing
sP_4 sP_3 sP_2 sP_1 sP_1' sP_2' sP_3' sP_4'
splice-reactants
c
substrate
N- sP_{-4}' sP_{-3}' sP_{-2}' sP_{-1}' sP_1' sP_2' sP_3' sP_4' sP_4 sP_3 sP_2 sP_1 sP_{-1} sP_{-2} sP_{-3} sP_{-4} -C
product
reverse cis peptide splicing
sP_4 sP_3 sP_2 sP_1 sP_1' sP_2' sP_3' sP_4'
d
substrate
N- sP_4 sP_3 sP_2 sP_1 sP_{-1} sP_{-2} sP_{-3} sP_{-4} -C
N- sP_{-4}' sP_{-3}' sP_{-2}' sP_{-1}' sP_1' sP_2' sP_3' sP_4' -C
product
trans peptide splicing
sP_4 sP_3 sP_2 sP_1 sP_1' sP_2' sP_3' sP_4'

Figure 10. – Peptide hydrolysis and types of proteasomal splicing.

The event in a) represents the normal peptide hydrolysis reaction by the proteasome whereas b) and c) depict proteasomal-catalyzed cis peptide splicing whereby fragments different parts of the same protein/substrate get ligated together either In the same order as in the substrate or in reverse order. Ligation of fragments derived from two different proteins in shown in d) *Adapted from (Specht et al., 2020).*

Interestingly, Ebstein (2016) et al. And Faridi et al. both found that multiple patients harbored T cells specific against (Ebstein et al., 2016; Faridi et al., 2020). This meant that these spliced peptides are spliced similarly across patients, also hinting at the non-randomness of peptide cis splicing.

1.8. Bio-informatic tools

Hybrid peptides are a novel source of neoantigens and thus constitute key potential targets for immunotherapies. Their identification is evidently of great importance and bio-informatic tools could facilitate this task.

1.8.1. The need for open-source and accessible software to identify potential hybrid peptide by MS

The identification of proteasome-catalyzed spliced peptides from MS is not straightforward since their hybrid nature would not allow for their direct identification with database search of the human proteome. Thus, bio-informatic tools are warranted for the identification of these.

Five algorithms/workflows have been developed for the purpose of identifying spliced peptides: SpliceMet, HybridFinder, SPI-delta (Spliced Peptide Identifier - delta version), TagPep, Neo-Fusion.

SpliceMet, knowing the protein substrate a database of all possible theoretical spliced peptides is created and theoretical m/z values are calculated and for all charge states (z=+1, +2, +3) and compared against their measured counterpart MS. Furthermore, a kinetic aspect was considered whereby the authors determined that the higher quality in-vitro proteasome catalyzed peptide splicing happens up until a maximum timepoint after which the peptides produced are due to a re-entry of the fragments into the proteasome. This aids in the formation of an “inclusion list” based on the m/z and this kinetic aspect which then are confirmed by comparison with synthetic peptides followed by validation with MALDI TOF (Liepe et al., 2010).

The HybridFinder (2018) workflow, is based on analysis of both database search results and ‘de novo’ peptide sequencing in PEAKS software from MS data. ‘De novo’ sequencing refers to a bottom the identification of peptide sequences in PEAKS software (independently of a protein database) and relies on forming the best possible combinations of amino acids based on b- and y-ions(Ma et al., 2003). HybridFinder uses both results to determine unassigned spectra from de novo and then goes through a sequence of checks for whether the sequence is non-spliced/linear or not. If not, it divides the sequence and checks whether both fragments are from the same protein (cis), or from two different proteins (trans). Then, one sequence from each spectrum is picked. Generally, non-spliced/linear sequences (linear > cis > trans) are more likely and therefore in a given spectrum if there are identified linear and spliced candidates, the spliced peptide candidates are discarded. Otherwise, in cases of ties the ALC score (Average Local Confidence score: a PEAKS-specific score that attributes confidence values for the predicted amino acids at each position) for ranking the peptide sequences. Then, a database is formed by concatenating the original database used for the search and the one with the spliced peptide candidate followed by a second database search (Faridi et al., 2018).

Mylonas et al. (2018) employed TagPep workflow to compare the estimation of the contribution of spliced peptides to the HLA-I ligandome with the original estimations around ~30% (Liepe et al., 2016; Mylonas et al., 2018). Tagpep follows a similar workflow to HybridFinder however it does not search for trans-spliced peptides and the second database is done in other tools (MaxQuant and Comet) rather than in PEAKS again. Their re-analysis of the HLA-I ligandome with Tagpep revealed that the actual contribution is between 1 and 3% (Mylonas et al., 2018).

In 2018, Rolfs et al. developed Neo-Fusion, the first open-source software for the identification of spliced peptides (Rolfs et al., 2018). Non-spliced peptides are first identified, and then potential decoys. Next, Neo-fusion creates two databases: an N-terminal ion database consisting of b-ions and a C-terminal one with y-ions. Then, to determine the peptide fragments that form a spliced peptide, it uses open mass search and keeps the sequences that match while discarding those that do not. Neo-Fusion also identifies trans-spliced peptides.

Lastly, another tool developed by Liepe's lab (2020), is SPI-Delta, this tool was especially developed to identify spliced and non-spliced peptides derived from given substrates. For their database they have used 55 substrates. First, a custom database is built containing all linear cis-spliced sequence possibilities of the substrate. Then, after a search using this database, Mascot software ranks and ion scores are used to filter peptides. and linear peptides assigned within a spectrum are prioritized to spliced peptides (linear>spliced) (Specht et al., 2020).

Finally, SpliceMet and SPI-Delta rely on a forward approach and require an initial knowledge of the substrates to be used. Therefore, with no prior knowledge of the substrate the workflows of Faridi, Mylonas and Rolfs are well-suited for the identification of spliced peptides, as they allow for a reverse immunology approach. Nevertheless, it is also clear that this identification cannot be done solely computationally but should be done in two steps: 1) a computational identification approach followed by 2) the high confidence confirmation of the source and presentation of these peptides which is only possible through wet lab experiments (likened to earlier experiments for the identification of these) as suggested in our paper and, presented in the next chapter (Saab et al., 2021). Keeping in mind that one previously identified spliced peptide also contained another post-translational modification(Dalet et al., 2011). Also, the intervening distance (the number of amino acids that separate the peptide fragments in the original sequence) is still an issue that was not solved. Neo-Fusion uses a default intervening of 25 amino acids. However, the initial spliced peptide had an intervening distance of 41 amino acids, whereas others had lower numbers. Therefore, their computational identification should rely on a broad approach followed by thorough validation. Moreover, the use of multiple search engines in order to increase confidence is a valuable approach before experimental validation, as is done in the workflow of Mylonas et al.(Mylonas et al., 2018).

Nevertheless, the availability of the computational workflows for other researchers in the community is indispensable. To our knowledge, no software was widely distributed: open-source and available on multiple OS at the time of publication. Moreover, using the HybridFinder algorithm (or a similar workflow), has allowed several researchers to identify and confirm the source and immunogenicity of their identified spliced peptides (Faridi et al., 2020; Kato et al., 2021). In summary, the bio-informatic tools for the identification of spliced peptides are indispensable and spliced peptides could have great clinical potential as novel anti-cancer targets for the development of new immunotherapies targeting cancer cells using these.

1.8.2. The need for bio-informatic tools to predict gene fusion derived neoantigens

Two great tools have been developed until now for the prediction of gene fusion-derived neoantigens from patient transcriptomic data. These tools first identify the gene fusions and the patient's HLA typing. Then, they apply selection filters based on expression and predicted effects of frame retention or shift, followed by peptide sequence prediction and then peptide-MHC-I predictions. Integrate-Neo uses an older version of netMHCpan (4.0) and uses the binding affinity results while applying a threshold of 500nM for the selection of neoantigens(Zhang et al., 2017). On the other hand, pVACtools and specifically pVACfuse (module for gene fusion) follows a similar workflow to IntegrateNeo, and even integrates it. However, pVacfuse uses multiple peptide-MHC-I prediction algorithms. It also uses binding affinity as predictor and emphasizes a 500nM threshold. However, it also incorporates interesting additional metrics to score neoantigens, such as dissimilarity from self, proteasomal cleavage or abundance. The binding prediction algorithms used by pVACtools have two disadvantages: (1) the versions used are trained on older data and (2) certain binding prediction algorithms used which are provided by IEDB (Immune Epitope DataBase) cannot perform predictions on all peptide lengths and for all alleles (this is also indicated on the IEDB website: http://tools.iedb.org/mhci/help/). Furthermore, additional tools used include netCTL for which also as the netCTL homepage indicated is that it is optimized for 9-mer peptides (https://services.healthtech.dtu.dk/service.php?NetCTLpan-1.1), and only provides approximations for non 9-mer peptides. The recent versions of both netMHCpan and MHCflurry not only emphasize the presentation (elution aspect) but are also trained on newer data. Next, both older and newer versions of netMHCpan, have predicted a fusion peptide (RIAECILGM) derived from the ETV6-RUNX1 gene fusion to be a weak binder to HLA-A*02:01 (surpassing the threshold adopted by either tool), however tetramer assay showed strong MHC-restricted recognition by T cells (Zamora et al., 2019a). Taken together, we can conclude that 1) applying stringent thresholds could lead to missing out valuable fusion peptide candidates, 2) these tools do not consider the recurrence of gene fusions, 3) these tools use old versions of the algorithms and 4) are not relying on the eluted ligand predictions.

1.9. Hypotheses & objectives

The objective of my master’s thesis was to develop bio-informatic tools that would aid in the identification of hybrid peptides that bind the MHC-I molecule. Essentially, these hybrid peptides are attractive targets for targeted immunotherapy strategies.

We were particularly interested in the identification of MHC-I peptides derived from gene fusion events and proteasomal peptide splicing. We hypothesized that it is possible to develop bio-informatic tools for the identification of these.

Of note, because the peptides that undergo proteasome catalyzed splicing cannot be identified beforehand, we have hypothesized that using a bottom-up approach through their identification from mass spectrometry data of presented MHC-I peptides would constitute an unbiased method. Therefore, we aimed to develop a tool in the form of an R package that would be easily accessible for the analysis of Mass spectrometry data in order to identify hybrid peptides based on a validated workflow. The workflow by Faridi et al. (2018) or a variation of it has been instrumental to the discovery of several spliced peptides, (Faridi et al., 2018, 2020; Kato et al., 2021) Therefore, we aimed to develop a R package for the systematic discovery of spliced peptides based on this workflow.

On the other hand, gene fusion transcripts are commonly detected by RNAseq (Haas et al., 2019), therefore, we hypothesized that it would be possible to do that in a forward approach. This would entail predicting MHC-I peptides from the gene fusion protein sequence. To do that, we developed FusionChoppeR (FCR) that predicts MHCI-fusion junction peptides using a combination of two peptide-MHC-I prediction algorithms and a set of HLA-I alleles covering all supertypes. Moreover, we hypothesized that due to their hybrid nature, gene fusion-derived peptides would bind more than one HLA-I allele. To test that, we took advantage of a database of gene fusion transcripts identified in cancer cell lines (LiGEA). Nevertheless, we expected that peptide-MHC-I predictions might reveal that fusion peptides could bind more than a single HLA-I allele. Therefore, we also sought to take advantage of an immunopeptidomic dataset, specifically the one by Sarkizova et al. (Sarkizova et al., 2020a) in order to analyze to whether peptides could bind multiple HLA-I alleles. Finally, to test our hypothesis further on clinically relevant gene fusions, our aim was to select a few candidates for computational (clustering, 3D modeling) as well as in-vitro peptide-HLA-I binding, after applying FCR on 1) a published fusion panel of gene fusions found in pediatric patients and 2) a recurrent gene fusion (CBFA2T3-GLIS2) with an unfavorable prognosis identified in a cell line (M0E7 cell line).

Chapter 2: Published Article

Title : RHybridFinder: An R Package to Process Immunopeptidomic Data for Putative Hybrid Peptide Discovery

Authors

Frederic Saab[1,5], David J. Hamelin[1], Qing Ma[4], Kevin Kovalchik[1], Isabelle Sirois[1], Pouya Faridi[2], Chen Li[2], Anthony Purcell[2], Peter Kubiniok[1,5]*, Etienne Caron[1,3,6]*

Affiliations

[1]CHU Sainte-Justine Research Center, Montreal, QC H3T 1C5, Canada

[2]Infection and Immunity Program and Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia

[3]Department of Pathology and Cellular Biology, Faculty of Medicine, Université de Montréal, QC H3T 1J4, Canada

[4]School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, ON K1N 6N5, Canada

[5]Technical Contact

[6]Lead Contact

*Correspondence: Peter Kubiniok and Etienne Caron.

Author Contributions

F.S. & P.K.: R code, conceptualization & authorship, Q.M., D.H., K.K., I.S : validation & conceptualization, P.F., C.L., A.P. :conceptualization, authors of the original workflow, E.C.: Conceptualization & authorship

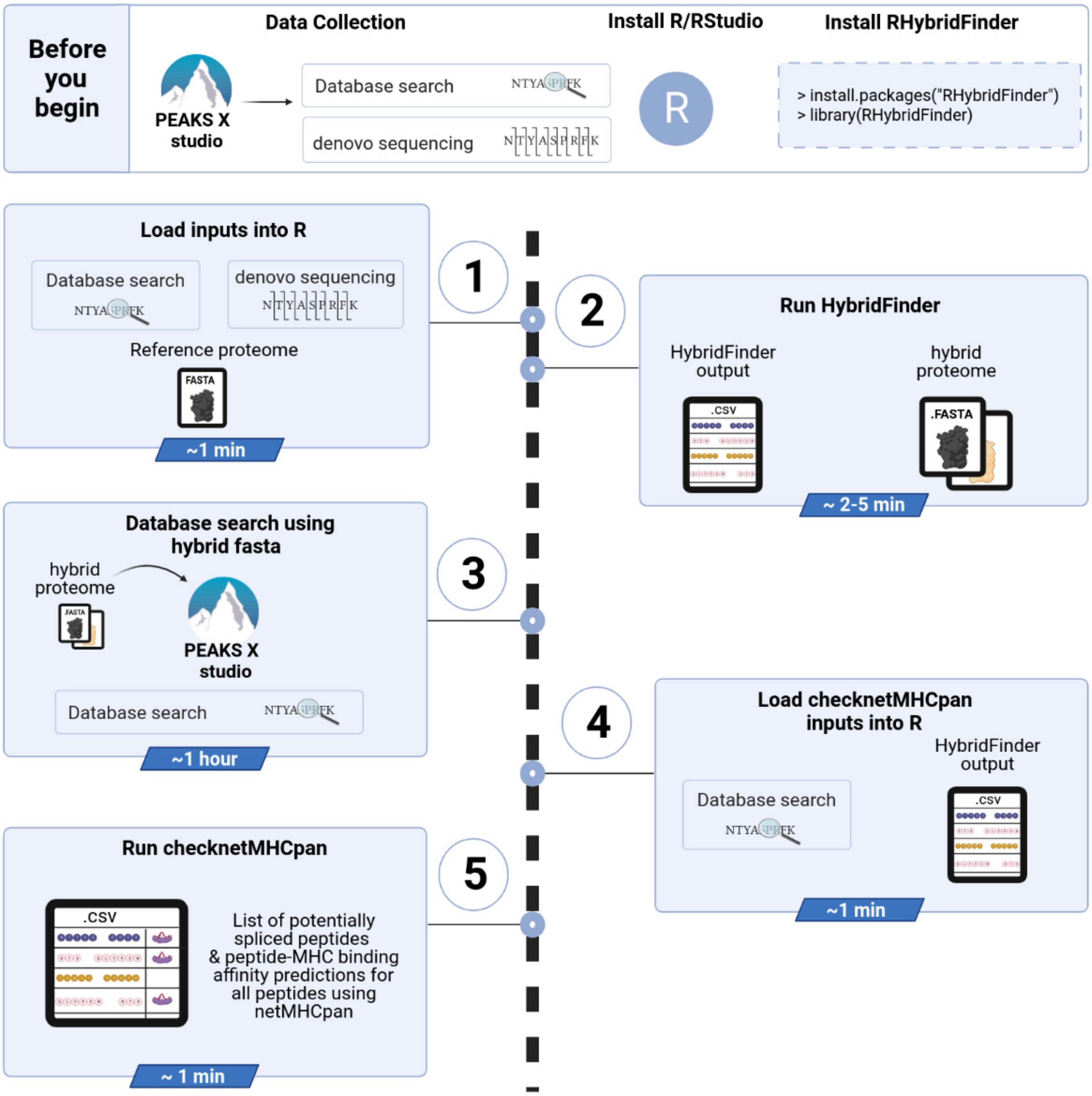


Figure 11. – RHybridFinder graphical abstract.

2.1. Abstract

The identification of proteasomal spliced peptides (PSPs) in mass spectrometry (MS) data is not straightforward since PSP sequences do not fully reflect the same sequence as the parent protein. RHybridFinder sequentially searches in candidate spectral sequences for linear cis- and trans- potential matches in the proteome from PEAKS results. Here, we provide a protocol for running RHybridFinder for

MS-based identification of PSPs, built using R and parallel computing in order to compute the results in an efficient manner.

For complete details on the use and execution of this protocol, please refer to Faridi et al. (2018).

2.2. Before You Begin

The proteasome is recognized as the core enzymatic machinery of the antigen processing and presentation pathway wherein peptides derived from proteasomal proteolysis are selectively presented on the cell surface by MHC (Major Histocompatibility Complex)-I molecules (Neefjes et al., 2011). Following Hanada's (2004) groundbreaking discovery of a FGF-5 (Fibroblast Growth Factor-5)-derived proteasomally-spliced peptide from a renal cancer patient (Hanada et al., 2004), other research groups have been able to uncover additional spliced peptides presented by MHC class I molecules, referred in this protocol as proteasomal spliced peptides (PSPs) (Dalet et al., 2011; Ebstein et al., 2016; Michaux et al., 2014; Vigneron et al., 2004; Berkers et al., 2015).

More recently, MS-based immunopeptidomics has been used to expedite their identification in a systematic manner, including cis- and trans-spliced peptides (Liepe et al., 2010; Berkers et al., 2015; Liepe et al., 2016; Faridi et al., 2018; Rolfs et al., 2019; Specht et al., 2020). However, MS-based studies using different approaches have led to a debate around the proportion of these PSPs in the MHC class I immunopeptidome (Liepe et al. 2016; Mylonas et al., 2018; Lichti, 2021; Willhelm et al., 2021).

Here, we provide an open access and improved R package built upon the computational workflow developed by Faridi et al. (2018) for the analysis of MS data to systematically identify putative PSPs (Faridi et al., 2020). Importantly, we do wish to emphasize throughout this protocol that while RHybridFinder enables the identification of putative PSPs, we do not claim that it can identify high-confidence PSPs that are genuinely spliced by the proteasome in vivo. Therefore, PSPs obtained using this protocol should then be experimentally validated using rigorous biochemical and immunological assays (**Figure 12**).

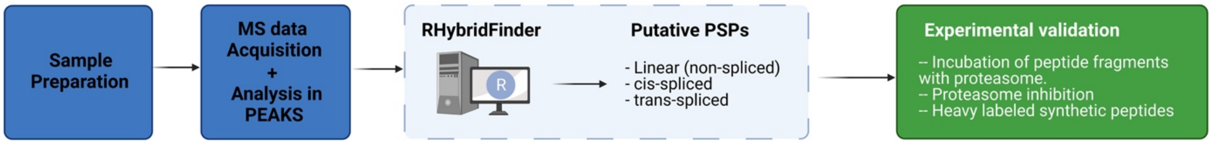


Figure 12. – Overview of suggested workflow for the discovery of PSPs.

We propose a four-step workflow for the identification of PSPs. The first three steps (blue squares: sample preparation, MS data acquisition and RHbridFinder enable computational exploration of putative PSPs followed by experimental validations (green square). A non-exhaustive list of possible experiments is shown for validating/gaining confidence in the identification of MHC-I peptides that are genuinely catalyzed by proteasomal splicing.

In order to facilitate and optimize how the protocol would be run, we propose a folder structure as shown in Figure 13.

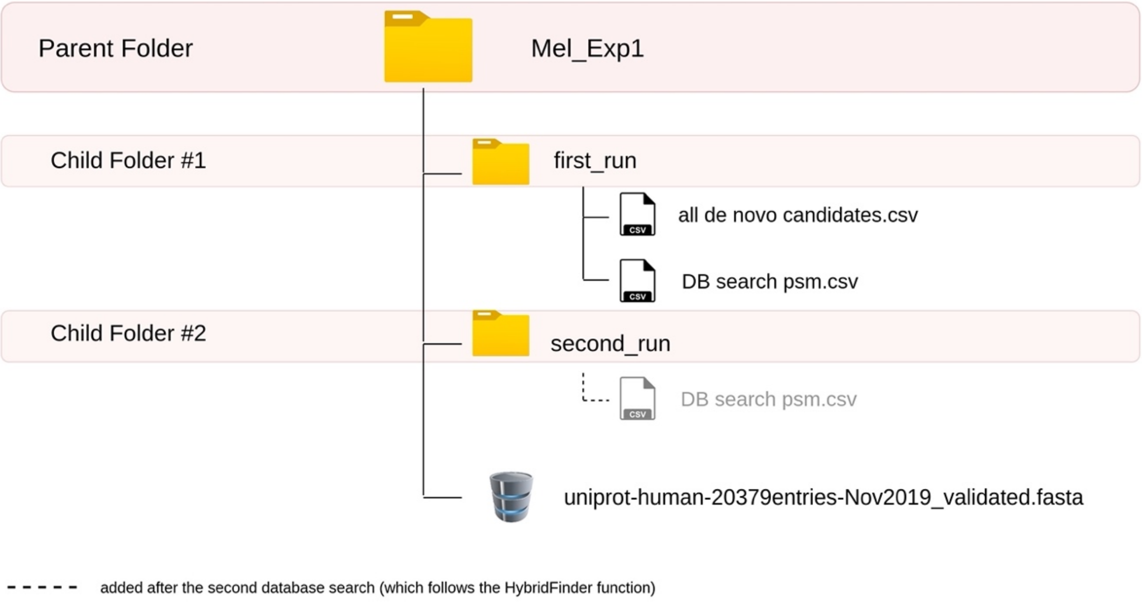

Parent Folder
Mel_Exp1
Child Folder #1
first_run
all de novo candidates.csv
DB search psm.csv
Child Folder #2
second_run
DB search psm.csv
uniprot-human-20379entries-Nov2019_validated.fasta
CSV
- - - - - added after the second database search (which follows the HybridFinder function)

Figure 13. – Recommended folder structure.

The parent folder includes two child folders. The child folders include the various files that are necessary for running RHybridFinder. The dotted line (second_run) indicates that the DB search psm.csv file is added after the second DB search.

RHybridFinder is available on CRAN (https://cran.r-project.org/package=RHybridFinder) to enable more researchers to explore those debated peptides.

2.2.1. Data collection

For demonstration of the output of the different RHybridFinder functions, we have used datasets from the HLA Ligand Atlas (Marcu et al., 2021) deposited in PRIDE (Proteomics IDentification Database) PXD019643.

1. Download the following mzML files and analyzed them in PEAKS:
   171002_AM_AUT01-DN17_Liver_W6-32_10%_DDA_3_400-650mz_msms4,
   171002_AM_AUT01-DN17_Liver_W6-32_10%_DDA_3_400-650mz_msms5,
   171002_AM_AUT01-DN17_Liver_W6-32_10%_DDA_3_400-650mz_msms6.

2.2.2. Analyze these files in PEAKS Installing Rstudio/R

RHybridFinder package has been developed in RStudio (Rstudio Team, 2020) and implemented in R (R Core Team, 2017) programming language.

2. Download & install Rstudio if not already installed: (https://www.rstudio.com/products/rstudio/download/).

2.2.3. Installing and loading RHybridFinder

Below are the lines needed to install the RHybridFinder package from CRAN (the Comprehensive R Archive Network) and then load it.

2.2.3.1. Install and load RHybridFinder by typing “install.packages(“RHybridFinder”) in the R console.

```
> install.packages(“RHybridFinder“)
```

2.2.3.2. Load RhybridFinder by typing “library(RHybridFinder)” in the R console

```
> library(RHybridFinder)
```

CRITICAL: if you copy the lines of code from here, keep in mind that you might have to re-write the quotation marks yourself.

2.3. Step-by-step method details

2.3.1. Step 1: Load inputs into R

Timing: 1 minute

Before running HybridFinder, the inputs need to be loaded into R. We propose the following way of loading the files into R in order to facilitate the process

10. Create an object (folder_Exp1) for the path to the parent folder (Mel_Exp1) (but both can be named otherwise).

```
> folder_Exp1 <- file.path(“/Users/YOURUSERNAME/Desktop/Mel_Exp1”)
```

11. Import the *de novo* sequencing as well as the database results, both of which are located in the first_run child folder.

    a. *de novo* sequencing results file

    ```
    > denovo_Exp1 <- read.csv(file = file.path(folder_Exp1, “first_run”,
    "all_denovo_candidates.csv"), header=TRUE, sep=",", stringsAsFactors = FALSE)
    ```

    b. database search results file

    ```
    > db_search_Exp1<- read.csv (file=file.path(folder_Exp1, “first_run”,“DB
    seach psm.csv”), header=TRUE, sep=“,”, stringsAsFactors=FALSE)
    ```

12. Create an object for the path to the proteome file, located in the parent folder (folder_Exp1) (see refproteome_Exp1, in the example below). The fasta proteome will be imported in R during the HybridFinder function.

```
> refproteome_Exp1 <- file.path(folder_Exp1, “uniprothuman-20379entries-
Nov2019_validated.fasta”)
```

CRITICAL: Please note that if you copy the file access path (in windows), you will need to switch the backslash (“\”) to a normal slash (“/”).

Access the datasets included in the R package: The RHybridFinder package also includes demonstration datasets from the HLA Ligand Atlas that have already been analyzed in PEAKS. These datasets include PEAKS de novo sequencing results and PEAKS database search results.

```
# access denovo dataset

> data(package= “RHybridFinder”, “denovo_Human_Liver_AUTD17”)

# access database search dataset

> data(package=”RHybridFinder”, “db_Human_Liver_AUTD17”)
```

Note: that due to size constraints the proteome database (.fasta) file is not included in the package. It can be downloaded from the Uniprot database.

Note: In the environment tab, the denovo_Human_Liver_AUTD17 and db_Human_Liver_AUTD17 should appear. Note that if you see <promise>, after clicking on the objects, the data would appear.

2.3.2. Step 2: Run HybridFinder

Timing: 2-5 minutes (with parallelism, 8 cores) - 10-15 minutes (without parallelism)

In order to have a relatively short runtime, we have implemented an option to use parallel computing. However, please note that because parallel computing requires a certain amount of processing units for proper functioning, it has been made possible to also run HybridFinder without parallel computing.

Based on default parameters in the HybridFinder function, the “all de novo candidates.csv” file contains 16,286 peptide sequences and the runtime (parallelism with 8 cores) is of 2 minutes 17 seconds. ~5 minutes are required for double the number of peptides. Without parallelism, the runtime ranged between 10 and 15 minutes for 16,286 peptide sequences.

13. Run Hybridfinder (Please refer to table 1 in order to know more about the inputs needed) and export the results in the parent folder.

```
> HybridFinder_results_Exp1<- HybridFinder(denovo_candidates = denovo_Exp1, db_search
= db_search_Exp1, proteome_db = refproteome_Exp1,customALCcutoff = NULL,
with_parallel=TRUE, customCores = 8, export_files= TRUE, export_dir = folder_Exp1)
```

CRITICAL: if you use the datasets included in the package, please note that they are named differently so for instance the “denovo_candidates” and “db_search” parameters should be set to the datasets loaded from the package: denovo_Human_Liver_AUTD17 and db_Human_Liver_AUTD17, respectively.

CRITICAL: Make sure to store the HybridFinder results in an object (i.e HybridFinder_results_Exp1), as the HybridFinder output dataframe will come in handy in the second function.

Note: At the end of the hybrid proteome will be the concatenated hybrid fake proteins with the name pattern ‘sp|denovo_HF_fake_protein_[#]’.

Parameter	**Description**	**Default value**
de novo_candidates	the dataframe containing the *de novo* sequencing results	No defaults. Necessary input.
db_search	the data frame containing the database search results	No defaults. Necessary input.

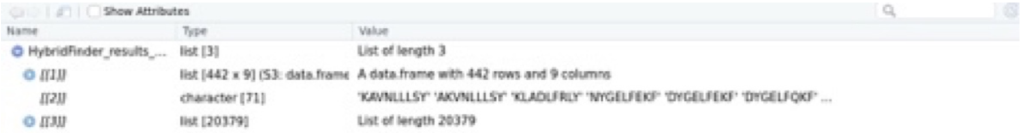
db_search	the data frame containing the database search results	No defaults. Necessary input.
proteome_db	the file path to the proteome used for the database search	No defaults. Necessary input.
(Optional) customALCcutoff	A custom score cutoff that can be set by the user as long as it would be at least 85	NULL. (ALC cutoff calculated automatically as median of matching peptide sequences of assigned spectra). If set manually, minimum is 85.
with_parallel : boolean (True or False)	representing whether parallel computing should be employed for running the function.	TRUE
(Optional) customCores	If with_parallel is set to TRUE and the PC has >5 cores, the user can set a custom amount of cores to be used by the function.	6
(Optional) export_files : boolean (True or False)	by default it is set to False, however, if set to True, then the following input is essential.	FALSE
(Optional) export_dir	file path to the directory where the output files should be stored. This parameter is necessary for the export.	NULL

Tableau 2. – HybridFinder function parameters

Note: with_parallel is activated if set to true and if the PC has more than 5 cores.

CRITICAL: Please ensure to have a minimal number of other windows open and to save any work in other softwares prior to using HybridFinder with parallelism.

The function will output a list (**Figure 14**) containing: (1) the HybridFinder output containing all the denovo peptides along with their potential splice type explanation cis-/trans-, (2) a list of the step1 hybrid candidate peptides, (3) the hybrid proteome (merged proteome: the original user proteome along with the hybrid proteome composed of the concatenated candidate hybrid peptide sequences).



Name	Type	Value
HybridFinder_results_...	list [3]	List of length 3
[[1]]	list [442 x 9] (S3: data.frame	A data.frame with 442 rows and 9 columns
[[2]]	character [71]	'KAVNLLLSY' 'AKVNLLLSY' 'KLADLFRLY' 'NYGELFEKF' 'DYGELFEKF' 'DYGELFQKF' ...
[[3]]	list [20379]	List of length 20379

Figure 14. – Screenshot of the HybridFinder function results

In the results list you will find 3 items: 1) a dataframe containing the HybridFinder output. 2) a character vector containing the candidate spliced peptides. 3) a list which is in a seqinr class (Charif et al., 2007) containing the merged hybrid proteome.

Note: In the example above, export_files have been set to TRUE and the export_dir has been defined which means that the files are also automatically exported. If these two parameters were not specified or were set to FALSE & NULL, the results are only stored in the Exp1_HybridFinder_results. In this case, you can still use "export_HybridFinder_results" as in the code below, where HybridFinder_results_Exp1 is the object created above for the storage of HybridFinder results.

```
> export_HybridFinder_results(HybridFinder_results_Exp1, export_dir= folder_Exp1)
```

Pause Point: If you would like to conduct the rest of the protocol at a later time, either use the export functionality and then load the hybridfinder output in order to use it for the second step. Alternatively, save the objects in R in a .rda file as follows and once you want to use it again for the step 4, load checknetMHCpan inputs into R.

```
>         save         (HybridFinder_results_Exp1,          file=file.path(folder_Exp1,
”HybridFinder_results_Exp1.rda”)
```

```
>load (file.path(folder_Exp1, ”HybridFinder_results_Exp1.rda”))
```

2.3.3. Step 3: Database search using hybrid Fasta

Timing: 1 hour

An essential interim step must follow the HybridFinder function and consists of running a database search in PEAKS with the merged proteome. Importantly, now that a merged hybrid proteome has been obtained from the hybridfinder function, it can be used to obtain potential PSPs whose quality is comparable with all other database search peptides while filtering all peptides at the same FDR (False Discovery Rate) cutoff which can be adjusted by the users in PEAKS. In the original workflow by Faridi et al. (2018), the database search peptides in both runs were filtered in PEAKS at a 1% FDR.

14. Perform a database search in PEAKS using the original raw MS file (while using the same settings as in the beginning) however, this time while using the merged hybrid proteome (.fasta) file generated with the HybridFinder function.

2.3.4. Step 4: Load checknetMHCpan inputs into R

Timing: 1 minute

Prior to running checknetMHCpan, please ensure that netMHCpan (versions 4.0 or 4.1) is installed. checknetMHCpan is the last step of the hybrid finder workflow, the function uses the database search results from the second PEAKS analysis and provides the binding affinity results of all the peptides along with their categorizations.

15. Create an object for the location of the netMHCpan executable

```
> netmhcpan_dir <- file.path(“/usr/local/bin”)
```

16. Create an object (vector) for storing the HLA-I alleles that you would like to have binding affinity predictions for.

```
> alleles_Exp1 <- c("HLA-A*02:01", "HLA-A*03:01", "HLA-B*07:02")
```

17. Retrieve the hybridfinder output from the HybridFinder function results

```
> HF_output_Exp1 <- HybridFinder_results_Exp1[[1]]
```

18. Import the database search results (from step 3: Database search using hybrid fasta)

```
> rerun_db_search_Exp1 <- read.csv(file.path(folder_Exp1, "second_run", "DB search
psm.csv"), sep=",", head = TRUE, stringsAsFactors = FALSE)
```

Note: in case your computer's OS is "Windows" (netMHCpan is not compatible with Windows) the web version of netMHCpan (http://www.cbs.dtu.dk/services/NetMHCpan-4.1/instructions.php) would come in handy. In this case, we propose to use a separate function from this package instead (step2_wo_netmhcpan) which outputs a netMHCpan-ready input of sequences in .pep format.

Access the datasets included in the R package: The demonstration datasets from the HLA Ligand Atlas included in this package also include datasets for the checknetMHCpan/step2_wo_netMHCpan functions. After having run the HybridFinder function and stored the results in HyrbidFinder_results_Exp1, PEAKS was run using the merged hybrid proteome. Below is a way to retrieve the second PEAKS run dataset included in the package:

```
> data(package= "RHybridFinder", "db_rerun_Human_Liver_AUTD17")
```

Note: The merged proteome used for the second database search is based on the customALCcutoff being set to NULL (default parameter value).

CRITICAL: The merged proteome database would change between different samples, and if the customALCcutoff parameter is changed. The same merged hybrid proteome cannot be used for separate analyses.

2.3.5. Step 5: Run checknetMHCpan

Timing: ~ 1 minute

The checknetMHCpan function embodies the second major step of the workflow. The categorizations of the hybrid peptides from the hybridfinder output are retrieved for matched peptides found in the second PEAKS database results. Then, peptide-MHC class I binding predictions for the entire database search results (for peptides between 9 and 12 amino acids) are computed using netMHCpan and are tidied in order to summarize the results.

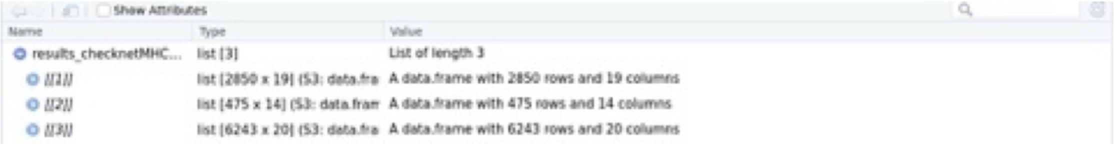
19. Run checknetMHCpan using the code below (Please refer to table 2 in order to know more about the inputs needed) and export the results in the same folder:

```
> checknetMHCpan_results_Exp1 <- checknetMHCpan(netmhcpan_directory = netmhcpan_dir,
netmhcpan_alleles  =  alleles_Exp1,  peptide_rerun  =  rerun_db_search_Exp1,
HF_step1_output  = HF_output_Exp1, export_files= TRUE, export_dir = folder_Exp1)
```

Note: checknetMHCpan is compatible with the exports from both netMHCpan 4.0 & netMHCpan 4.1.

CRITICAL: if you use the datasets included in the package, please note that they are named differently so for instance the "peptide_rerun" parameter should be set to dataset loaded from the package db_rerun_Human_Liver_AUTD17.

After running the code above, a results list should be returned (**Figure 15**).



Name	Type	Value
results_checknetMHC...	list [3]	List of length 3
[[1]]	list [2850 x 19] (S3: data.fra	A data.frame with 2850 rows and 19 columns
[[2]]	list [475 x 14] (S3: data.fram	A data.frame with 475 rows and 14 columns
[[3]]	list [6243 x 20] (S3: data.fra	A data.frame with 6243 rows and 20 columns

Figure 15. – Screenshot of the checknetMHCpan results list.

In the results list you will find 3 items: 1) a dataframe containing the netMHCpan results. 2) a dataframe containing the tidied netMHCpan results. 3) the database search results with the "Potential_spliceType" for the hybrid peptides retrieved from step1.

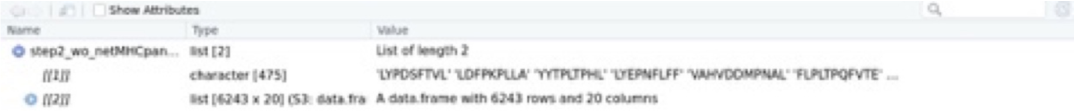
Table 2. checknetMHCpan function parameters

Parameter	Description	Default value
netmhcpan_directory	the directory where netMHCpan is installed (i.e '/usr/bin' or '/usr/local/bin', depending on where you have it installed)	No defaults. Necessary input.
netmhcpan_alleles	a vector composed of the alleles the peptides will be tested against.	No defaults. Necessary input.
peptide_rerun	the database search results from the second peaks run	No defaults. Necessary input.
HF_step1_output	the data frame from the HybridFinder function of the containing the spliced peptide potential explanations as well as RT, m/z, ALC, Scan & Fraction	No defaults. Necessary input.
(Optional) export_files : boolean (True or False)	by default it is set to False, however, if set to True, then the following input is essential.	FALSE
(Optional) export_dir	file path to the directory where the output files should be stored. This parameter is necessary for the export.	NULL

These results are also exportable with the export_checknetMHCpan_results function.

```
> export_checknetMHCpan_results(step2_RHF_results_Exp1 , export_dir = folder_Exp1)
```

Note: If you intend on using the web version of netMHCpan (especially useful for windows OS users) or another software for peptide binding affinity, the step2_wo_netMHCpan function does the same as checknetMHCpan but without running netMHCpan. The function should return a list (**Figure 16**) containing the updated database search results as well as a list of the peptides which can be used as input in the web version of netMHCpan.



Name	Type	Value
step2_wo_netMHCpan...	list [2]	List of length 2
[[1]]	character [475]	'LYPDSFTVL' 'LDFPKPLLA' 'YYTPLTPHL' 'LYEPNFLFF' 'VAHVDDMPNAL' 'FLPLTPQFVTE' ...
[[2]]	list [6243 x 20] (S3: data.fra	A data.frame with 6243 rows and 20 columns

Figure 16. – Screenshot of the step2_wo_netMHCpan results list.

In the results list you will find 2 items: 1) a character vector containing the netMHCpan-ready input. 2) the database search results with the "Potential_spliceType" for the hybrid peptides retrieved from step1.

Expected Outcomes

HybridFinder

The HybridFinder function follows the same rationale as indicated in Faridi et al. (2018). After high-confidence *de novo* peptides are extracted, these are searched sequentially for an exact hit, followed by a search of pair fragments within one protein and then within two proteins (**Figure 17**). Finally, the sequences of all hybrid peptides are concatenated to create fake proteins, which are added at the bottom of the proteome database in order to constitute a merged hybrid proteome.

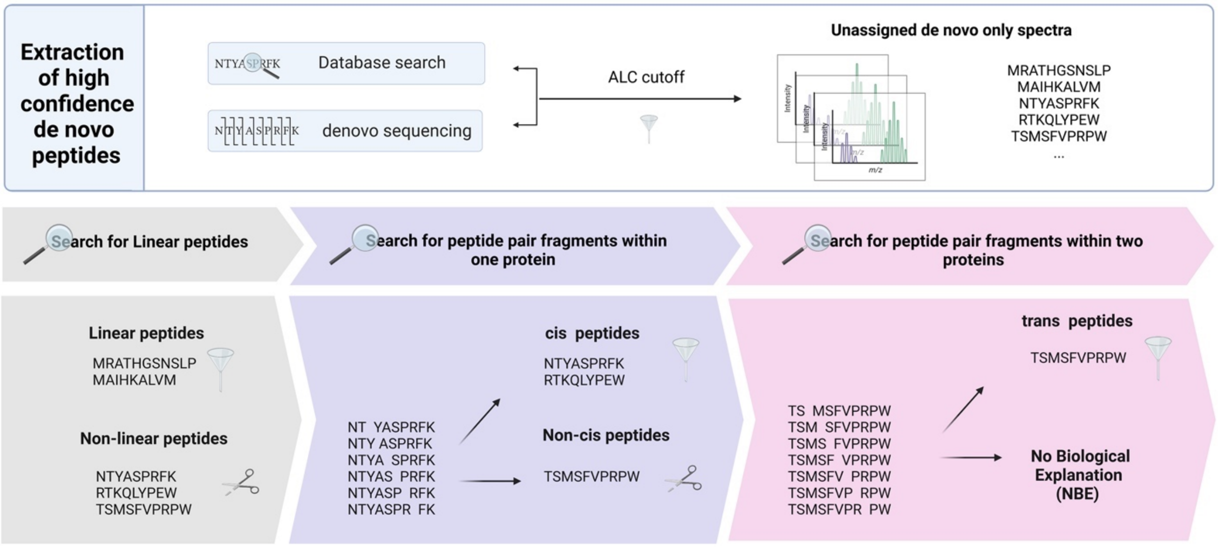
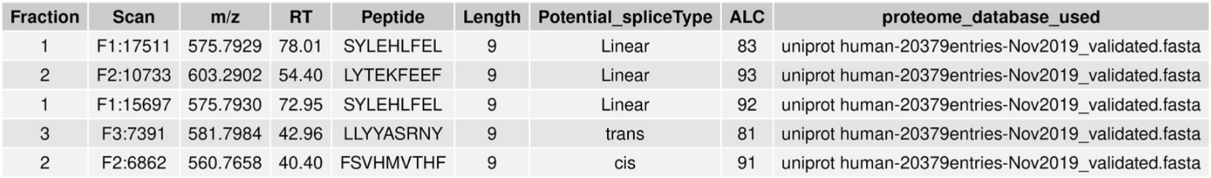


Figure 17. – HybridFinder function.

HybridFinder extracts high confidence de novo peptides by using a ALC cutoff based on the median ALC of common spectrum groups & sequence of peptides between the de novo and the database search. The ALC cutoff is used to filter unassigned de novo spectrum groups in order to obtain high confidence de novo spectra. All sequences are then searched in the proteome for the entire sequence, those that match are filtered and considered "Linear", the remainder of the peptide spectrum groups are "cut" in order to create peptide fragment combinations. These are then searched in the proteome for whether fragment combinations exist within a same protein, matches are considered as cis-spliced and further filtered. Finally, fragment combinations are created from those that didn't match in the previous step and are searched whether they exist in two proteins. If there is a match, these are considered as trans-spliced peptides. The remaining uncategorized spectrum groups are considered not to have a biological explanation (NBE) and are therefore discarded.

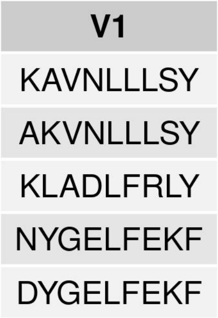
Typically, when the HybridFinder function is run, 3 messages are printed representing each major stage of the algorithm and finally 'Done!' is printed once the processing is finished. The function returns a list containing 3 items: the hybridfinder output (**Figure 18**) where the predicted splice type is displayed, a character vector containing only the list of hybrid candidates (**Figure 19**) and finally the merged hybrid proteome (**Figure 20**) where the hybrid peptide candidates have been concatenated as fake proteins.



Fraction	Scan	m/z	RT	Peptide	Length	Potential_spliceType	ALC	proteome_database_used
1	F1:17511	575.7929	78.01	SYLEHLFEL	9	Linear	83	uniprot human-20379entries-Nov2019_validated.fasta
2	F2:10733	603.2902	54.40	LYTEKFEEF	9	Linear	93	uniprot human-20379entries-Nov2019_validated.fasta
1	F1:15697	575.7930	72.95	SYLEHLFEL	9	Linear	92	uniprot human-20379entries-Nov2019_validated.fasta
3	F3:7391	581.7984	42.96	LLYYASRNY	9	trans	81	uniprot human-20379entries-Nov2019_validated.fasta
2	F2:6862	560.7658	40.40	FSVHMVTHF	9	cis	91	uniprot human-20379entries-Nov2019_validated.fasta

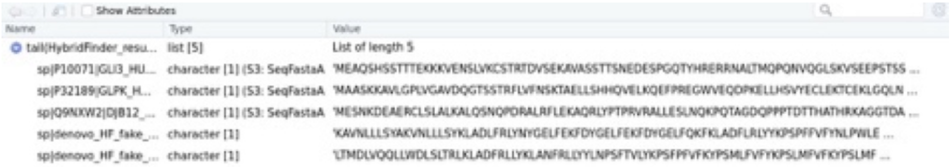
Figure 18. – Screenshot of the HybridFinder output dataframe (5 rows)

The Fraction column represents the LC-MS run, the Scan column is a number representing a unique index for the tandem mass spectra (F[Fraction#]:Scan#), m/z is the precursor mass-to-charge ratio, RT is the Retention Time (elution time) for the spectrum, Peptide corresponds to the peptide sequences. The Length column represents the number of amino acids for a given peptide, ALC (Average Local Confidence), is a score calculated in PEAKS as the total of the residue local confidence scores in the peptide divided by the peptide length. These columns are not provided by the HybridFinder function, they are columns found in any PEAKS de novo sequencing export. For more information, please visit the PEAKS user manual. The Potential_spliceType corresponds to the resulting categorization from the HybridFinder function. Finally, the proteome_database_used is the filename of the fasta proteome provided by the user (this column is mainly for helping the user keep track of the proteome used) in the HybridFinder function.



V1
KAVNLLLSY
AKVNLLLSY
KLADLFRLY
NYGELFEKF
DYGELFEKF

Figure 19. – Screenshot of the HybridFinder hybrid peptide candidates vector (5 rows)



Name	Type	Value
tail(HybridFinder_resu...	list [5]	List of length 5
sp|P10071|GLI3_HU...	character [1] (S3: SeqFastaA	'MEAQSHSSTTTEKKKVENSLVKCSTRTDVSEKAVASSTTSNEDESPGQTYHRERRNALTMQPQNVQGLSKVSEEPSTSS ...
sp|P32189|GLPK_H...	character [1] (S3: SeqFastaA	'MAASKKAVLGPLVGAVDQGTSSTRFLVFNSKTAELLSHHQVELKQEFPREGWVEQDPKELLHSVYECLEKTCEKLGQLN ...
sp|Q9NXW2|DJB12_...	character [1] (S3: SeqFastaA	'MESNKDEAERCLSLALKALQSNQPDRALRFLEKAQRLYPTPRVRALLESLNQKPQTAGDQPPPTDTTHATHRKAGGTDA ...
sp|denovo_HF_fake_...	character [1]	'KAVNLLLSYAKVNLLLSYKLADLFRLYNYGELFEKFDYGELFEKFDYGELFQKFKLADFLRLYYKPSPFFVFYNLPWLE ...
sp|denovo_HF_fake_...	character [1]	'LTMDLVQQLLWDLSLTRLKLADFRLLYKLANFRLLYYLNPSFTVLYKPSFPFVFKYPSMLFVFYKPSLMFVFKYPSLMF ...

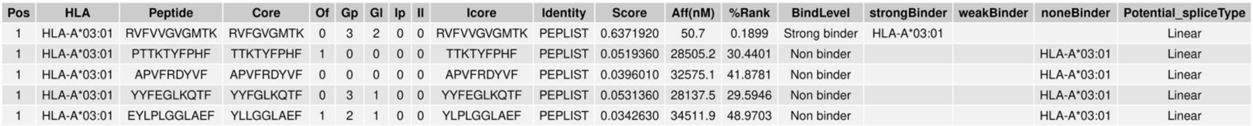
Figure 20. – Screenshot of the bottom of the HybridFinder merged hybrid proteome (5 proteins)

The results might differ if the customALCcutoff score parameter is changed. If the results are exported, these are stored in a folder as .csv files and the merged proteome database is saved as .fasta file. The peptide sequences predicted as spliced are considered as preliminary candidates. Performing the rest of the steps is essential in order to obtain the final list.

checknetMHCpan & step2_wo_netMHCpan

The checknetMHCpan & step2_wo_netMHCpan functions represent the last step in Faridi et al.'s (2018) workflow. After a database search is performed using the merged hybrid proteome in step 1, these two functions can be used. Both of these functions retrieve the potential splice type categorization established in step 1. However, with checknetMHCpan the user can directly obtain MHC-I binding affinity predictions computed for all peptides between 9 and 12 amino acids using netMHCpan (Jurtz et al., 2017; Reynisson et al., 2020).

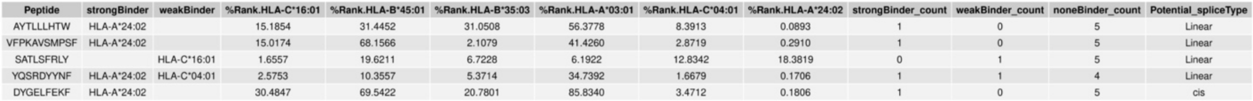
The checknetMHCpan function returns two formats of the netMHCpan results and the updated database search results from the second run with the potential splice type. The first format of the netMHCpan represents the results as they are (**Figure 21**). The second format is a tidied version of the netMHCpan results (**Figure 22**), where the rows are summarized into different columns, to allow quick analysis of the netMHCpan results (especially when more than one HLA-I allele is used); in these columns are summed the number of HLA-I alleles that a given peptide is a strong or weak binder to as well as the corresponding alleles. Finally, the database search results dataframe (from the second PEAKS run) updated with the potential splice type determined in the HybridFinder function for each peptide (**Figure 23**). Additionally, any sequence not identified in the hybridfinder output and solely attributed to the fake proteins created is removed. If exported, these are stored in a folder containing 2 .csv files and a .tsv (tab-separated values) corresponding to these different outputs.



Pos	HLA	Peptide	Core	Of	Gp	Gl	Ip	Il	Icore	Identity	Score	Aff(nM)	%Rank	BindLevel	strongBinder	weakBinder	noneBinder	Potential_spliceType
1	HLA-A*03:01	RVFVVGVGMTK	RVFGVGMTK	0	3	2	0	0	RVFVVGVGMTK	PEPLIST	0.6371920	50.7	0.1899	Strong binder	HLA-A*03:01			Linear
1	HLA-A*03:01	PTTKTYFPHF	TTKTYFPHF	1	0	0	0	0	TTKTYFPHF	PEPLIST	0.0519360	28505.2	30.4401	Non binder			HLA-A*03:01	Linear
1	HLA-A*03:01	APVFRDYVF	APVFRDYVF	0	0	0	0	0	APVFRDYVF	PEPLIST	0.0396010	32575.1	41.8781	Non binder			HLA-A*03:01	Linear
1	HLA-A*03:01	YYFEGLKQTF	YYFGLKQTF	0	3	1	0	0	YYFEGLKQTF	PEPLIST	0.0531360	28137.5	29.5946	Non binder			HLA-A*03:01	Linear
1	HLA-A*03:01	EYLPLGGLAEF	YLLGGLAEF	1	2	1	0	0	YLPLGGLAEF	PEPLIST	0.0342630	34511.9	48.9703	Non binder			HLA-A*03:01	Linear

Figure 21. – Screenshot of the checknetMHCpan netMHCpan results. (5 rows)

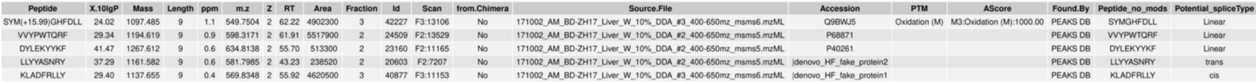
HLA/MHC is the allele, Peptide is the amino acid sequence of the potential ligand, Core is the minimal 9 amino acid sequence core to enable HLA binding, Of is the starting position of the Core within the peptide, Gp and Gl are the position and the length of the deletions (respectively), if any. Ip and Il are the position and the length of the insertions (respectively), if any. Icore is the interaction core, Identity is PEPLIST (which indicates that peptides were used as input as opposed to proteins in fasta-format). Score is the raw prediction, Aff(nM) is the predicted IC50 value in nanoMolar units, %Rank is the percentile rank of the predicted affinity compared to a set of random natural ligands. BindLevel is designated by 3 qualifiers: Strong binder, Weak binder, None binder. Potential_spliceType is the categorization retrieved from the HybridFinder output on the potential splice type explanation of the peptide (i.e linear, cis, trans).



Peptide	strongBinder	weakBinder	%Rank.HLA-C*16:01	%Rank.HLA-B*45:01	%Rank.HLA-B*35:03	%Rank.HLA-A*03:01	%Rank.HLA-C*04:01	%Rank.HLA-A*24:02	strongBinder_count	weakBinder_count	noneBinder_count	Potential_spliceType
AYTLLLHTW	HLA-A*24:02		15.1854	31.4452	31.0508	56.3778	8.3913	0.0893	1	0	5	Linear
VFPKAVSMPSF	HLA-A*24:02		15.0174	68.1566	2.1079	41.4260	2.8719	0.2910	1	0	5	Linear
SATLSFRLY		HLA-C*16:01	1.6557	19.6211	6.7228	6.1922	12.8342	18.3819	0	1	5	Linear
YQSRDYYNF	HLA-A*24:02	HLA-C*04:01	2.5753	10.3557	5.3714	34.7392	1.6679	0.1706	1	1	4	Linear
DYGELFEKF	HLA-A*24:02		30.4847	69.5422	20.7801	85.8340	3.4712	0.1806	1	0	5	cis

Figure 22. – Screenshot of the checknetMHCpan tidied netMHCpan results (5 rows)

Peptide is the amino acid sequence of the potential ligand, the strongBinder, weakBinder, noneBinder (this column not shown in this figure) columns correspond to the alleles to which a given peptide is a strong/weak/none binder to, respectively. If more than one allele, these are separated by commas. For each peptide, there will be %Rank columns per allele (eg. If 3 alleles were specified in the checknetMHCpan command, then each peptide will have 3%Rank columns). strongBinder_count, weakBinder_count, noneBinder_count represent the number of alleles to which a peptide is a strong/weak/none binder to. Lastly, the Potential_spliceType column is the categorization retrieved from the HybridFinder output on the potential splice type explanation of the peptide (i.e linear, cis, trans).

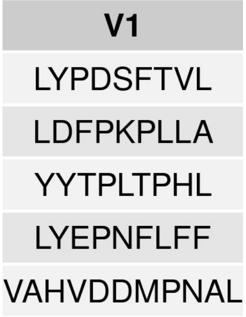


Peptide	X.10lgP	Mass	Length	ppm	m.z	Z	RT	Area	Fraction	Id	Scan	from.Chimera	Source.File	Accession	PTM	AScore	Found.By	Peptide_no_mods	Potential_spliceType
SYM(+15.99)GHFDLL	24.02	1097.485	9	1.1	549.7504	2	62.22	4902300	3	42227	F3:13106	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#3_400-650mz_msms6.mzML	Q9BWJ5	Oxidation (M)	M3:Oxidation (M):1000.00	PEAKS DB	SYMGHFDLL	Linear
VVYPWTQRF	29.34	1194.619	9	0.9	598.3171	2	61.91	5517900	2	24509	F2:13529	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#2_400-650mz_msms5.mzML	P68871			PEAKS DB	VVYPWTQRF	Linear
DYLEKYYKF	41.47	1267.612	9	0.6	634.8138	2	55.70	513300	2	23160	F2:11165	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#2_400-650mz_msms5.mzML	P40261			PEAKS DB	DYLEKYYKF	Linear
LLYYASNRY	37.29	1161.582	9	0.6	581.7985	2	43.23	238520	2	20603	F2:7207	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#2_400-650mz_msms5.mzML	|denovo_HF_fake_protein2			PEAKS DB	LLYYASNRY	trans
KLADFRLLY	29.40	1137.655	9	0.4	569.8348	2	55.92	4620500	3	40877	F3:11153	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#3_400-650mz_msms6.mzML	|denovo_HF_fake_protein1			PEAKS DB	KLADFRLLY	cis

Figure 23. – Screenshot of the checknetMHCpan database search results updated with the Potential_spliceType column (5 rows)

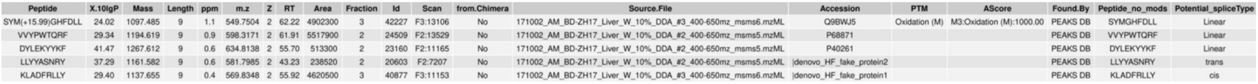
Peptide is the amino acid sequence of the potential ligand, X.log10P represents the best -10logP identification score for the corresponding peptide. Mass represents the monoisotopic mass of the peptide, Length is the number of amino acid residues that constitute the given peptide, ppm is the precursor mass error, the m.z is the precursor mass-to-charge ratio, Z is the precursor charge, RT is the Retention Time (elution time) for the spectrum, Area represents the area underthe curve of the peptide feature found at the same m/z and retention time as the MS/MS scan, Fraction is the LC-MS run, id represents the precursor ID associated with the PSM, Scan is a number representing a unique index for tandem mass spectra (F[Fraction#]:Scan#), from.Chimera (this column is not shown in this figure) displays whether the identified peptide is from chimeric spectra, Source.File is the mzML/mzXML file used in the PEAKS analysis, PTM is the type of the post-translational modification, Ascore is the localization score assigned to modifications on the peptide, Found.By represents the analysis (in this case PEAKS DB). Peptide_no_mods represents the peptide sequence without modifications, Potential_spliceType is linear, cis or trans and is retrieved from the HybridFinder function.

The step2_wo_netMHCpan is the equivalent of checknetMHCpan with the exception of computing binding affinity. The function returns a netMHCpan-ready list of peptides (**Figure 24**), as well as the updated the database search results (**Figure 25**). If exported, the results are exported into a folder containing a .pep file and a .csv file.



V1
LYPDSFTVL
LDFPKPLLA
YYTPLTPHL
LYEPNFLFF
VAHVDDMPNAL

Figure 24. – Screenshot of the step2_wo_netMHCpan netMHCpan-ready input (5 rows)



Peptide	X.10lgP	Mass	Length	ppm	m.z	Z	RT	Area	Fraction	Id	Scan	from.Chimera	Source.File	Accession	PTM	AScore	Found.By	Peptide_no_mods	Potential_spliceType
SYM(+15.99)GHFDLL	24.02	1097.485	9	1.1	549.7504	2	62.22	4902300	3	42227	F3:13106	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#3_400-650mz_msms6.mzML	Q9BWJ5	Oxidation (M)	M3:Oxidation (M):1000.00	PEAKS DB	SYMGHFDLL	Linear
VVYPWTQRF	29.34	1194.619	9	0.9	598.3171	2	61.91	5517900	2	24509	F2:13529	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#2_400-650mz_msms5.mzML	P68871			PEAKS DB	VVYPWTQRF	Linear
DYLEKYYKF	41.47	1267.612	9	0.6	634.8138	2	55.70	513300	2	23160	F2:11165	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#2_400-650mz_msms5.mzML	P40261			PEAKS DB	DYLEKYYKF	Linear
LLYYASNRY	37.29	1161.582	9	0.6	581.7985	2	43.23	238520	2	20603	F2:7207	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#2_400-650mz_msms5.mzML	|denovo_HF_fake_protein2			PEAKS DB	LLYYASNRY	trans
KLADFRLLY	29.40	1137.655	9	0.4	569.8348	2	55.92	4620500	3	40877	F3:11153	No	171002_AM_BD-ZH17_Liver_W_10%_DDA_#3_400-650mz_msms6.mzML	|denovo_HF_fake_protein1			PEAKS DB	KLADFRLLY	cis

Figure 25. – Screenshot of the checknetMHCpan database search results updated with the Potential_spliceType column. (5 rows)

The dataframe contains the same columns as in Figure 13.

After running checknetMHCpan or step2_wo_netMHCpan the final list of hybrid candidate peptides should be explored for further experimental validation (**Figure 9**).

Limitations

The presented package was developed and optimized for exports from PEAKS software. Therefore, results from other search engines or *de novo* sequencing softwares might not work while using this package. Limitations related to the workflow include the possible introduction of bias towards having results containing a higher proportion of Leucine residues. This is due to the workaround proposed by Faridi et al. (2018) which is also used in this package, entailing a switch of all Isoleucines to Leucines in the database search and the proteome since *de novo* sequencing does not differentiate between Isoleucine and leucine. As mentioned above, it is also important to emphasize that this protocol does not enable the direct identification of high-confidence PSPs that are genuinely spliced by the proteasome in vivo. However, this protocol enables the computational identification of putative PSPs, which should then be validated experimentally in a rigorous manner as shown in **Figure 9**.

Troubleshooting

Problem 1:

While installing the .tar.gz file for the package, in case you run into the following error: ”Error in install.packages : type == ”both” cannot be used with ’repos = NULL’”

Potential Solution:

The solution would be to simply invoke the install.packages function while specifying where the package is located, setting the repository (repos) to NULL and setting the type as source (source package).

```
> install.packages("~/Downloads/RHybridFinder_0.1.0.tar", repos = NULL, Type="source")
```

Problem 2:

While running HybridFinder, in case you run into the following error: ”Error in prepare_input_for_HF(de novo_candidates, db_search): Please make sure you have the right input. N.B: The *de novo* results data frame should be the first input”.

Potential Solution:

- Verify the *de novo* data frame has been correctly imported. Since the *de novo* results file is in .csv format, the separator should be a comma “,”, stringsAsFactors should be set to FALSE and lastly the header should be set to TRUE. Please refer to step1: Loading inputs into R.
- Verify that the HybridFinder parameters are properly typed. The *de novo* sequencing results data frame is indicated first and then the database search

results. Alternatively, write the parameters and assigned them their appropriate objects (i.e de novo_candidates = de novo_results_human_liver_Exp1). Please refer to step2: Run HybridFinder.

Problem 3:

While running HybridFinder, in case you run into the following error: ”Error in $<-.dataframe`(`*tmp`, “db_id”, value = character (0)) : replacement has 0 rows, data has[…]”

Potential Solution:

Verify the database search data frame, make sure it has been correctly imported. Since the database search results file is in .csv format the separator should be a comma “,”, stringsAsFactors should be set to FALSE and lastly the header should be set to TRUE. Please refer to step1: Loading inputs into R.

Problem 4:

While running checknetMHCpan, if the following error is displayed: ”Error in checknetMHCpan[…]:Please provide the proper input”

Potential Solution:

Verify the that the *de novo* and database search data frames are not switched. Please refer to step 5: Run checknetMHCpan.

Problem 5:

While running checknetMHCpan, if the following error is displayed: ”Please check the input alleles: […]”

Potential Solution:

Ensure that the alleles are in the right format, or that the allele is written correctly (i.e HLA-A03:01, HLA-A*03:01). Please refer to step 4: Load checknetMHCpan inputs into R.

Problem 6:

While running checknetMHCpan, if the path to netMHCpan is not correct, the following error might appear: sh: 1: [/temporary directory/netMHCpan]: not found error in running command

Potential Solution:

The issue could either be that the directory does not contains the netMHCpan file or that the directory was not well written I.e('usr/bin/' vs. '/usr/bin' or '/usr/bin/, where the first example is wrong and the other two are correct). Please refer to step 4: Load checknetMHCpan inputs into R

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Etienne Caron etienne.caron@umontreal.ca.

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The package is available on CRAN and includes data (PEAKS analyses) from HLA Ligand Atlas (Marcu et al., 2021) deposited in PRIDE (Proteomics IDentification Database) PXD019643 (were analyzed in PEAKS and used in this protocol for demonstration purposes only).

Acknowledgments

This work was supported by funding from the Funds de recherche du Québec - Santé (FRQS), the Cole Foundation, CHU Sainte-Justine and the Charles-Bruneau Foundations, Canada Foundation for Innovation, the National Sciences and Engineering Research Council (NSERC) (#RGPIN-2020-05232) and the Canadian Institutes of Health Research (CIHR) (#174924). KK is a recipient of IVADO's postdoctoral scholarship (#3879287150). CL is currently supported by a National Health and Medicine Research Council (NHMRC) of Australia CJ Martin Early Career Research Fellowship (1143366). AWP is supported by a NHMRC Principal Research fellowship (1137739). PF is supported by a Victorian Cancer Agency (Australia) Mid-Career Fellowship.

Declaration of Interests

The authors declared no conflicts of interest.

References

Admon, A. 2021. Are there indeed spliced peptides in the immunopeptidome?. Mol Cell Proteomics *20*, p.100099.

Berkers, C., Jong, A., Schuurman, K., Linnemann, C., Meiring, H., Janssen, L., Neefjes, J., Schumacher, T., Rodenko, B., and Ovaa, H. 2015. Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. J. Immunol. *195*, p.4085–4095.

Charif, D., and Lobry, J. 2007. Structural Approaches to Sequence Evolution, Molecules, Networks, Populations. Biological and Medical Physics, Biomedical Engineering, p.207–232.

Dalet, A., Robbins, P., Stroobant, V., Vigneron, N., Li, Y., El-Gamil, M., Hanada, K.i., Yang, J., Rosenberg, S., and Eynde, B. 2011. An antigenic peptide produced by reverse splicing and double asparagine deamidation. Proc. Natl. Acad. Sci. U S A *108*, p.E323–E331.

Ebstein, F., Textoris-Taube, K., Keller, C., Golnik, R., Vigneron, N., Eynde, B., Schuler-Thurner, B., Schadendorf, D., Lorenz, F., Uckert, W., Urban, S., Lehmann, A., Albrecht-Koepke, N., Janek, K., Henklein, P., Niewienda, A., Kloetzel, P., and Mishto, M. 2016. Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. Scientific Reports *6*, p.24032.

Faridi, P., Li, C., Ramarathinam, S., Vivian, J., Illing, P., Mifsud, N., Ayala, R., Song, J., Gearing, L., Hertzog, P., Ternette, N., Rossjohn, J., Croft, N., and Purcell, A. 2018. A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. Sci. Immunol. *3*, p.eaar3947.

Faridi, P., Woods, K., Ostrouska, S., Deceneux, C., Aranha, R., Duscharla, D., Wong, S., Chen, W., Ramarathinam, S., Sian, T., Croft, N., Li, C., Ayala, R., Cebon, J., Purcell, A., Schittenhelm, R., and Behren, A. 2020. Spliced Peptides and Cytokine-Driven Changes in the Immunopeptidome of Melanoma. Cancer Immunol. Res. *8*, p.1322–1334.

Hanada, K.i., Yewdell, J., and Yang, J. 2004. Immune recognition of a human renal cancer antigen through post-translational protein splicing. Nature *427*, p.252–256.

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*, p.3360–3368.

Lichti, C. 2021. Identification of spliced peptides in pancreatic islets uncovers errors leading to false assignments. Proteomics *21*, p.e2000176.

Liepe, J., Marino, F., Sidney, J., Jeko, A., Bunting, D., Sette, A., Kloetzel, P., Stumpf, M., Heck, A., and Mishto, M. 2016. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science *354*, p.354–358.

Liepe, J., Mishto, M., Textoris-Taube, K., Janek, K., Keller, C., Henklein, P., Kloetzel, P., and Zaikin, A. 2010. The 20S Proteasome Splicing Activity Discovered by SpliceMet. PLoS Comput. Biol. *6*, p.e1000830.

Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D., Freudenmann, L., Backert, L., Mühlenbruch, L., Szolek, A., Lübke, M., Wagner, P., Engler, T., Matovina, S., Wang, J., Hauri-Hohl, M., Martin, R., Kapolou, K., Walz, J., Velz, J., Moch, H., Regli, L., Silginer, M., Weller, M., Löffler, M., Erhard, F., Schlosser, A., Kohlbacher, O., Stevanović, S., Rammensee, H.G., and Neidert, M. 2021. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. J. Immunother. Cancer *9*, p.e002071.

Michaux, A., Larrieu, P., Stroobant, V., Fonteneau, J.F., Jotereau, F., Eynde, B., Moreau-Aubry, A., and Vigneron, N. 2014. A Spliced Antigenic Peptide Comprising a Single Spliced Amino Acid Is Produced in the Proteasome by Reverse Splicing of a Longer Peptide Fragment followed by Trimming. J. Immunol. *192*, p.1962–1971.

Mishto, M., Goede, A., Taube, K., Keller, C., Janek, K., Henklein, P., Niewienda, A., Kloss, A., Gohlke, S., Dahlmann, B., Enenkel, C., and Kloetzel, P. 2012. Driving Forces of Proteasome-catalyzed Peptide Splicing in Yeast and Humans. Mol. Cell. Proteomics *11*, p.1008–1023.

Mishto, M. 2020. What We See, What We Do Not See, and What We Do Not Want to See in HLA Class I Immunopeptidomes. Proteomics *20*, p.2000112.

Mylonas, R., Beer, I., Iseli, C., Chong, C., Pak, H.S., Gfeller, D., Coukos, G., Xenarios, I., Müller, M., and Bassani-Sternberg, M. 2018. Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome. *Mol. Cell.* Proteomics *17*, p.2347–2357.

Neefjes, J., Jongsma, M., Paul, P., and Bakke, O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology, 11*(12), p.823–836. Purcell, A. 2021. Is the Immunopeptidome Getting Darker?: A Commentary on the Discussion around Mishto et al., 2019. Frontiers in Immunology *12*, p.720811.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL: http://www.rstudio.com/.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, p.W449–W454

Rolfs, Z., Solntsev, S., Shortreed, M., Frey, B., and Smith, L. 2018. Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion. J. Proteome Res. *18*, p.349–358.

Rowley, D., and Fitch, F. 2012. The road to the discovery of dendritic cells, a tribute to Ralph Steinman. Cell. Immunol. *273*, p.95–98.

Specht, G., Roetschke, H., Mansurkhodzhaev, A., Henklein, P., Textoris-Taube, K., Urlaub, H., Mishto, M., and Liepe, J. 2020. Large database for the analysis and prediction of spliced and non-spliced peptide generation by proteasomes. Scientific Data *7*, p.146.

Steinman, R., and Witmer, M. 1978. Lymphoid dendritic cells are potent stimulators of the primary mixed leukocyte reaction in mice. Proc. Natl. Acad. Sci. U S A *75*, p.5132–5136.

Wilhelm, M., Zolg, D., Graber, M., Gessulat, S., Schmidt, T., Schnatbaum, K., Schwencke-Westphal, C., Seifert, P., Krätzig, N., Zerweck, J., Knaute, T., Bräunlein, E., Samaras, P., Lautenbacher, L., Klaeger, S., Wenschuh, H., Rad, R., Delanghe, B., Huhmer, A., Carr, S., Clauser, K., Krackhardt, A., Reimer, U., and Kuster, B. 2021. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. Nat. Commun. *12*, p.3346.

Chapter 3: Materials and methods

Next generation sequencing as well as the development of various bio-informatic tools, annotation datasets, etc. have tremendously advanced the identification of gene fusions, especially on the genomic level. On the immunopeptidomic level, gene fusion-derived peptides are a promising class of tumour specific neoantigens. It is possible to identify immunologically relevant peptides derived from a chimaeric protein given the predicted polypeptide sequence, as well as assess the heterogeneity of the HLA-I binding of these gene fusion-derived peptides.

3.1. Develop FusionChoppeR

In order to investigate the antigenic potential of gene fusions, we have developed FusionChoppeR. FusionChoppeR allows the prediction of potential HLA-I peptides (peptides of 8-12 amino acids) derived from gene fusion proteins (spanning the junction region exclusively) by predicting HLA-I binding affinity of fusion junction peptides against 35 HLA-I alleles (Fig. 27) covering all the groups (or supertypes) (Annex Table1).

3.1.1. In silico peptide-HLA-I binding predictions

In order to predict peptide-HLA-I binding, two algorithms were used: netMHCpan and MHCflurry (Fig. 26). While netMHCpan is often used in the immunopeptidomic community, the last benchmark performed of peptide-MHC-I prediction algorithm has shown the powerful performance of mhcflurry (Zhao and Sher, 2018). NetMHCpan is modeled through an artificial neural Network (ANN) and uses the NNAlign_MA algorithm in order to establish the peptide-MHCI binding specificities from the integration of single-and multi-allele data(Alvarez et al., 2019). On the other hand, MHCflurry has been trained through a two-step process: an allele-dependent and an allele-independent one. The allele-dependent step predicts the binding affinity of peptide-MHC-I through a neural network ensemble. Then, a predictor trained to distinguish between hits and decoy peptides in MS data and is meant to determine cleavability (especially if N- and C-terminal flanks are provided). Finally, a logistic regression model is trained on the outputs from both steps and generates a presentation score between 0 and 1(O'Donnell et al., 2020). In conclusion, the utilization of peptide-MHC-I predictions from both of these tools could boost confidence in the predictions obtained.

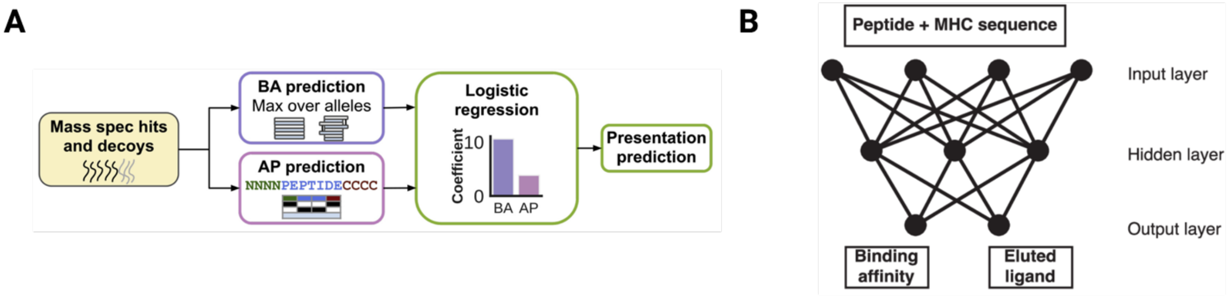


Figure 26. – Architectural structure of MHCflurry and netMHCpan.

Panel A represents the architecture of MHCflurry. Figure adapted from (O'Donnell et al., 2020). Panel B represents netMCpan architecture. Figure adapted from (Jurtz et al., 2017)

3.1.2. Prioritization of gene fusion-derived neoantigens

NetMHCpan and MHCflurry generate a percentile rank that corresponds to a value between 0 and 100. This corresponds to the rank of the predicted binding score compared to a set of random natural peptides The developers of netMHCpan suggest a percentile rank threshold of 0.5, to identify strong peptide-MHC-I binders and 2.0 for weak binders. Integrate-Neo which utilizes netMHCpan applies a stringent threshold of 0.5 to select peptide-MHC-I binders (Zhang et al., 2017). By applying a low threshold (such as 0.5), it is inevitable that there will be potential false negatives (which would lead to missing out on potential binders) and in turn, by increasing the threshold the introduction of false positives is also inevitable (Annex Table 3). Therefore, FusionChoppeR sets a threshold of 2.0. However, it applies it while integrating the results of two peptide-MHC-I binding prediction tools: netMHCpan and MHCflurry. Thus, peptide-MHC-I pairs are considered by FusionChoppeR (FCR) if they have a percentile rank less or equal to 2.0 by both tools. This ensures to balance out the false predictions while boosting confidence in the predicted binders.

3.1.2.1. General scores

To compare and select peptides based on their predicted binding of HLA-I alleles, we have calculated 2 scores:

- Strong binder ratio: each peptide-MHCI is predicted by both algorithms, which means that based on the percentile threshold (see equation below & Annex Table 4, SBO: Strong binder by at least One algorithm, SB: Strong Binder, WB: Weak Binder) generated by each of the algorithms an assignment “SB”, ‘WB” is obtained. Since we only consider a peptide as binder to an HLA allele if it got a percentile by both algorithms that does not exceed 2.0, the results can be summarized in 3 different scenarios: 1) either both algorithms agree that a peptide is SB to the HLA, or 2) only one of them, or 3) both algorithms agree that a peptide is WB.

$$SBratio = \frac{\sum \mathrm{SBO}}{\sum SB + \sum WB}$$

- Local Contribution to Beta Diversity (LCBD). Considering the multivariate nature of the analysis, the LCBD was employed to detect the peptides derived from a fusion that create heterogeneity in the HLA-I molecules predicted to be bound. LCBD is defined in ecological terms as the contribution of a site to the beta diversity (Borcard et al., 2018). Beta diversity was originally developed to assess diversity while working with multivariate data (as ecological data usually are, for example assessing the compositional diversity in species across sites and the contribution of different sites to it). In the case of gene fusions, because from each gene fusion we generate 45 peptides and evaluate the binding potential of each one of these to various HLA-I alleles, we obtain a matrix of 45 rows by 35 columns, where the ‘sites’ are the junction peptides and the species (columns) is the set of HLA-I alleles. In a given matrix where *i* is a given row, *j* is a given column, *p* is the total number of columns and *n* is the total number of sites:

$$LCBD_i = \frac{SS_i}{SS_{total}}$$

Where SS_i is the total sum of squares of the row *i* (sum of the centred and squared values for row *i*),

$$SS_i = \sum_{j=1}^{p} s_{ij}$$

And the SS_{total} is total sum of squares of the entire matrix

$$SS_{total} = \sum_{i=1}^{n} SS_i$$

The LCBD value is calculated with the package adespatial (Dray et al., 2012), permutation tests followed by adjustment of p value. The null hypothesis of the permutation tests posits that the abundance is preserved and that the species are distributed at random. Therefore, a p adjusted <0.05 (assuming a 95% probability of non-randomness) leads to the rejection of the above-mentioned null hypothesis.

3.1.2.2. Score for the extraction of the richest non-diverse peptide

The extraction of the richest non-diverse peptide for section 4.3.2 was based on the proportion of categorizations for all HLAs for each peptide by both algorithms (SBO) multiplied by the number of supertypes predicted to cover. This score is meant to give a relative quality of the peptide in binding all the alleles. (BNDP, Best non-diverse peptide)

$$BNDP = SBO * supertypes\_covered$$

3.1.3. FusionChoppeR workflow

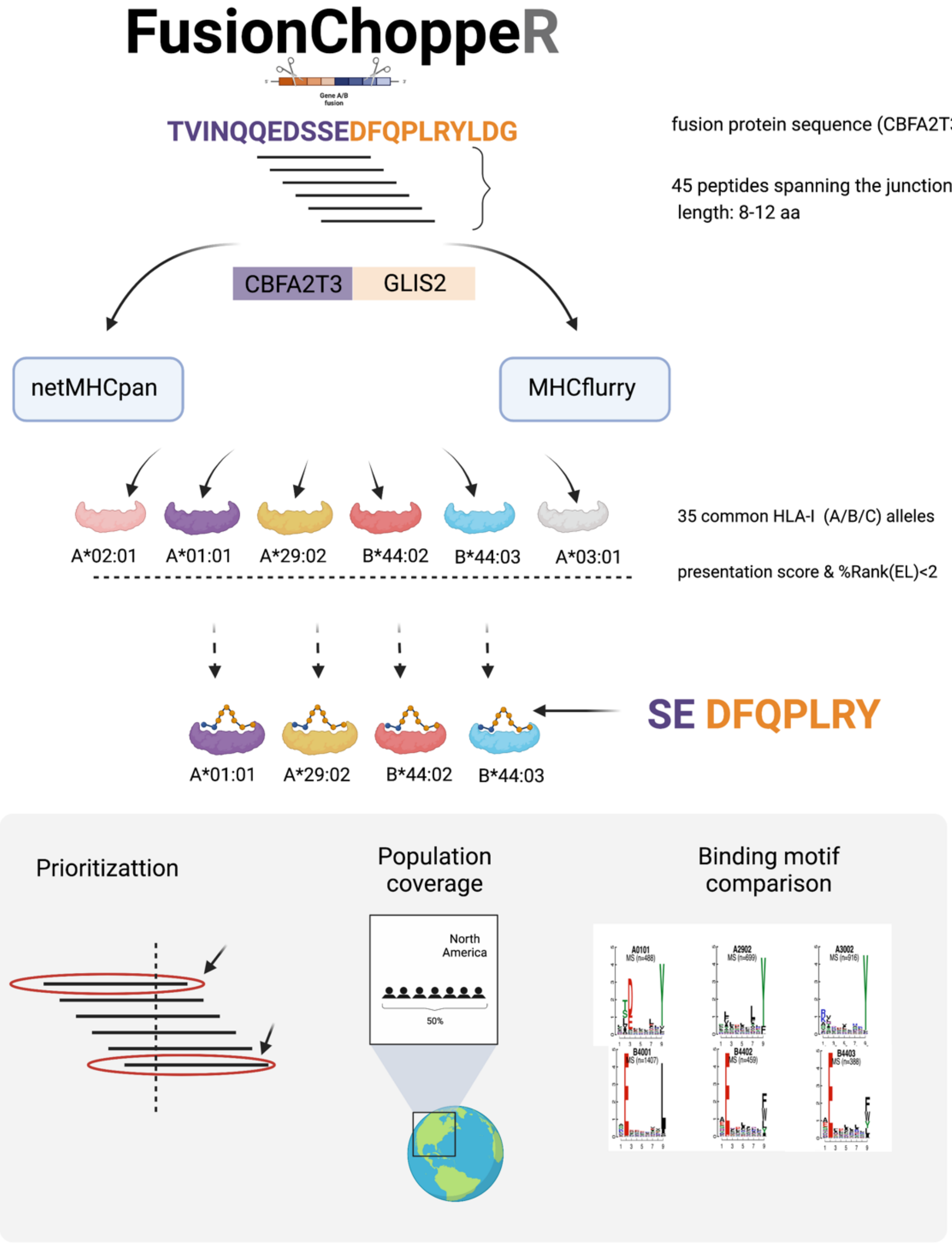


Figure 27. – The workflow of FusionChoppeR.

A fusion protein sequence is 'chopped up' into 45 peptides of 8 to 12 amino acids spanning the junction region. Then, MHC-I binding predictions are computed on pairs of each of these peptides with 35 of the most common HLA-I alleles. Applying a common threshold on both algorithms, if both algorithms agree that a peptide would bind an HLA-I molecule, the peptide-HLA pair makes it to the next step. Finally, prioritization is applied to determine the best candidate(s), their population coverage for different *populations is computed, and the appropriate motifs from MS experiments are shown to aid with validation (Sarkizova et al., 2020a)*

3.1. Database of gene fusions in cancer cell lines

With the purpose of understanding whether gene fusion-derived peptides would be predicted bind multiple HLA-I alleles, we aimed to apply FCR on a large set of gene fusions. We decided to use a database of gene fusions identified in cancer cell lines. The LiGeA database is the collection of gene fusions identified from paired-end RNA-seq data of 935 cell lines of the Cancer Cell Line Encyclopedia (CCLE) (Gioiosa et al., 2018). The database also includes the translated protein sequence of gene fusions as well as different annotations relating to the matching of the fusion in other databases. We have used the output of FusionCatcher. In order to limit our prospective analyses to fusions not found in healthy individuals we have removed all fusions also found in GTEX (Genotype Tissue Expression). Additionally, the remaining fusions were queried in fusion hub (Panigrahi et al., 2018) and those matched with further fusions found in healthy individuals, were removed as well as those in the newer version of the banned list of fusions from the 'generate_banned.py' script in fusionCatcher(Nicorici et al., 2014) on the GitHub repository. Next, only gene fusions with a translated protein were kept. Of note, we chose not to use the number of algorithms that have detected the fusion as a filter, as we found that certain fusions identified by only one algorithm are validated and known to exist in these cell lines.

To analyse the predicted population and HLA-I allele coverage of gene fusions that are clinically relevant. The output from FusionCatcher includes an annotation column that has information on where else the fusion is matched as well as whether it involves an oncogene. We used this information to filter the database for gene fusions that 1) involve oncogenes and 2) were matched with those found in COSMIC (Catalogue Of Somatic Mutations in Cancer) database. COSMIC database is manually curated primary tumour data on gene fusions identified in different samples. Therefore, this filtering would allow us to explore the predicted HLA-I binding and population coverage of gene fusions in a clinically relevant context.

Finally, FCR predictions were being performed on 35 HLA-I alleles (that cover all supertypes), therefore we decided to restrict the predictions to each cell line's encoded HLA-I molecules. This would help us explore the predicted gene fusion-derived peptides in a more realistic setting where HLA-I molecules encoded are 3 up to 6 per individual. To do that, we found and retrieved from the TRON (Scholtalbers et al., 2015) web portal the HLA-I typings of only 70 out of 392 cell lines (not all cell lines were found).

3.2. Population Coverage analysis

HLA-I frequencies of different populations (collected from ~66800 individuals) were obtained from public datasets covering 5 populations and an overall world population (Solberg et al., 2008; Weingarten-Gabbay et al., 2021), the calculation of the Cumulative Phenotypic Frequency (CPF) was done with Hardy-Weinberg's assumption of proportions of the HLA genotypes (Dawson et al., 2001):

$$CPF = 1 - \left(1 - \sum_{i \in H} p_i\right)^2$$

H corresponds to the HLA group (A/B/C) which consists of different HLA-I alleles and p_i represents the frequency of the i alleles belonging to the HLA group set (in a given population). The calculation of the CPF for a given peptide predicted to bind multiple HLA-A, -B & C alleles, follows (Poran et al., 2020; Weingarten-Gabbay et al., 2021):

$$CPF = 1 - \left(1 - \sum_{i \in A} p_i\right)^2 \times 1 - \left(1 - \sum_{j \in B} p_j\right)^2 \times 1 - \left(1 - \sum_{k \in C} p_k\right)^2$$

3.3. Immunopeptidomic data of mono-allelic cell lines

In order to analyze the existence of peptides binding multiple HLA-I alleles we have taken advantage of a recently published dataset of HLA-I immunopeptidomes of 95 HLA-I alleles (HLA-A/B/C/E/G) (Sarkizova et al., 2020a). The number of alleles considered in the experiment expands the HLA-I immunopeptidomes profiled to date. In this thesis, we have excluded results from the HLA-E & -G alleles. The dataset we have used is published by the authors and has already been filtered for contaminants of different sources (other organisms and negative controls), tryptic peptides and peptides appearing in more than 20 alleles.

Additionally, because we were interested in understanding whether our selected candidates could justifiably bind/elute with multiple HLA-I alleles. To do that, we have applied an existing protocol for clustering peptides based on sequence similarity(Venema et al., 2021). Firstly, we extracted a subset of this large dataset containing only 9-mer peptides that bound the alleles predicted and/or tested against experimentally. Of note, is that all these peptides were single HLA-I binders (i.e. they did not appear in the immunopeptidome of more than allele). To this, we added our candidate peptides. Then, peptide sequence distances were calculated using the distPMBEC matrix for MHC-I peptides (Kim et al., 2009) and accounting for molecular entropy. Then, non-metric dimensional scaling (NMDS) was computed to reduce the dimensionality of our data into two. Next, we clustered the 2-dimensional data (the workflow is depicted in the results section, Fig. xfw) using two methodologies: k-means and density-based scanning. To determine the ideal number of clusters, k-mean simulations of partitions of 2 to 25 groups (cascadeKM function from the vegan package in R) were run. Then, the ideal number of groups was based on the partition that optimizes the Calinski-Harabasz criterion. We also used density-based scanning (dbscan from fpc package in R) in order to compare results and understand whether this methodology has classified any peptides as noise. Lastly, to 1) understand the composition of the clusters in which our peptides were grouped, we plot the logo of the peptides within these clusters (ggseqlogo from ggseqlogo package in R). And to 2) determine which HLA-I alleles they would bind based on sequence similarity we used KNN algorithm. We retrieved which HLA-I alleles the 50-closest peptides were bound to in the original immunopeptidome dataset.

3.4. Gene fusion sequences

The sequences of the gene fusions KIAA1549-BRAF, GIT2-BRAF, BCR-ABL1, CENPC-ABL1, ATF7IP-PDGFRB, ZNF274-JAK2, NUP153-ABL1, SPTAN1-ABL1, NOTCH1-ROS1, GOLGA5-JAK2 gene fusions were derived from the sequencing results reported in the Children's Hospital Of Philadelphia fusion panel (Chang et al., 2019a). On the other hand, the CBFA2T3-GLIS2 was obtained from the LIGEA database (Annex Table 2).

3.5. In vitro peptide-HLA-I binding measurement

Binding affinity was performed by our collaborators at La Jolla Institute for Allergy (Dr. Alessanndro Sette & Dr. Sidney and lab). The methodology employed by Dr. Sidney and Dr. Sette is the standard for peptide-MHC-I binding measurement. The assay is a competitive inhibition of binding assay. The binding of peptide to an MHC-I molecule is assessed through its ability to inhibit the binding of a radiolabeled probe peptide to the MHC-I molecule. The IC50, or the concentration of the peptide needed to inhibit the binding of the labeled peptide can be deduced from plotting the concentrations and inhibition (%). The IC50 values constitute reasonable approximations of true K_d (Sidney et al., 2013). The thresholds for peptide-MHC-I binding are defined as 500nM for strong binders and 2000nM for weak binders(Alessandro Sette et al., 1994).

3.6. 3D in silico peptide-MHC I prediction and visualization

Recently, a tool (RosettaMHC/ MHCpepthreader) for the 3D structure in silico prediction of the conformation of the peptide-HLA-I complex has been published in the context SARS-COV2-derived antigens bound to HLA-A2 (Nerli and Sgourakis, 2020). It has also been used in the context of pediatric neuroblastoma antigens presented on different HLA-I alleles (Yarmarkovich et al., 2021). The tool helped to understand how the conformation of the peptide on the different HLA-I alleles affected the result of their peptide-centric immunotherapy. RosettaMHC/ MHCpepthreader is based on homology modeling and utilizes the python wrapper for Rosetta (Chaudhury et al., 2010). In other words, the tool is fed a crystal structure of a peptide-HLA-I complex which it uses as template, in order to predict the 3D structural conformation of target peptide sequence within the HLA-I binding grove. First, the PDB structure is treated to only keep the chain containing the α1 and α2 domains (181 residues) and forces ideal bond lengths and angles according to Rosetta Methods. Then, the template and target peptide-HLA-I sequences are aligned using ClustalOmega. Afterwards, the target peptide is “threaded” onto the HLA-I structured which then goes through a “relax” procedure in order to optimize the structure. Finally, binding affinity of peptide-HLA-I is calculated using the InterfaceAnalyzer protocol. Here, the selection of the 3D templates was on the methodology described by the original authors (Nerli and Sgourakis, 2020) based on the peptide and the HLA-I allele (Fig. 26). After running the simulation, a PDB structure is generated and a binding energy is computed. This structure is then visualized and analyzed for peptide-MHC polar interactions using ChimeraX software(Pettersen et al., 2020)

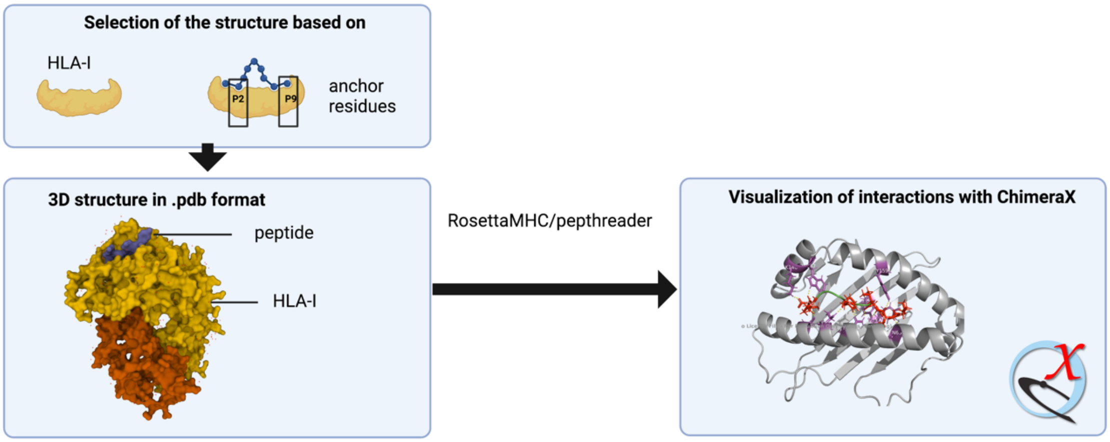

Selection of the structure based on
HLA-I
P2
P9
anchor residues
3D structure in .pdb format
peptide
HLA-I
RosettaMHC/pepthreader
Visualization of interactions with ChimeraX

Figure 28. – The process followed for the 3D in silico modeling.

The modeling starts with ensuring the selection of the most appropriate template(s). To do that, I have archived crystal structures in the Protein data bank (Burley et al., 2018), then based on the peptide-HLA-I pairs, the most appropriate PDB structure is chosen based on the criteria defined in the original article of RosettaMHC (Nerli and Sgourakis, 2020). Once the model is chosen, the simulation is run in RosettaMHC and then the structure is visualized in ChimeraX.

3.7. Other

3.7.1. Statistics

All statistical computations and graphs were performed and created using R programming language R v. 4.1.3 on R studio v.1.4.1717 on Mac OS.

R packages: vegan, adespatial, shiny, reactable, htmltools, plotly, ggpubr, ggplot2.

3.7.2. Figures

Figures were created using the online version of Biorender (https://www.biorender.com/).

3.7.3. Code Availability

Zenodo download: https://doi.org/10.5281/zenodo.7049144.

Chapter 4: Results

4. Applying FusionChoppeR to explore the neoantigenic potential of gene fusions

4.1. Immunogenic gene fusions

Firstly, we wanted to explore whether we would be able to predict potential peptides of 8-12 amino acids containing the fusion junction and with predicted HLA binding affinity using FusionChoppeR. We only considered peptides for which immunogenic potential has been published previously. Therefore, we only considered the ETV6-RUNX1 and CBFB-MYH11 fusions (Biernacki et al., 2020a; Yotnda, Garcia, et al., 1998; Zamora et al., 2019a). We used polypeptide sequences of these with FusionChoppeR and ran them against our list of 35 HLA-I alleles. From ETV6-RUNX1, two peptide sequences were selected MPIGRIAEC and RIAECILGM and REEMEVEHEL from CBFB-MYH11. These were the exact same peptides tested in the literature (Fig. 29-30).

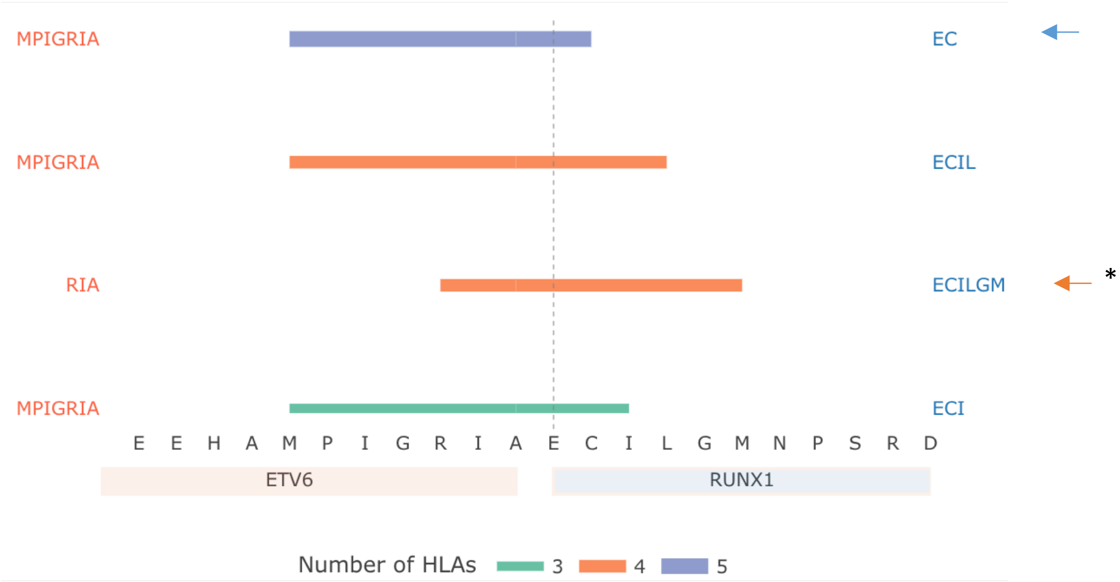


Figure 29. – Analysis of ETV6-RUNX1 using FusionChoppeR GUI.

The x-axis displays the amino acids around the junction region, on the y-axis the sequences of the fusion peptides are displayed – on the left are the amino acids from ETV6 (protein 1), on the right are the ones from RUNX1 (protein 2). The horizontal bars represent the sequences of the peptides with relation to the x-axis. The color and thickness of these horizontal bars is related to the number HLA-I molecules the peptides are predicted to bind by both algorithms. The dashed line represents the junction. The arrows were added in this document. The arrows represent peptides derived from ETV6-RUNX1 that were tested in the literature. The red arrow with an asterisk designates the diverse peptide that FCR has generated a significant LCBD (p adj<0.05) for.

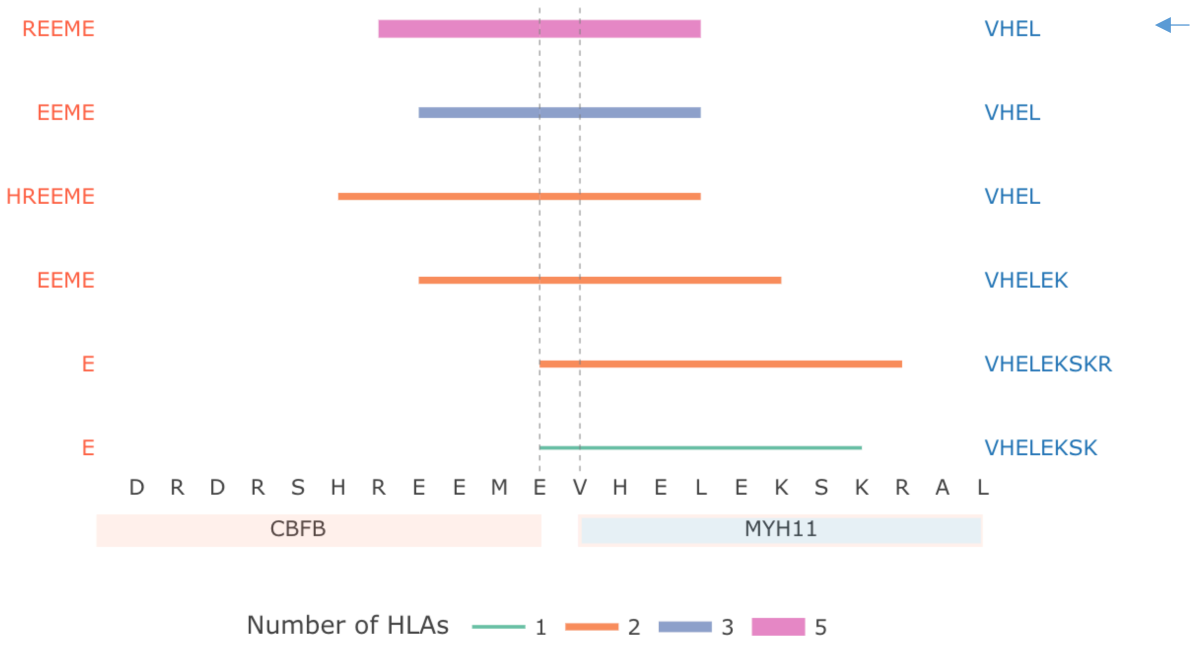


Figure 30. – Analysis of CBFB-MYH11 using FusionChoppeR GUI.

The x-axis displays the amino acids around the junction region, on the y-axis the sequences of the fusion peptides are displayed – on the left are the amino acids from CBFB (protein 1), on the right are the ones from MYH11 (protein 2). The horizontal bars represent the sequences of the peptides with relation to the x-axis. The color and thickness of these horizontal bars is related to the number HLA-I molecules the peptides are predicted to bind by both algorithms. The dashed lines represent the junction. The arrow was added in this document.

The FCR analyses (Fig.29-30) for ETV6-RUNX1 and CBFB-MYH11, show that are predicted to generate 4 and 6 HLA-I peptides, respectively. The peptides predicted to bind the most of HLA-I alleles (MPIGRIAEC, REEMEVHEL) as well as a peptide with a significant LCBD (p adj<0.05) (RIAECILGM) (LCBD, explained in section the methods section 3.1.2.1 and results section 4.3.2.2) are peptides for which immunogenic potential has been published previously peptides.

4.2. Prepare the gene fusions database

To test FCR on a larger dataset of gene fusions, we used the LiGeA database of gene fusions identified in cancer cell lines (Gioiosa et al., 2018). The database contains the results of gene fusion calling on 935 paired-end RNA seq experiments. For each cell line different fusions are detected and their transcript sequence is included as well as whether the tool (FusionCatcher (Nicorici et al., 2014)) predicts the fusion to be a True or False positive, the predicted effect of the fusion (in-frame or out-of-frame or intronic) and other metadata (whether it involves an oncogene, or it has been reported in other databases). First, gene fusions identified in healthy tissues were removed (found in GTEX and datasets published by different authors, FusionCatcher's updated gene fusion list) in order to focus on those that are more prone to be oncogenic (Oliver et al., 2020). Then, only the fusions that are predicted to be translated into a protein were kept. Finally, to add robustness to our analyses we applied a stringent approach by removing gene fusions predicted as False positives as well as a polyQ polypeptide derived from a reciprocal gene fusion found across >10 cell lines (Fig. 31). At the end of our processing of the database, we had 523 fusions spanning across 337 cell lines and 20 different cancer types.

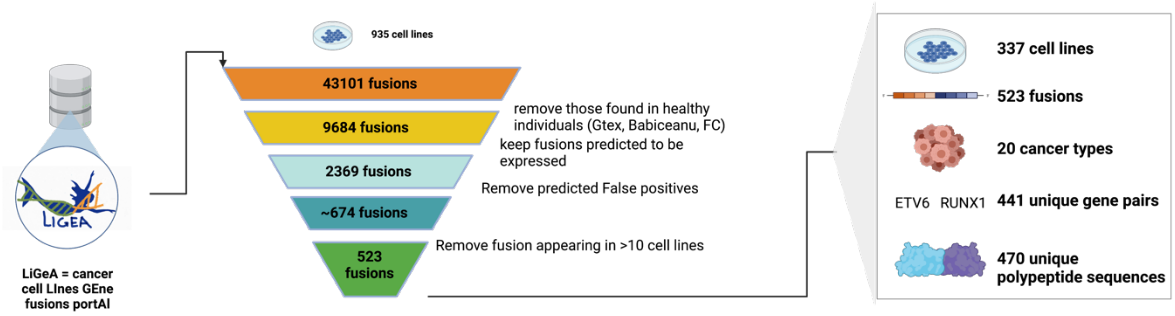

935 cell lines
43101 fusions
9684 fusions
2369 fusions
~674 fusions
523 fusions
remove those found in healthy individuals (Gtex, Babiceanu, FC)
keep fusions predicted to be expressed
Remove predicted False positives
Remove fusion appearing in >10 cell lines
LiGeA = cancer cell LInes GEne fusions portAl
LIGEA
337 cell lines
523 fusions
20 cancer types
ETV6 RUNX1
441 unique gene pairs
470 unique polypeptide sequences

Figure 31. – Processing LiGeA Cancer cell line database.

To evaluate the predicted HLA-I binding of gene fusions, the database was curated by removing fusions identified in healthy individuals and removing potential false positive identifications and keep protein expressing gene fusions. Additional filtering has led to the removal of EP400#NCOR2 and NCOR2#EP400 fusions which are present in many cell lines. Finally, after the processing the database included 523 fusions (of which 441 are unique) found in 337 cell lines and spanning 20 cancer types. That is, the total number of fusions left is 523, however, because certain gene fusions could exist in more than one cell line (i.e EWSR1-FLI1, see Fig. 32), we distinguished between this number and the number of unique pairs (i.e EWSR1 + FLI1 is 1 unique pair, ETV6 + RUNX1 is another unique pair). Furthermore, because gene fusions could be rearranged differently, we also accounted for the fact that this could create distinct fusion polypeptide sequence which could alter the predicted HLA-I peptides, therefore a fusion can have 1 or more unique polypeptides (i.e EWSR1-FLI1, see Fig. 32). Lastly, because each cell line is associated with a disease, we also enumerate the cancer types covered.

Certain fusions were identified were identified in more than one cell line with the same or different translated polypeptide sequence, thus a lower number of unique gene pairs totalling 441. To explain the differences in the numbers in Fig. 31, we illustrate this with EWSR1-FLI1 fusion from the database (Fig.32). EWSR1-FLI1 fusion which is essential for the malignant transformation in Ewing's Sarcoma is present in 4 sarcoma cell lines, therefore these are counted as 4 fusions. Despite that, it is still considered as being 1 gene pair (EWSR1 and FLI1). Moreover, for the peptide-MHC-I predictions, the polypeptide sequence around the junction derived from the EWSR1-FLI1 fusion is important. In the case of EWSR1-FLI1, the same sequence is found in 3 out of the 4 cell lines which accounts for two unique (different) polypeptide sequences (Fig 32).

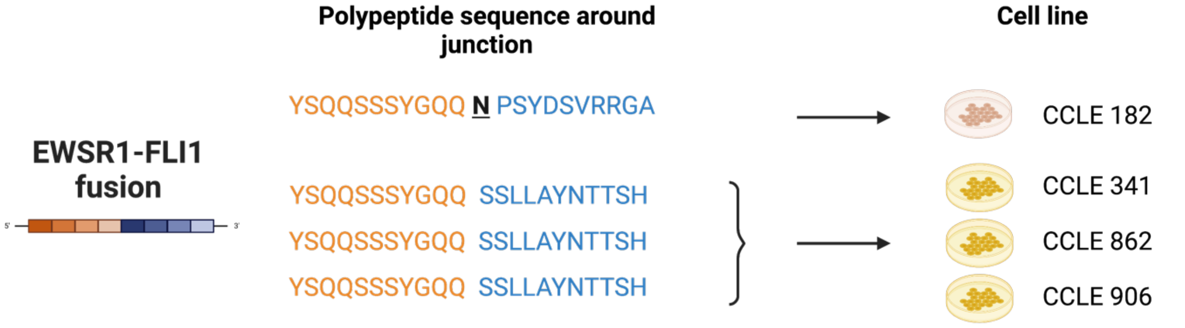


Figure 32. – Example of fusions present in more than one cell line.

The example shows a gene fusion EWSR1-FLI1 present in 4 sarcoma cell lines (CCLE 182, 341, 862, 906). The example also outlines the fact that certain fusions can have a different sequence based on the breakpoint. However, as can also be seen the same arrangement is present in 3 out of 4 cell lines.

4.3. The landscape of predicted gene fusion-derived neoantigens in LiGeA

From each gene fusion protein, FCR generated through a sliding window approach 45 peptides of 8 to 12 amino acids, the range of peptide length capable of binding HLA-I (Rock et al., 2016). The gene fusions considered span 20 different cancer types (Fig. 33A) and are predicted to generate mainly in-frame (85.53%) fusion proteins. Around 13% of the gene fusions are predicted to generate out-of-frame fusion proteins (Fig. 33B). Predictions were then computed in FCR for each of the peptides with 35 of the most common HLA-I alleles covering all HLA-I supertypes (Annex, Table 1**)**. This process was iteratively done for each fusion protein sequence.

Then, we were interested in exploring the cross-HLA-I binding potential of these gene fusion-derived neoantigens, we found that 1 up to 3 out of 35 alleles predicted to be bound represented the largest proportion of peptides (78.26%). However, there were peptides predicted to bind more than 3 alleles (~21%). Remarkably, one peptide (RVKPPWMAF, from RUNX1#PRDM7 fusion) was predicted to bind 25 out of the 35 HLA-I alleles (Fig. 33D).

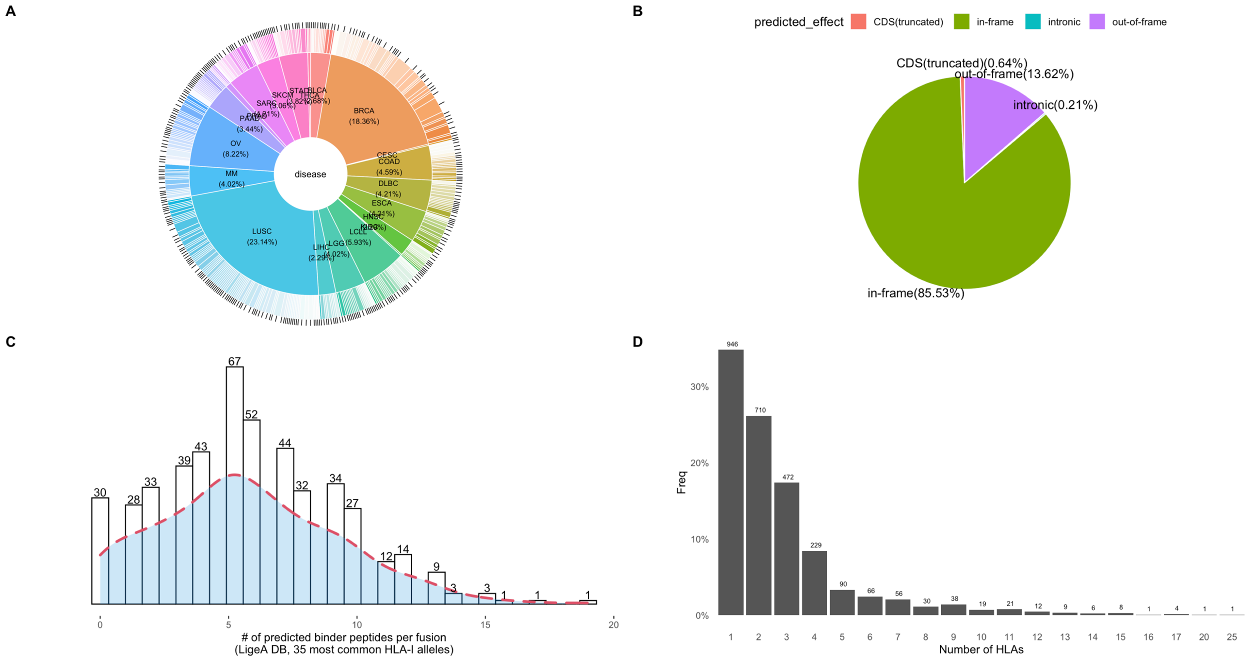


Figure 33. – The database statistics and the antigenic landscape of the gene fusion-derived peptides.

The donut chart in A) displays the proportion of the diseases (associated with the cancer cell lines) observed in the database. The predicted effect pie chart in B) is representative of the results of FusionCatcher's prediction of whether the gene fusion-derived peptide conserves the reading frame or not. C) shows the amount of predicted binder peptides per fusion whereas D) is a histogram of the number of HLA-I alleles predicted to be bound per fusion peptide. BRCA : Breast invasive carcinoma, SARC : Sarcoma, SKCM : Skin Cutaneous Melanoma, OV : Ovarian serous cystadenocarcinoma, PAAD : Pancreatic adenocarcinoma, COAD : Colon adenocarcinoma, LUSC : Lung squamous cell carcinoma, BLCA : Bladder Urothelial Carcinoma, LGG : Brain Lower Grade Glioma, LCLL : Chronic Lymphocytic Leukemia, HNSC : Head and Neck squamous cell carcinoma, STAD : Stomach adenocarcinoma, DLBC : Lymphoid Neoplasm Diffuse Large B-cell Lymphoma, MM : Multiple Myeloma Plasma cell leukemia, ESCA : Esophageal carcinoma, LIHC : Liver hepatocellular carcinoma, PRAD : Prostate adenocarcinoma, THCA : Thyroid carcinoma, KIRC : Kidney renal clear cell carcinoma, CESC : Cervical squamous cell carcinoma and endocervical adenocarcinoma .

4.3.1. The neoantigen landscape of a fusion polypeptide

4.3.1.1. Richness difference

We also wanted to understand the variation of predicted HLA binding across the polypeptide sequence surrounding the junction region. This polypeptide sequence consists of 22 amino acids, 10 amino acids from each side of the junction (Fig. 34).

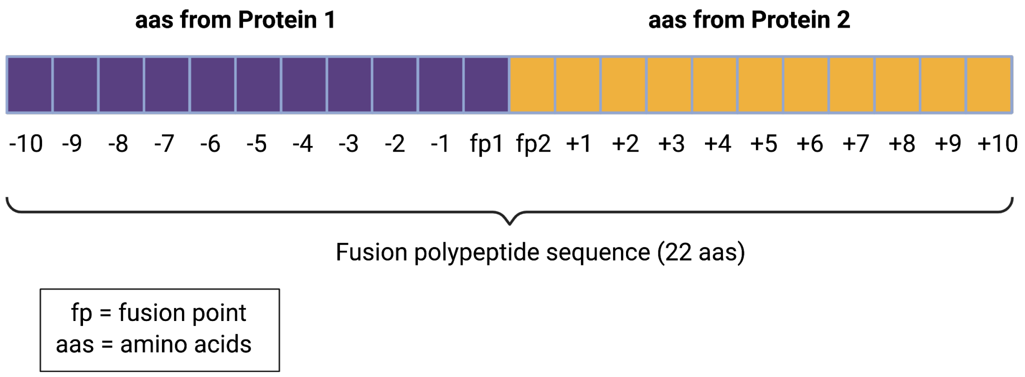
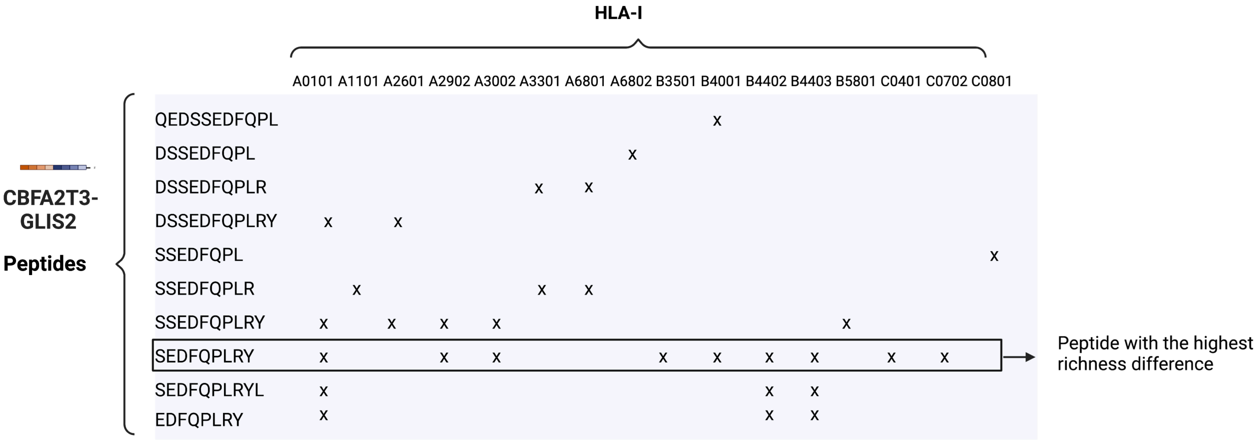

aas from Protein 1
aas from Protein 2
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 fp1 fp2 +1 +2 +3 +4 +5 +6 +7 +8 +9 +10
Fusion polypeptide sequence (22 aas)
fp = fusion point
aas = amino acids

Figure 34. – Polypeptide sequence surrounding the fusion region.

The 45 peptides surrounding the fusion region are generated by FCR through a sliding approach taking 10 amino acids from each side of the fusion points (fp). This allows to generate peptides of 8-12 amino acids (aas) that strictly include the junction region.

Considering the sliding approach adopted (Fig. 27), we expected for the fact that the predictions would generate some redundancies. These redundancies are related to the fact that these peptides have a sequence overlap and therefore they are predicted to bind to the same HLA-I molecule(s) to some extent. For example, as shown in Fig. 35, based on the HLA-I molecules the different peptides are predicted to bind (by two algorithms) the SEDFQPLRY is the peptide with the highest richness difference. It shares sequence overlap with the other peptide sequences and is predicted to bind the same HLA-I molecules as the other peptides as well as additional ones.


CBFA2T3-GLIS2
Peptides
HLA-I
Peptide with the highest richness difference

	A0101	A1101	A2601	A2902	A3002	A3301	A6801	A6802	B3501	B4001	B4402	B4403	B5801	C0401	C0702	C0801
QEDSSEDFQPL										x						
DSSEDFQPL								x								
DSSEDFQPLR						x	x									
DSSEDFQPLRY	x		x													
SSEDFQPL																x
SSEDFQPLR		x				x	x									
SSEDFQPLRY	x		x	x	x								x			
SEDFQPLRY	x			x	x				x	x	x	x		x	x	
SEDFQPLRYL	x										x	x				
EDFQPLRY	x										x	x				

x next to a peptide and below a given allele means that the peptide is predicted to be binder to this allele by 2 algorithms

Figure 35. – Example of richness difference (CBFA2T3-GLIS2 fusion).

The peptide sequences from the CBFA2T3-GLIS2 fusion predicted to bind at least one HLA-I molecule by 2 algorithms are shown in a tabular format. For each peptide sequence, an x corresponds to it being a binder to the HLA-I molecule mentioned on the top of the table. The peptide sequence 'SEDFQPLRY' is considered as having the highest richness as it is the peptide that is predicted to bind the same molecules as the other peptide as well as additional ones.

Therefore, in order to understand where within the polypeptide sequence (Fig. 34) is the best source for the richest peptides in HLA-I binding, we have plotted the richness difference proportion of peptides per fusion. In Fig. 36, the proportion of richness difference per location within the 22-aa polypeptide sequence are displayed for each fusion. Our results show that the highest richness difference across fusions generally is mapped to peptides that start 7 amino acids before the breakpoint from the protein 1's side until 3 amino acids before the breakpoint. This could mean that the optimal anchor residues (2 and 9) are generally obtained from both proteins at such residues within the polypeptide sequence (Fig. 36). This alludes to the optimal region around the fusion junction from protein 1 that would generate potential strong cross-HLA-I binding peptides.

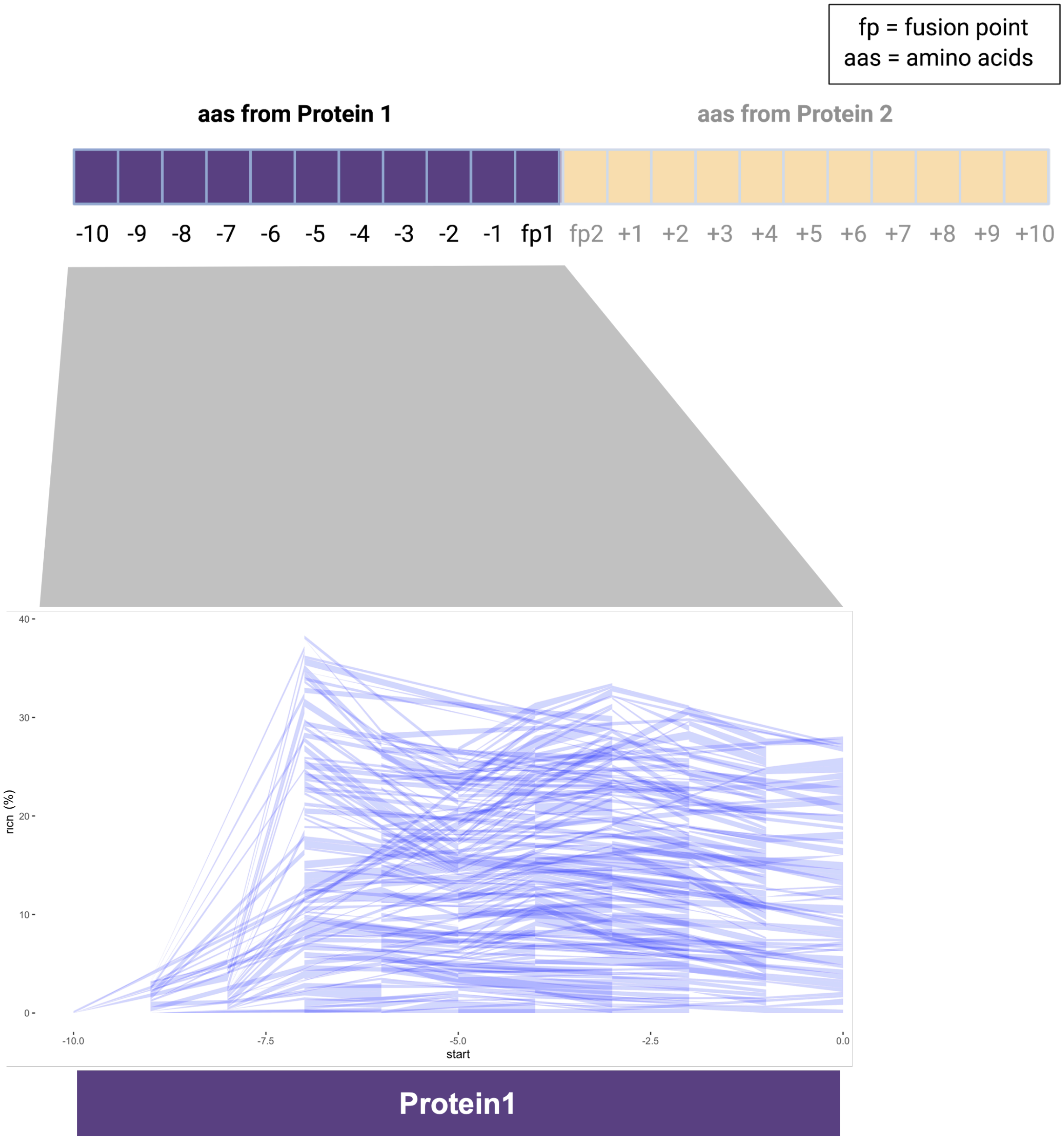


Figure 36. – Analysis for understanding the position of the predicted most HLA-I binding with regards to the fusion point from Protein 1 side.

The upper panel shows the area of interest for the analysis of the source within the 22-aa polypeptide sequence. The line plot represents the proportion of richness difference per fusion, across fusions according to where the starting amino acid within the 22-aa polypeptide sequence is. The x-axis is the starting position, the y-axis is the proportion of richness difference.

4.3.2.2. Diversity of predicted HLA-I binding of fusion-derived peptides

Then, to understand the diversity of predicted HLA-I-peptides derived from each gene fusion, we have used the LCBD indicator with p adj<0.05 to select these (Legendre, 2014). We used the LCBD indicator to identify peptides that have a diverse predicted HLA-I binding in comparison with the rest of the gene fusion-derived peptides. For example, in the case of CBFA2T3-GLIS2 (Fig. 35) the predicted HLA-I binding of its derived peptides is rather "homogeneous". In contrast, we have found that ETV6-RUNX1 consists of two groups of peptides, first group is predicted to bind a set of HLA-I molecules, whereas the second is predicted a completely different set of HLA-I molecules (Fig. 37). The RIAECILGM peptide has a significant LCBD (p adj <0.05) which means that its predicted HLA-I binding is diverse in comparison with the other peptides from the same fusion. This means that the ETV6-RUNX1 fusion could have two potentially strong candidates, one with highest richness difference (MPIGRIAEC) and one diversly-binding (RIAECILGM). Due to this increase in the sets of predicted HLA-I peptides from the ETV6-RUNX1, such gene fusions could be considered as "heterogeneous". Interestingly, immunogenic potential of both peptides with patient T cells has been published (Zamora et al., 2019a)

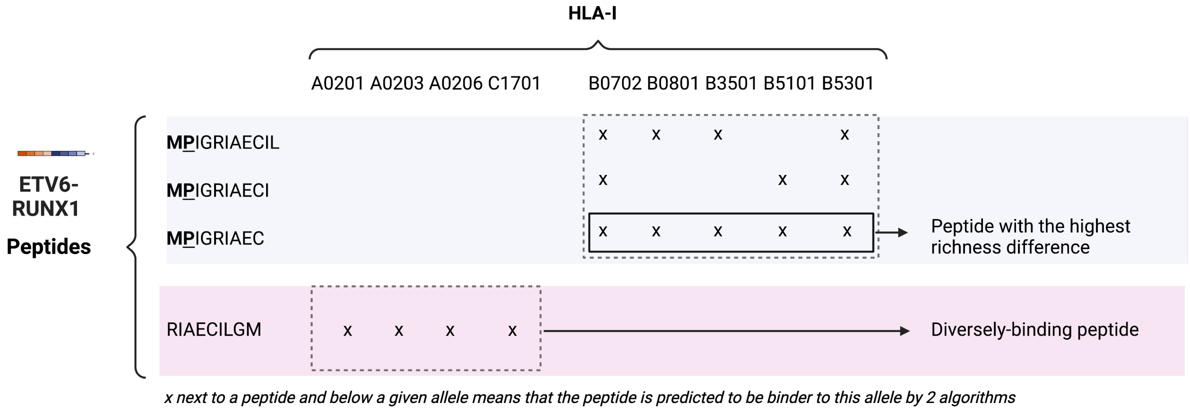

HLA-I
A0201 A0203 A0206 C1701
B0702 B0801 B3501 B5101 B5301
ETV6-RUNX1 Peptides
MPIGRIAECIL
MPIGRIAECI
MPIGRIAEC
RIAECILGM
x
Peptide with the highest richness difference
Diversely-binding peptide
x next to a peptide and below a given allele means that the peptide is predicted to be binder to this allele by 2 algorithms

Figure 37. – Example of diverse predicted HLA-I binding (ETV6-RUNX1).

*The plot shows the predicted binder peptides (to at least one HLA-I molecule by 2 algorithms) from the ETV6-RUNX1 fusion. The set of peptides in the blue background: MPIGRAECIL, MPIGRIAECI, MPIGRIAEC, all start from the same Methionine amino acid, and have proline as anchor residue. All three of these are predicted to bind HLA-B*07:02 and HLA-B*53:01 which is explained by the common proline anchor residue which is characteristic of HLA-B*07:02 (Fig. 5). Among these MPIGRIAEC is the peptide with the highest richness difference. On the other hand, the fusion is also predicted to generate a peptide RIAECILGM that is predicted to bind a different set HLA-I molecules. There is no overlap between this set of HLA-I molecules and those predicted to be bound by the other peptides from the same fusion. This peptide is shown in a pink background. The RIAECILGM has a significant LCBD value (p adj<0.05) and is considered as diversely-binding.*

To understand whether certain positions within the 22 aa polypeptide sequence are important for generating diversely-binding peptides, we firstly identified the heterogeneous fusions from the LIGEA database. We found that 44 out of 523 fusions are heterogeneous (~8.4%) and can generate one or more peptides with significant LCBD (p adj<0.05). Next, we generated a sequence logo the 22 aa polypeptide sequences of these 44 gene fusions in order to understand whether there are certain positions that are more prominent for these polypeptides.

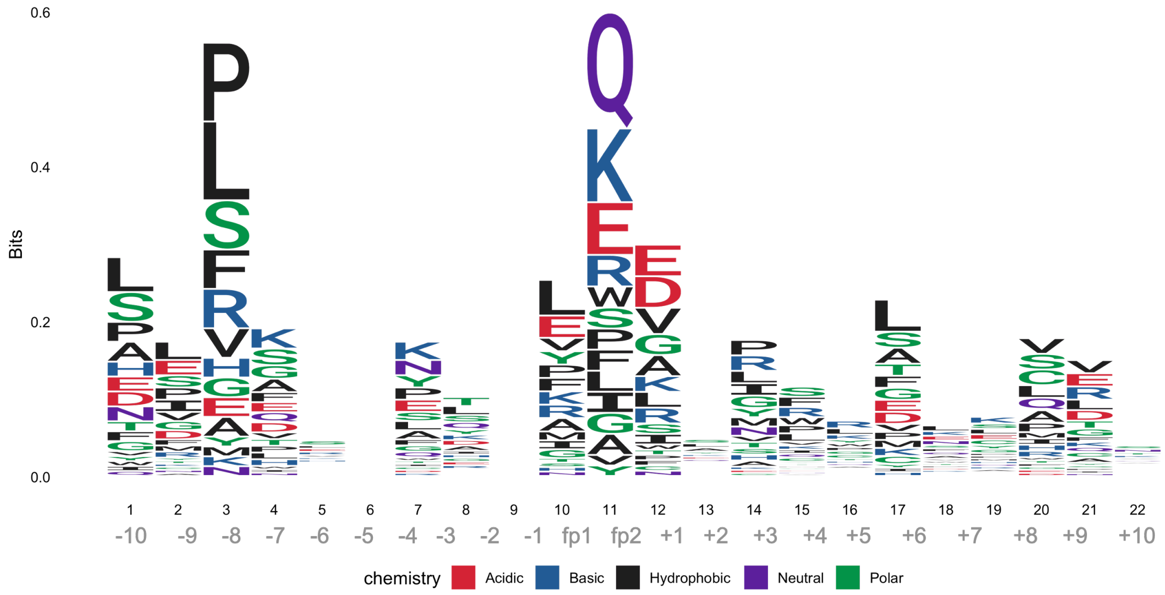


Figure 38. – Logo plot of the polypeptides that have at least one predicted diversely binding peptide across fusions.

The plot shows a sequence logo of the polypeptides that were shown to generate a peptide with a diverse HLA-I profile in comparison to the rest of the predicted HLA-I-peptides

Fig. 38 show to that 1) the -8 position from fp1 as well as 2) fp2 could be important for this (Fig. zb). This is possibly accounts for the optimal position of certain residues so as to generate peptides with more than 1 set of HLA-I alleles they would be predicted to bind. To sum up, gene fusions that generate (or are predicted to generate) diverse HLA-I-peptides, might have certain residues at given positions so as to create the generate multiple candidates. Furthermore, these gene fusions have great potential for the selection of multiple peptides from these. Also, due to the predicted diverse HLA-I binding nature it is possible to consider that these could boost the potential population coverage.

4.3.2. The predicted HLA-I and population coverage of gene fusion-derived peptides

To explore the potential population coverage as well as the cross HLA-I binding of gene fusions. To do that, we firstly selected subset of gene fusions, those involving oncogenes and those found in COSMiC cancer database (from primary samples - COSMIC, as annotated by FusionCatcher (Nicorici et al., 2014; Tate et al., 2019)). From 120 fusions, we randomly selected 20 fusions: 7 that are heterogeneous and 13 that are homogeneous in their HLA-I predicted binding. From each heterogeneous gene fusion, all peptides with significant LCBD (p adj<0.05) were selected as well as the peptide with the best supertype coverage score (see materials and methods). In the plot below (Fig. 39), for each peptide derived from a gene fusion (in the format: [Fusion]_[Peptide sequence]) the HLA-I molecules predicted to be bound are shown on the right and their associated estimated population coverage worldwide is shown on the left. The results in Fig. 39 clearly show that these clinically relevant gene fusions are predicted to generate HLA-I binding peptides that cover a relatively high estimated population and reaches up to at least 89% (YAP1#MAML2)(Fig. 39). Also, when looking at each heterogeneous fusions, it is possible to see that the diversely-binding peptide(s) boost(s) the estimated population coverage.

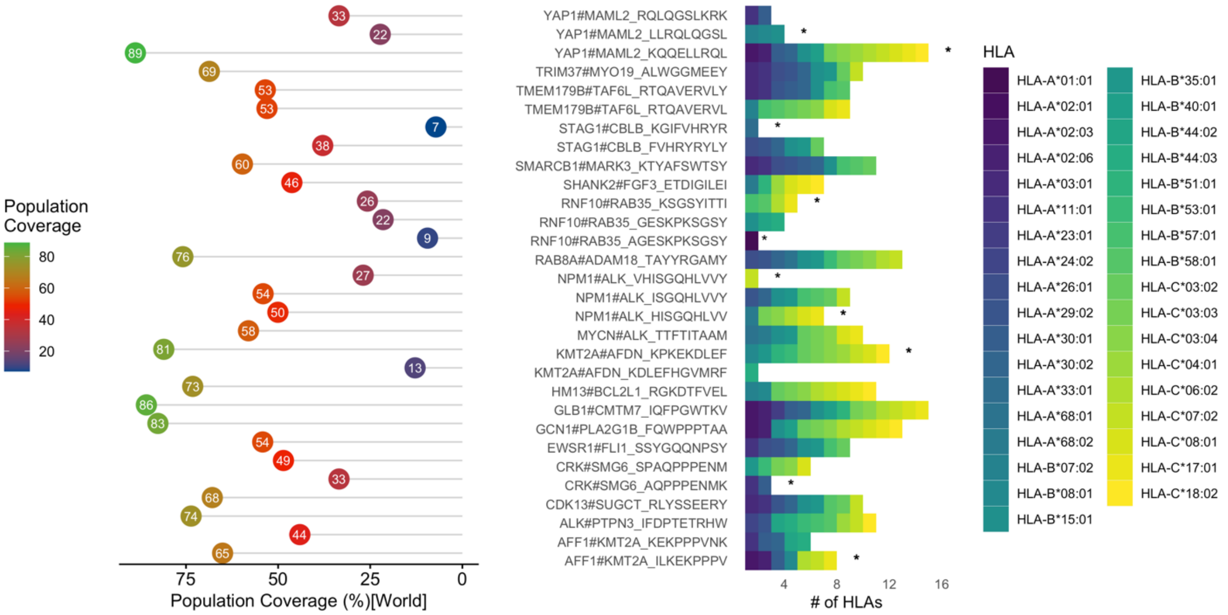


Figure 39. – The predicted binding of 27 fusion peptides (from 19 gene fusions) from gene fusions involving an oncogene and/or also found in COSMIC.

*(Left Panel) The estimated population coverage (world) of the HLA pairs predicted to bind the peptides shown in the middle (middle Panel) as an identifier representing the peptide sequence and the fusion separated by an "_" ([Peptide sequence] _[gene1#gene2]). (Right Panel) displays the HLA-I alleles predicted be bound by the peptides in the middle panel. Each square represents an HLA-I allele. The * represents a significant effect of heterogeneity of the peptide in comparison with the other peptides from the same fusion (adjusted p-value of the local contribution to beta diversity p.adj<0.05).*

4.4. Focusing on the gene-fusion derived neoantigens of HLA-typed cell lines

Applying FCR predictions on gene fusions found in COSMIC and/or involving an oncogene with 35 of the most frequent HLA-I alleles have demonstrated that GF-Neo are predicted to bind multiple HLA-I alleles and could have wide population coverage. Next, we wanted to apply the FCR predictions on the actual HLA-I molecules expressed on the surface of each of the cell lines. First, we have retrieved the HLA typing of the cell lines from the TRON database. We were able to retrieve the HLA typing of 62 cell lines spanning 20 cancer types. The total number of fusions across these is 100 fusions (96 unique gene pairs). We found that per disease, across cell lines, 25-30% of peptides derived from fusions are predicted to bind more than one HLA-I allele of those expressed on the surface whereas the biggest proportion of gene fusion-derived peptides are predicted to bind just one HLA-I molecule of those expressed (Fig. 40A). Of note, we did not take fusion heterogeneity into consideration, as the set of alleles per fusion was of 3 up to 6, which would not allow to conduct such analyses. We then, delved deeper by looking at the specific fusions, within each cell line and their binding potential with respect to the HLA-I alleles expressed on the cell. Our results show that 86% of the gene fusions (83 out of 95) generate a peptide that is predicted to bind at least one of the HLA-I alleles expressed on the surface of the cell lines (Fig. 40B). Consequently, the neoantigens found in these cell lines show strong potential in generating peptides that would bind the HLA-I alleles expressed on the surface.

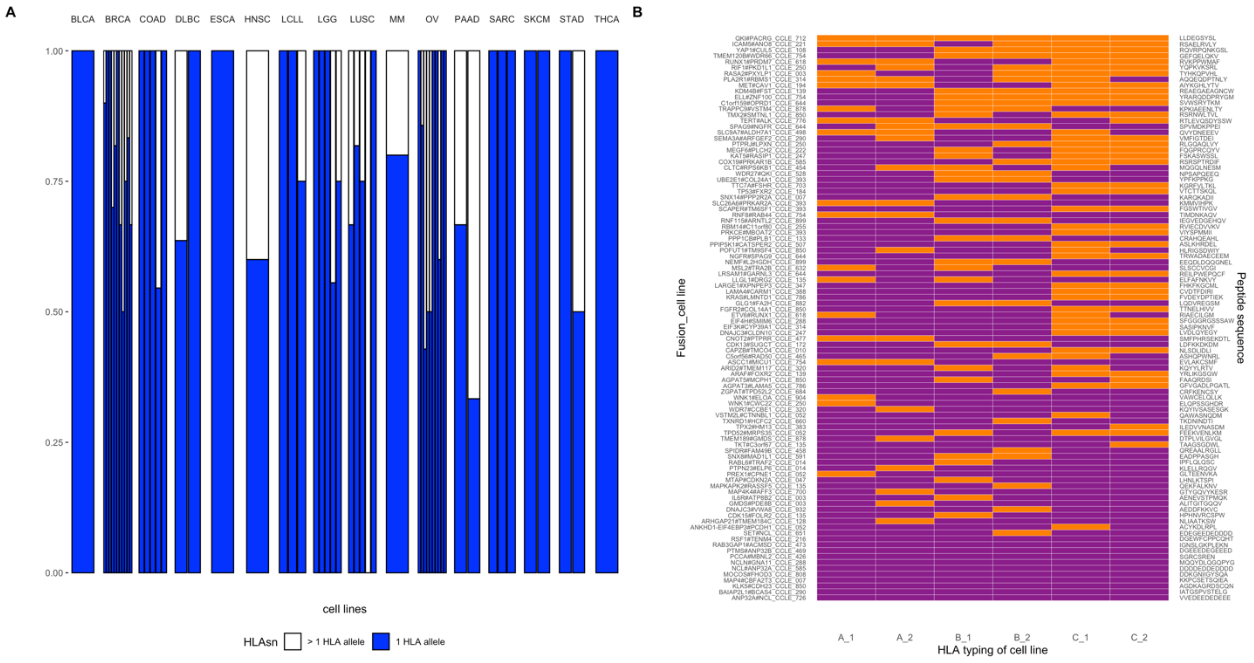


Figure 40. – The predicted (cross-) binding of gene fusion derived peptides of HLA-typed cells.

A) *The proportions of gene fusion-derived peptides binding 1 HLA allele (out of the 6 presented on the surface of the cell) and those binding more than one. The x-axis represents the cell lines. The proportions are calculated based on the peptides from the gene fusions within each cell line. The heatmap in B) shows the alleles expressed on the surface of the cell predicted to be bound by the peptides. The x-axis represents each HLA-I allele from the cell lines' HLA typing. Due to the possible homo-/heterozygosity of the A/B/C genes, cells could present on their surface 3 up to 6 different HLA-I alleles. The y-axis has two sides: on the left, the gene fusion and the cell line it was identified in are shown ([gene1#gene2]_[CCLE_]). On the right side, the peptide chosen from a given fusion corresponds to the peptide predicted to bind the highest number of the HLA-I alleles expressed out of the 45 possible fusion peptides.*

4.5. Analysis of immunopeptidomic data of 92 mono-allelic cell lines

Our peptide-HLA-I prediction results revealed that GF-Neo are predicted to bind multiple HLA-I alleles. Therefore, we asked whether HLA-I peptides can bind more than one HLA-I molecule. To do that, we used a recent and extensive dataset of immunopeptidomic (Sarkizova et al., 2020b). In their experiments, Sarkizova et al. (2020) generated cell lines expressing single HLA-I alleles and then sequenced the peptides presented on the HLA-I alleles of these cell lines. In total, the immunopeptidomes of 92 different HLA-I alleles (Fig.41). These 92 HLA-I alleles cover all 12 supertypes (see section 1.2.2 and Annex table 1) and include alleles that have been less studied such as HLA-A*34:01 and HLA-A*74:01 to name a few.

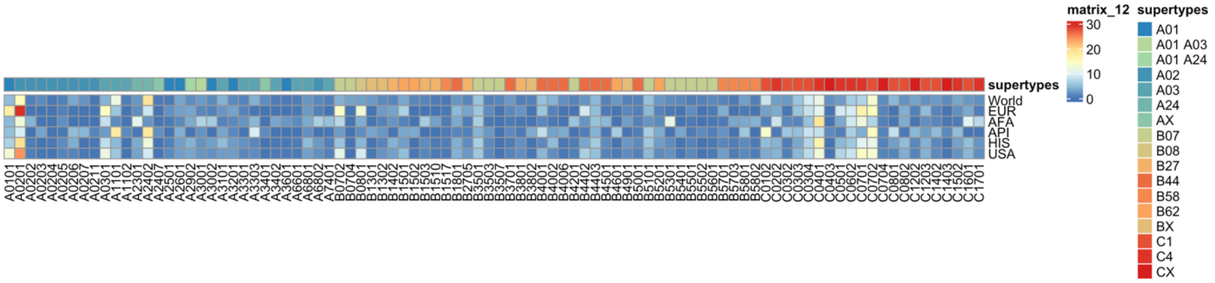


Figure 41. – The 92 alleles considered. Heatmap plot of the HLA-I frequency in different regions and supertype group of the 92 HLA-I alleles considered from the dataset.

4.5.1. Peptides could bind HLA-I alleles of different supertypes

Our results show that as expected the largest portion of the peptides (~70%) binds 1 HLA-I allele (Fig. 42A). Around 18% of the peptides bound (~22000 peptides) two HLA-I alleles. Examining the peptides that bound 2 HLA-I alleles (for simplification purposes, we will refer to these as di-allelic peptides (diHLAp) henceforth), it is possible to see that HLA-I alleles have different propensities of sharing peptide repertoires (fig. 42B). We chose 6 HLA-I alleles HLA-A*01:01, HLA-A*02:05, HLA-A*02:07, HLA-A*30:01, HLA-B*15:01, HLA-B*44:02 to display this. We wanted to explore the alleles bound by a same peptide. Did these diHLAp bind two HLA-I alleles from the same or different supertypes? These six were chosen as they show binding to the same and different supertypes as well as HLA-A and HLA-B molecules. The diHLAp that bound A*30:01 (belonging to the A01A24 supertypes) as well as those that bound A*02:05 and A*02:07 were found to also bind HLA- B and -C molecules. On the other hand, the pie chart for the B*44:02 molecule shows that the diHLAps especially (~97%) one other molecule (B*44:03) from the same supertype (Fig. 42C). This is also similar with A*0101 molecules. However, diHLAp bound to B*1501 bound mainly molecules from the same gene (HLA-B) but that belong to different supertypes. The focus here, on the supertypes is due to the fact HLA molecules of the same supertype have very similar binding preferences in comparison with those of different supertypes. Together, these results have shown that peptides could bind multiple HLA-I molecules of same and different supertypes.

Despite that this analysis was restricted to peptides binding to two HLA-I molecules but did reveal that peptides are able to bind HLA-I molecules belonging to different supertypes, we wanted to focus on the supertypes.

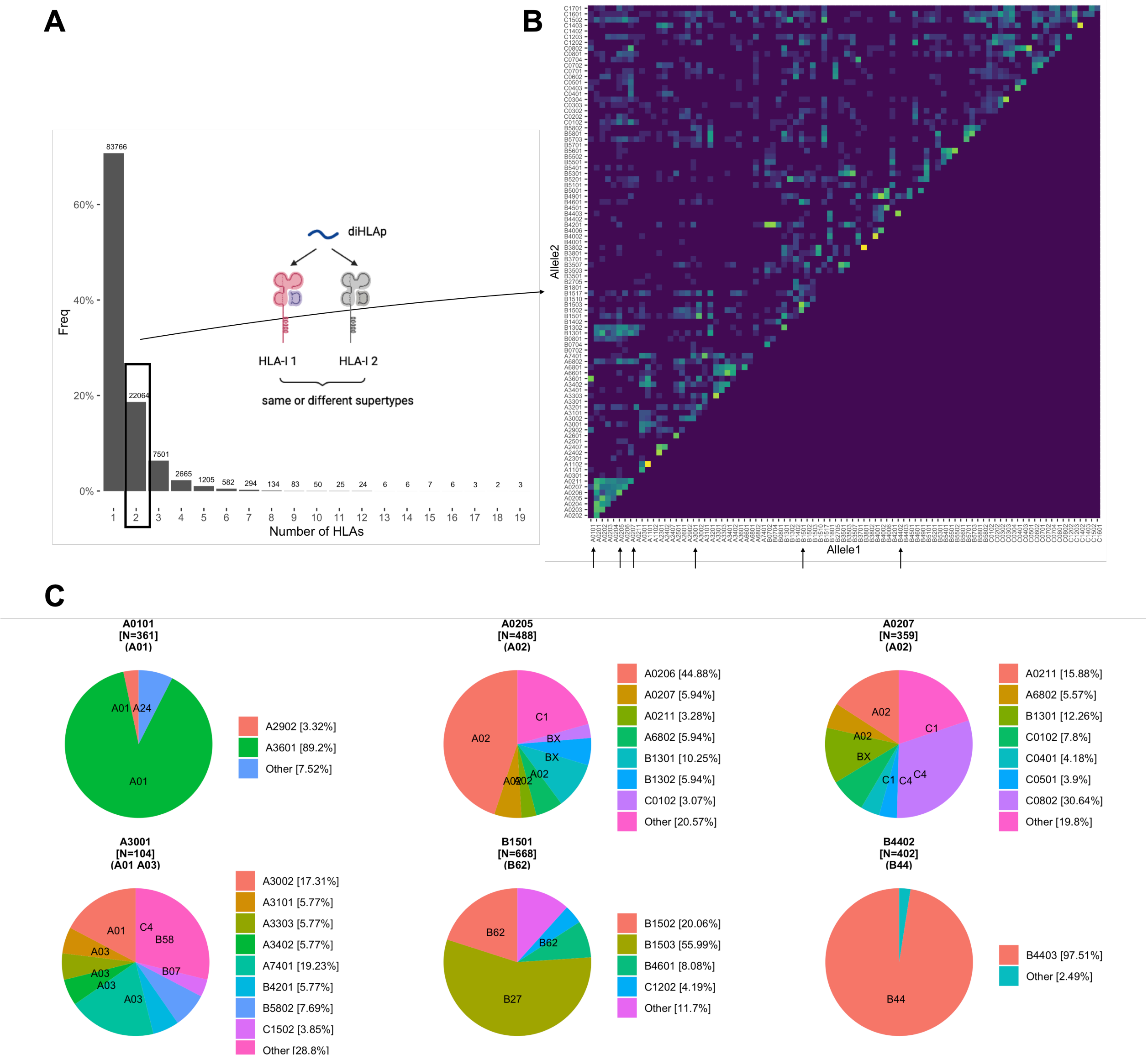


Figure 42. – The HLA-I binding landscape of eluted ligands from experiments on 92 mono-allelic cell lines.

A) histogram displaying the frequency of peptides binding/eluting with one or more HLA-I molecules. The x-axis represents the number of HLA-I alleles bound; the y-axis corresponds to the frequency. B) Focusing on peptides that eluted with 2 HLA-I alleles, the heatmap displays a log frequency (brightness of the color) of the combination of each allele pair eluting with the same peptide (x- and y-axes are each allele). C) Pie charts for 6 HLA-I alleles of the frequencies of HLA-I pairs by the same peptide. Above each pie is indicated the HLA-I allele, the number of observations and the supertype it belongs to. Within the pie charts, the pie slices are indicated based on their belonging to a supertype. On the left of each pie chart are indicated the percentages.

(i.e. Among the peptides binding two HLA-I molecules, of those that bound A0101 (A01 supertype), 89% of these peptides also bound A3601 (a01 supertype also). Lastly, 'Other' designates the grouping of any HLA-I allele with a percent less than 2%.

Based on that, we also wanted to evaluate the same analysis but from the perspective of supertypes. We see a similar trend when the alleles were grouped by supertype. Approximately 85% of peptides bound 1 supertype (100517 peptides), which means ~15% of peptides bound more than one supertype. The large majority bound 2 supertypes (12.3%, 14577 peptides) (Fig. 43A). We were also interested in exploring peptides that bind HLAs from two different supertypes referred to from now as di-Supertype peptides (diSup). Generally, disregarding the unclassified supertypes (that we encoded as BX, AX, CX, X refers to the unclassified nature of alleles belonging to this supertype) it is possible to see that a large number of diSups that were found to bind an A supertype, also bound B (i.e. A01 & B58, A01& B62) or C supertypes (A02 & C1) and same for diSups binding B & C supertypes (B07 & C1, B08 & C1, B58 & C1, B58 & C4) (Fig. 43B).Furthermore, we still see that the B44, B27, A24 and to a lesser extent A03 bound alleles belonging mainly to B supertypes, specifically the unclassified supertypes (Fig. 43C). This meant that for alleles belonging to certain supertypes have more restrictive specificities whereas other have more loose ones, as also described in the literature (Kaufman, 2018; Manczinger et al., 2021). Also, this further shows that there must be an overlap in the binding specificities of these molecules which allows for the peptide to bind.

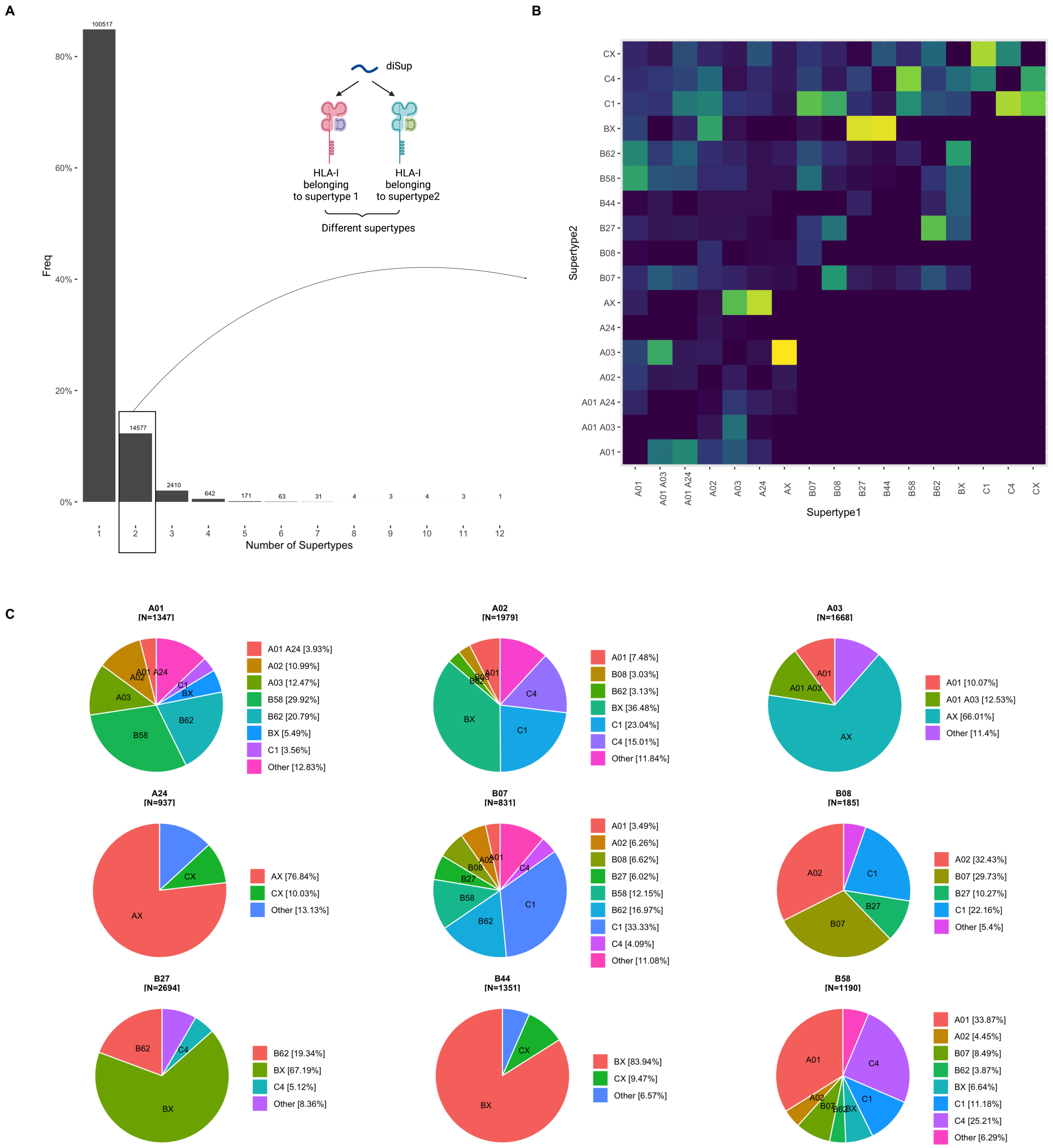


Figure 43. – The HLA-I supertypes binding landscape of eluted ligands from experiments on 92 mono-allelic cell lines.

A) histogram displaying the frequency of peptides binding/eluting with HLA-I molecules belonging to one or more HLA-I molecules. The x-axis represents the number of supertypes that the HLA-I molecules bound by the peptides belong to, the y-axis corresponds to the frequency. B) Focusing on peptides that eluted with HLA-I alleles, the heatmap displays a Hellinger transformed frequency (brightness of the color) of the combination of each supertype pair eluting the same peptide (x- and y-axes are each allele). C) Pie charts for the main supertypes of the frequencies of supertype pairs by the same peptide. Above each pie is indicated the supertype, *and the number of observations. The pie slices are labeled with the second supertype. On the left of each pie chart are indicated the percentages. Lastly, 'Other' designates the grouping of any HLA-I allele with a percent less than 2%.*

4.6. Analyze the potential of clinically relevant gene fusions to generate MHC-I super binders

4.6.1. Applying FCR on clinically relevant pediatric gene fusions

4.6.1.1. Selecting 3 cross-HLA-I peptides

We wished to analyse the cross-HLA-I binding of gene fusion-derived peptides with clinically relevant gene fusions experimentally (in vitro). Therefore, we selected sixteen clinically relevant gene fusions were selected. These included two positive controls from the literature, (2 fusions: ETV6-RUNX1, CBFB-MYH11) (Biernacki et al., 2020b; Yotnda, Firat, et al., 1998; Zamora et al., 2019b), 9 fusions from a validated fusion panel derived from patient sequencing (9 fusions: KIAA1549-BRAF, GIT2-BRAF, CENPC-ABL1, ZNF274-JAK2, NUP153-ABL1, SPTAN1-ABL1, GOLGA5-JAK2, NOTCH1-ROS1, ATF7IP-PDGFRB) (Chang et al., 2019b), and one gene fusion obtained from the M0E7 cell line in the LigeA database (CBFA2T3-GLIS2) (Gioiosa et al., 2018). CBFA2T3-GLIS2 fusion is found in 30% of non-Down syndrome Acute MegaKaryoblastic Leukemia (non-DS AMKL) and is recurrent and especially found in infants (Bolouri et al., 2018; Gruber et al., 2012; Masetti et al., 2013; Smith et al., 2020).

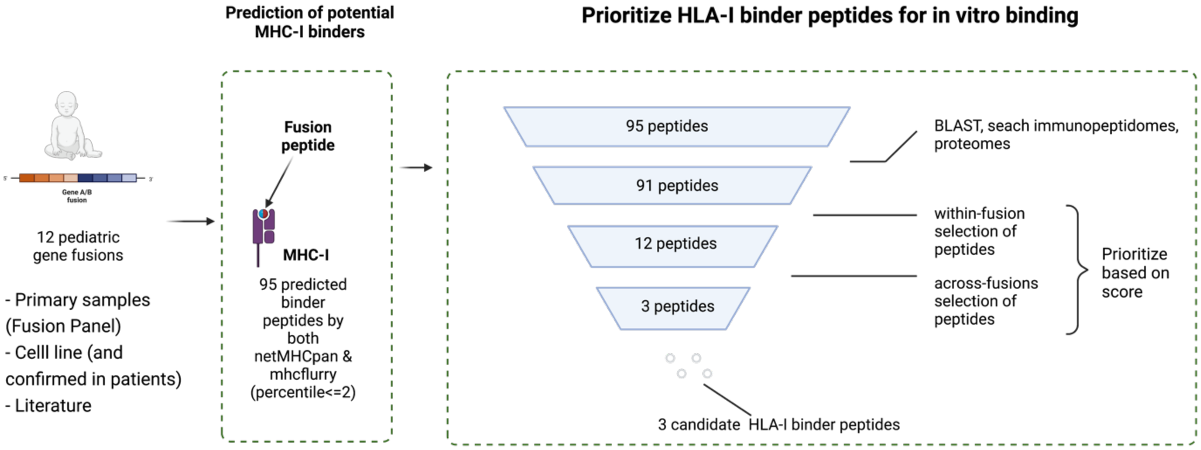

Prediction of potential MHC-I binders
Prioritize HLA-I binder peptides for in vitro binding
Gene A/B fusion
12 pediatric gene fusions
- Primary samples (Fusion Panel)
- Celll line (and confirmed in patients)
- Literature
Fusion peptide
MHC-I
95 predicted binder peptides by both netMHCpan & mhcflurry (percentile<=2)
95 peptides
91 peptides
12 peptides
3 peptides
BLAST, seach immunopeptidomes, proteomes
within-fusion selection of peptides
across-fusions selection of peptides
Prioritize based on score
3 candidate HLA-I binder peptides

Figure 44. – Prioritization of 3 candidate HLA-I binder peptides from 12 clinically relevant pediatric gene fusions.

The sequences of 12 recurrent and clinically relevant pediatric gene fusions were selected. These validated sequences and have different sources: Patient and cell line-derived xenografts, fusion panel based on primary samples or the literature. Fusion peptides were generated from these and MHC-I binding predictions were run with both algorithms, these were filtered based on algorithm agreement. This step was followed by a step aimed at making sure the sequences are unique in that there is no overlap with self-peptides. This was done by doing a BLAST search (similarity matching) with human and mouse non-redundant protein database, and exact matching by searching the human and mouse proteomes and immunopeptidomes/ One fusion was eliminated because it matched with a protein different from its parent proteins. Then, one up to two peptides were selected from each fusion. Then, these were narrowed based on the strength of their predictions (essentially through calculating a ratio), comparisons with binding motifs, cross-HLA-I binding potential.

All 12 fusion sequences went through FCR against 35 of the most common HLA-I alleles (Annex table 1, same as used for the database), this has allowed us to obtain 95 predicted peptide-HLA-I binders by both algorithms at a percentile threshold of 2. To assess the uniqueness of the peptides and thereby limiting risks of cross-reactivity risks, the peptide sequences were searched in the immunopeptidomes and proteomes of human and mouse (Consortium et al., 2020; Kubiniok et al., 2022). A BLAST (Basic Local Alignment Search Tool) search was also made for these peptides. As a result, we have excluded one gene fusion which has one peptide that has aligned (80%) with a protein that is not one of the parent proteins. Then, to narrow down the list of peptides to the most prominent ones, one up to 2 peptide sequences were selected from each fusion, totalling 12 peptides. These were further reduced by comparing the scores of the peptide sequences. Finally, we have selected 3 peptides based on the predicted ability of these to bind HLA-A & B alleles and the ability to bind different HLA-I alleles of the same supertype: SEDFQPLRY (CBFA2T3-

GLIS2), IALPFKVVV (ATF7IP-PDGFRB) and IPQDTIPVL (ZNF274-JAK2) (see Fig. 37, table 3 for FCR predicted for HLA-A&B alleles and Annex table 2 for complete FCR predictions (including C alleles)).

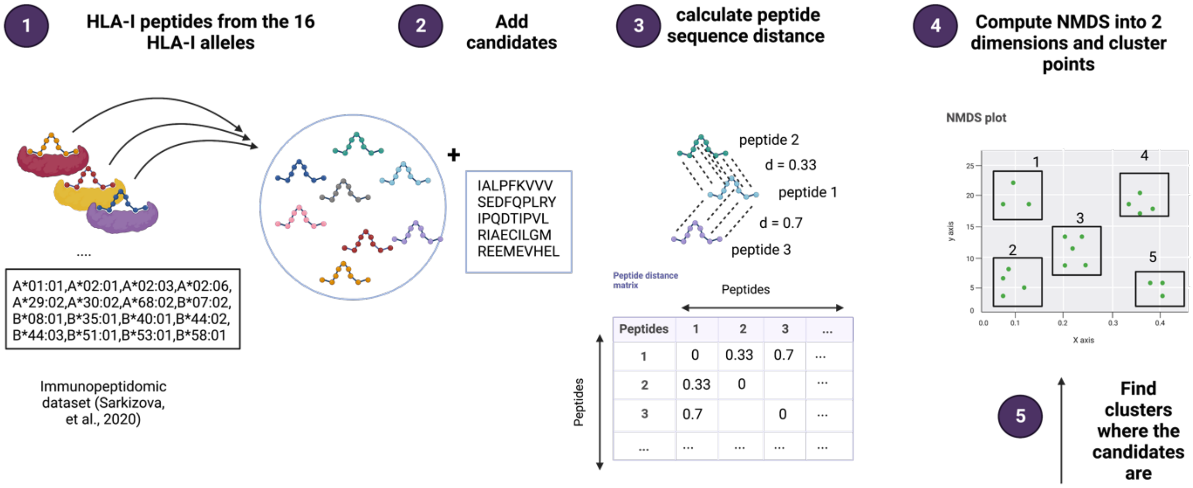
	Fusion	**Peptide Sequence**	**HLA-I predicted to be bound**	**# of HLA-I**	**Supertypes that the HLA-I alleles belong to**	**# of Supertypes**
🟦	**ATF7IP-PDGFRB**	IALPFKVVV	A*02:01, A*02:06, A*68:02, B*51:01, B*58:01	5	A02, B07, B58	3
🟦	**CBFA2T3-GLIS2**	SEDFQPLRY	A*01:01, A*29:02, A*30:02, B*35:01, B*40:01, B*44:02, B*44:03	7	A01, A01 A24, B07, B44	4
🟦	**ZNF274-JAK2**	IPQDTIPVL	B*07:02, B*08:01, B*35:01, B*51:01, B*53:01	5	B07, B08	2
🟨	**ETV6-RUNX1**	RIAECILGM	A*02:01, A*02:03, A*02:06	3	A02	1
🟨	**CBFB-MYH11**	REEMEVHEL	B*40:01, B*44:02, B*44:03	3	B44	1

Tableau 3. – FCR predictions of our candidate peptides with HLA-A & HLA-B alleles.

The table above shows the FCR predicted HLA-I alleles to be bound by the candidate peptides selected and the control peptides. Color-coding used to differentiate between the candidate peptides and the control peptides.

4.6.1.2. Using sequence similarity between candidate peptides and HLA-I peptides from the immunopeptidome dataset to hypothesize binding

As complement to the FCR predictions, we wished to use another computational approach in order to determine whether these peptides are likely to bind the predicted HLA-I alleles. This approach is based on the dataset of eluted ligands from mono-allelic cell lines analysed in section 4.5. Essentially, our approach relied on clustering peptides based on peptide sequence similarity (Fig. 45). Firstly, we have extracted all 9-mer peptides from these 16 HLA-I alleles (N = 6856 peptides). These 16 alleles (in step 1 of Fig. 45) are based on the total set of alleles from the FCR predictions (table 3). Importantly, for this analysis, the peptide pool used was derived from peptides identified in only one allele.


1
HLA-I peptides from the 16 HLA-I alleles
2
Add candidates
3
calculate peptide sequence distance
4
Compute NMDS into 2 dimensions and cluster points
....
A*01:01,A*02:01,A*02:03,A*02:06,
A*29:02,A*30:02,A*68:02,B*07:02,
B*08:01,B*35:01,B*40:01,B*44:02,
B*44:03,B*51:01,B*53:01,B*58:01
Immunopeptidomic dataset (Sarkizova, et al., 2020)
+
IALPFKVVV
SEDFQPLRY
IPQDTIPVL
RIAECILGM
REEMEVHEL
peptide 2
d = 0.33
peptide 1
d = 0.7
peptide 3
Peptide distance matrix
Peptides
Peptides
NMDS plot
1
2
3
4
5
y axis
X axis
0
5
10
15
20
25
0.0
0.1
0.2
0.3
0.4
5
Find clusters where the candidates are

Peptides	1	2	3	...
1	0	0.33	0.7	...
2	0.33	0		...
3	0.7		0	...
...	...	...	...	...

Figure 45. – Using sequence similarity between candidates and eluted MHC-I ligands as complement to hypothesize potential HLA-I alleles bound.

The figure above shows the overall workflow employed to achieve the analysis. Our predictions for our candidates and positive controls with FCR yielded a set of 16 HLA-I alleles for the peptides to be tested against. Therefore, we decided to firstly (1) extract HLA-I peptides from the immunopeptidomic dataset (used in section 4.5) of these 16 HLA-I alleles. Then, (2) we added our candidates and positive controls. Next, we (3) generated a square distance matrix of the pairwise distances between all the sequences (6861 rows, 6861 columns). Then, (4) to reduce this matrix into a 2-dimensional one, we compute Non-Metric Dimensional Scaling (NMDS) followed by a k-means clustering of the data points. Finally, we (5) looked for our candidates and controls within these clusters. By retrieving the HLA-I alleles that the peptides within each cluster bound in the immunopeptidomic experiments it would be possible to have an idea as to which HLA-I alleles our candidates and controls would potentially bind.

To this list were added the 3 candidate peptides along with the 2 positive control peptides (REEMEVHEL from CBFB-MYH11 fusion and RIAECILGM from ETV6-RUNX1 fusion) (N = 6856 + 5 = 6861 peptides). Then, we implemented the protocol (Venema et al., 2021) consisting of calculating a distance matrix based on an established peptide-MHC similarity matrix (Kim et al., 2009) while accounting for molecular Entropy. This yielded a matrix of 6861 rows and 6861 columns. Next, to reduce this to 2-dimensions we applied non-metric dimensional scaling (NMDS), this would allow to show these peptides in a two-dimensional plot and display peptides whose sequences are more similar closer and those that are more different, farther. The optimal clustering is a partition of the points on the NMDS axes into 23 groups (see Methods). The peptides were then clustered into 23 groups and plotted; each point has a different color based on the cluster it belongs to. Finally, we looked at the alleles that the peptides sharing the clusters with the peptides we have added (in step 2) have been grouped in to. 1)To better understand the peptide sequences of each cluster we constructed logo plots of the peptides within each cluster and (2) to predict the HLA-I alleles that our candidates are likely to bind, we retrieved the HLA-I alleles of the peptides within these clusters (Fig. 46).

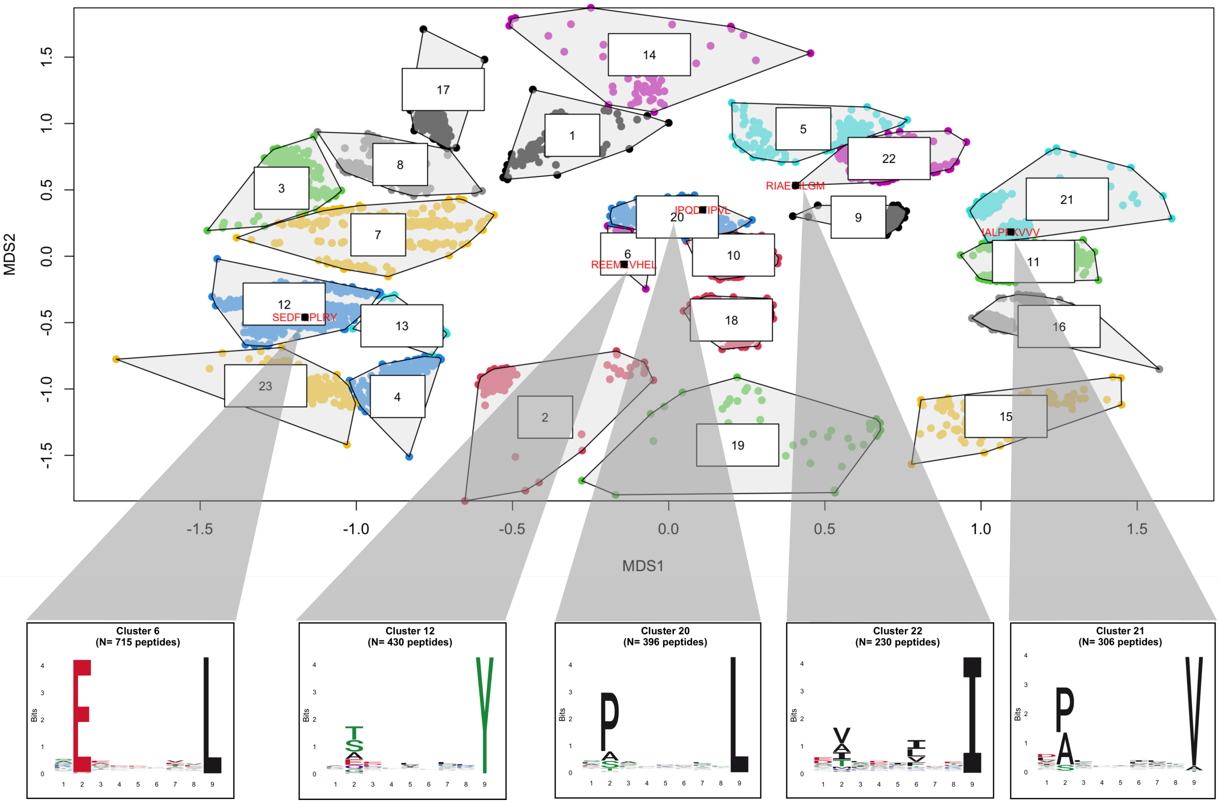
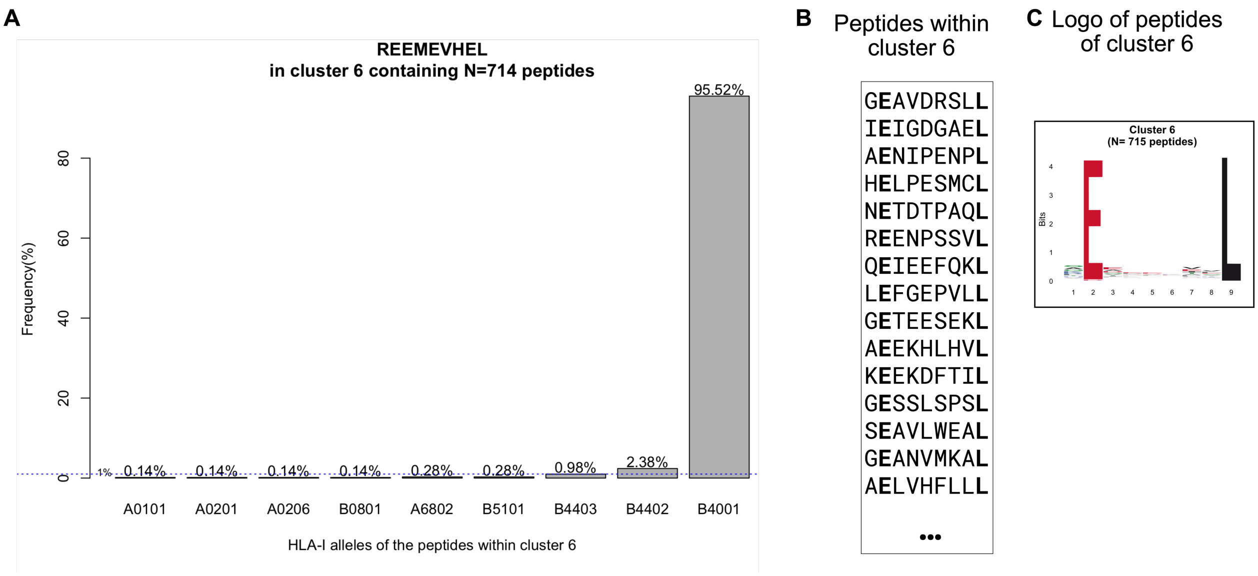


Figure 46. – NMDS of the HLA-I peptides of the 16 HLA-I alleles and the logos of the clusters containing the peptides to be tested in vitro.

The NMDS plot with polygons delimiting the 23 clusters and labeling. The x- and y-axes are the NMDS axes, the two-dimensional plot aims at representing the points closest to each by their similarity to each other and dissimilarity to others. The NMDS analysis is based on the 16 HLA-I alleles that the peptides were tested against. The peptides whose sequence is displayed based on their location on the plot are the peptides that were tested for their binding in vitro. At the bottom of the figures are the logos of the clusters that these peptides belong to.

For example, REEMEVHEL was found in cluster 6 (Fig. 46). Cluster 6 is made up of 714 peptides, if we retrieve the HLA-I alleles in which the peptides in this cluster were found (in the immunopeptidomic dataset), we would be able to understand which HLA-I alleles REEMEVHEL is likely to bind, based on the frequency of the HLA-I alleles of the peptides within the same cluster, that are higher than 1% (Fig. 47). This same process was done for the other peptides also (Table 4 and Annex Fig. 1). The analysis clearly shows that the fusion peptides share key residues with eluted ligands of different HLA-I alleles .



G**E**AVDRSL**L**
I**E**IGDGAE**L**
A**E**NIPENP**L**
H**E**LPESMC**L**
N**E**TDTPAQ**L**
R**E**ENPSSV**L**
Q**E**IEEFQK**L**
L**E**FGEPVL**L**
G**E**TEESEK**L**
A**E**EKHLHV**L**
K**E**EKDFTI**L**
G**E**SSLSPS**L**
S**E**AVLWEA**L**
G**E**ANVMKA**L**
A**E**LVHFLL**L**
•••

Figure 47. – Frequency (%) of HLA-I alleles of the HLA-I peptides within cluster 6.

Panel A: The plot describes the frequency of the HLA-I alleles of the peptides within cluster 6. A shaded line at 1% was used to determine the HLA-I alleles we considered as our predicted binders. Panel B: a sampled list of random peptides in cluster 6 drawn up, the second and last residue are in bold to highlight the shared residues also shown in Fig. 46 and common with the candidate peptide. Panel C: The logo plot of cluster 6 (of fig 46) is shown side-by-side with the list of peptides.

However, because we wanted to get information on the closest HLA-I-peptides to our candidate peptides we decided to utilize the K-nearest neighbors algorithm with the aim of focusing on the closest 50 peptides (Fig. 48). Doing that, we simply retrieved the HLA-I alleles of these (without relying on frequencies), resulting in table 6. This proved that also based on sequence similarities to eluted peptides, our candidates are still predicted to bind multiple HLA-I alleles. Of note, RIAECILGM's 50 closest peptides were farther than therefore, we excluded it in this analysis (Fig.48).

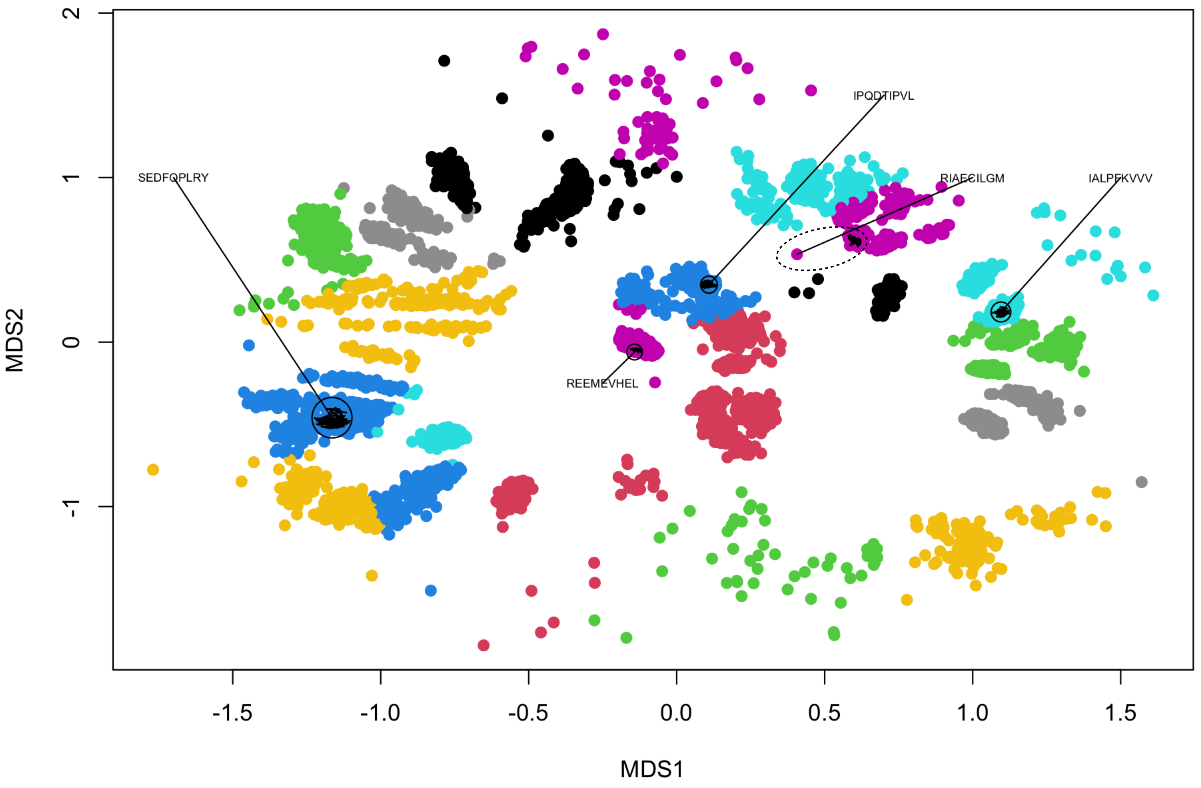


Figure 48. – Visualization of our candidate peptides on the NMDS plot with the K-nearest neighbors selected for the prediction.

This plot shows the locations of our candidate peptides on the NMDS plot, as well as the 50-nearest neighboring peptides (using the K-NN algorithm) in order to predict the alleles that our candidates will bind. Each candidate peptide location within the points is indicated with a segment. The black circles delimit the area from which the nearest neighbors are sampled. Inside the black circles are the neighbors shown in black polygons. All our candidate peptides have their nearest neighbors in proximity. RIAECILGM is far from its nearest neighbors, therefore, we have not included its predictions.

Peptide	HLA-I Alleles of the 50 nearest peptides to the candidate peptide
IALPFKVVV	A0206, A6802, B0801, B5101
SEDFQPLRY	A0101, A2902, A3002, B4402, B4403
IPQDTIPVL	B0702, B0801, B3501, B5101, B5301
REEMEVHEL	B4001, B4403

Tableau 4. – The alleles of the 50-nearest peptides to the candidate peptide

The table shows the alleles from of the 50-nearest HLA-I-peptides to each candidate peptide.

To sum up, the use of peptide sequence similarity between our candidate peptides and the HLA-I peptides of the sequenced immunopeptidomes of the 16 HLA-I alleles, has allowed us to 1) get an unbiased confirmation that gene fusion-derived peptides could bind more than 1 HLA-I allele, 2) obtain further validation on the predicted binding of the candidate peptides to HLA-I alleles.

4.6.2. In-vitro binding results for the cross-HLA-I peptides

4.6.2.1. In vitro binding results

To validate the predictions of the HLA-I alleles, an in-vitro binding assay experiment, was performed on the peptides selected as well as on the controls selected (see Methods section 3.5). The established thresholds for peptide-MHC-I binding are defined as 500nM for strong binders and 2000nM for weak binders and anything above 40000 nM is simply not measured (Alessandro Sette et al., 1994). Notably, to our surprise, the positive control peptide RIAECILGM with HLA-A*02:01 that whose immunogenic potential is documented (Zamora et al., 2019b)had a result of 6722 nM. We reasoned that the low binding affinity does not exclude its immunogenic potential as other factors such as stability of the peptide-HLA-I complex could play a big role (Harndahl et al., 2012). Also, a recent paper showed that a peptide-HLA-I binding affinity could reach up to 10000nM and still yield a positive result by tetramer assays with patient-derived T-cells.

Therefore, measured binding affinities less than 10000nM were not considered as non-binders. Nevertheless, the second control peptide REEMEVHEL of the CBFB-MYH11 bound strongly to B*40:01. This was another attestation to the above.

Despite that, the gene fusion candidates still were found to bind to multiple HLA-I alleles. The peptide IALPFKVVV derived from ATF7IP-PDGFRB bound divergent A & B alleles (A*02:06, B*08:01, B*51:01) very strongly (IC50<<500nM). It also bound weakly A*02:01 (IC50<2000nM) and B*58:01 (IC50<10000nM). On the other hand, the SEDFQPLRY peptide derived from CBFA2T3-GLIS2 fusion bound expected alleles but at variable affinities. However, the binding of SEDFQPLRY to B*44:02 was strong (IC50<500nM), weaker to A*29:02, B*44:03 (IC50<2000) and A*01:01(IC50<10000 nM). Additionally, the gene fusion-derived peptide (IPQDTIPVL) predicted to be presented on multiple alleles of the same supertype (B07) did in fact bind these alleles (B*07:02, B*35:01 and B*53:01). To sum up, FCR helped in predicting HLA-I binders, and gene fusion peptides did bind multiple HLA-I alleles in vitro.

Fusion	**Peptide**	**HLA**	**Predicted by FCR** (Fig.27)	**Fits MS HLA-binding motif** ϒ (visual inspection)	**Predicted by sequence similarity** (section 4.6.1.2)	**Measured binding affinity (nM)**
ATF7IP-PDGFRB	IALPFKVVV	A*02:01	Yes	Partially	No	1590
	IALPFKVVV	A*02:06	Yes	Yes	Yes	72
	IALPFKVVV	A*68 :02	Yes	Yes *	Yes	21825
	IALPFKVVV	B*08:01	No⊥	No	Yes	83
	IALPFKVVV	B*51:01	Yes	Yes	Yes	123
	IALPFKVVV	B*58:01	Yes	Partially	No	5073

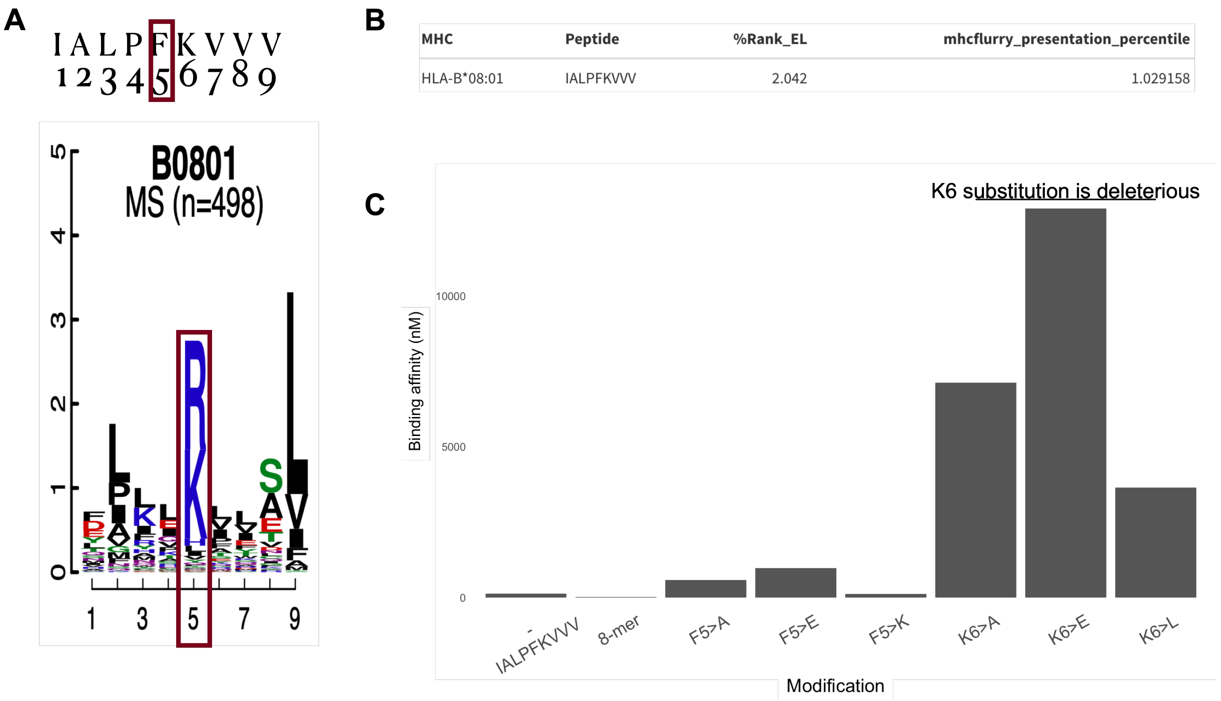
CBFA2T3-GLIS2	SEDFQPLRY	A*01:01	Yes	Yes * (3rd and 9th pos)	Yes	8729
	SEDFQPLRY	A*29:02	Yes	Yes	Yes	1636
	SEDFQPLRY	A*30:02	Yes	Yes	Yes	-
	SEDFQPLRY	B*35:01	Yes	No	No	-
	SEDFQPLRY	B*40:01	Yes	No	No	-
	SEDFQPLRY	B*44:02	Yes	Yes	Yes	1214
	SEDFQPLRY	B*44:03	Yes	Yes	Yes	408
ZNF274-JAK2	IPQDTIPVL	B*07:02	Yes	Yes	Yes	690
	IPQDTIPVL	B*08:01	Yes	No	Yes	9627
	IPQDTIPVL	B*35:01	Yes	Yes	Yes	224
	IPQDTIPVL	B*51:01	Yes	Yes	Yes	25400
	IPQDTIPVL	B*53:01	Yes	Yes	Yes	1067
ETV6-RUNX1	RIAECILGM	A*02:01	Yes	Partially	N/A	6722
	RIAECILGM	A*02:03	Yes	Partially	N/A	2171
	RIAECILGM	A*02:06	Yes	Partially	N/A	1675
CBFB-MYH11	REEMEVHEL	B*40:01	Yes	Yes	Yes	20
	REEMEVHEL	B*44:02	Yes	Yes	No	1840
	REEMEVHEL	B*44:03	Yes	Yes	Yes	16623

Tableau 5. – Peptide-HLA-I binding predictions and binding affinity measurement result.

*The table shows the results of the peptide-HLA-I binding * ambiguous. Ɏmotifs established in (Sarkizova et al., 2020b). Color-coding: green: Measured binding affinity <500n, blue: Measured binding affinity<2000nM, Red: Measured binding affinity<10000nM, Grey: Measured binding affinity<40000nM, White Measured binding affinity not measured (>40000nM). ⊥ predicted only by MHCflurry (peptide-MHC-I prediction software). N/A values for ETV6-RUNX1's predictions by sequence similarity are due to its exclusion because the closest peptides were less similar.*

4.6.2.2. In vitro binding affinity results reveal a potential non canonical binding to HLA-B*08:01

To our surprise, the ATF7IP-PDGFRB peptide (IALPFKVVV) strongly bound the HLA-B*08:01 molecule in vitro despite that 1) it was not predicted by both algorithms (netMHCpan (%RankEL) =2.042, 0.042 value over the %Rank threshold) (Fig. 49B) and 2) the peptide does not fit the defined binding motif, however, 3) it was predicted also by sequence similarity to HLA-I peptides that bound HLA-B*08:01. The HLA-B*08:01 binding motif has an important secondary anchor residue preference at position 5 for a basic amino acid – either an arginine or a lysine (Fig. 49A). In IALPFKVVV, a lysine is at the 6th position (Fig. 49A). To understand the binding of the peptide to the HLA-I allele a series of substitution experiments were performed on the amino acid residues at the 5th and 6th positions of the peptide sequence (Fig. 49C). The 10-fold increase of the binding affinity after the K ->E substitution (basic aa -> acidic aa) has revealed that the K in the 6th position of IALPFKVVV is indispensable for its binding to the B*08:01 molecule. This meant that the B*08:01 molecule is possibly accommodating the peptide by allowing the residue on the 6th position to bind as anchor residue. Taken together, our results for this peptide show on one hand 1) how thresholds could impede the discovery of true peptide-MHC-I binders, 2) binding motifs can also be limiting due to potential non-canonical binding and 3) the sequences of gene fusion peptides could allow them to have distinct conformations structurally.



MHC	Peptide	%Rank_EL	mhcflurry_presentation_percentile
HLA-B*08:01	IALPFKVVV	2.042	1.029158

Figure 49. – Binding motif and predictions comparisons for the IALPFKVVV peptide along with in vitro binding affinity results following a series of substitutions.

*The peptide sequence comparison in panel A with the HLA-B*08:01 binding motif obtained from (Sarkizova et al., 2020a) shows that the 9-mer peptide does not conform with the molecule's preference for a basic residue at position 5. In B) is shown a table of the presentation prediction results displayed in percentile values. The generally agreed percentile threshold is 0.5 for differentiating between strong and weak binders and 2.0 threshold for differentiating between binders and non-binders. The prediction results show that the peptide was not mutually considered as binder by both algorithms. C) represents the measurements of binding affinity of the peptide to the HLA-I molecule after a series of substitutions on the amino acids at the 5th and 6th positions of the peptide.*

4.6.3. In-silico 3D structural modeling of the peptide-HLA complexes

We aimed at understanding the conformation of the CBFA2T3- GLIS2 peptide (SEDFQPLRY) and its binding to the A*01:01 molecule and B*44:03, A*29:02 was not included due to the lack of a suitable PDB template to be used. We chose the CBFA2T3-GLIS2 protein to model its binding with A*01:01 because its measured binding affinity result was higher than the threshold for weak binders (>2000nM). however, it is known that for certain alleles in-vitro binding affinity of a peptide-MHC-I complex less than 10000nM is still capable of binding T cells (Reardon et al., 2021). Therefore, we modeled its binding with the B*44:03 molecule it strongly bound. Our modeling (Fig. 50A) revealed that HLA-B*44:03's B pocket residues E63, K45, Y9 interact with the anchor residue E2 of SEDFQPLRY (fig. 50B) whereas the F pocket residues N77, Y84, D116, K146 interact with Y9 of SEDFQPLRY (Fig. 50C) as similarly shown in the literature for HLA-B*44:03 allele (Rist et al., 2013) As for HLA-A*01:01, we also observed that the interactions between the B & F pockets (Fig. 51A) were the same as described in the literature for HLA-A*01:01 (Giam et al., 2015). The aspartate in third position of SEDFQPLRY acts as anchor residue and interacts with the B pocket (Fig. 51B). On the other hand, Y9 of SEDFQPLRY interacts with the same residues as those in the HLA-A*01:01 molecule (Fig. 51C). This modeling has allowed us to further validate the justified binding of the SEDFQPLRY peptide derived from the CBFA2T3-GLIS2 fusion to two alleles: HLA-A*01:01, HLA-B*44:03.

Fusion	**Peptide sequence**	**HLA**	**HLA-interacting residues**	**Peptide interacting residues**	**Polar Interactions**	**Pockets**	**Binding energies (RosettaMHC)**
CBFA2T3-GLIS2	SEDFQPLRY	HLA-A*01:01	E63, H70, N77, Y84, Y99, D116, K146, W147, R156, Y159	S1, E2, D3, Q5, L7, R8, Y9	S1 -> Y159 E2 -> E63 D3 -> H70, Y99 Q5 -> R156 L7 -> N77 R8 -> N77, W147 Y9 -> N77, Y84, D116, K146	A, B, C, D, F	-59.61971
CBFA2T3-GLIS2	SEDFQPLRY	HLA-B*44:03	Y7, Y9, R62, E63, K45, E76, N77, Y84, D116, K146, W147, Y159	S1, E2, R8, Y9	S1 -> Y7, R62, E63, Y159 E2 -> Y9, K45, R62 R8 -> E76, W147 Y9 -> N77, Y84, D116, K146	A, B, D, F	-59.4299

Tableau 6. – Results of the analyses from the 3D structural modeling

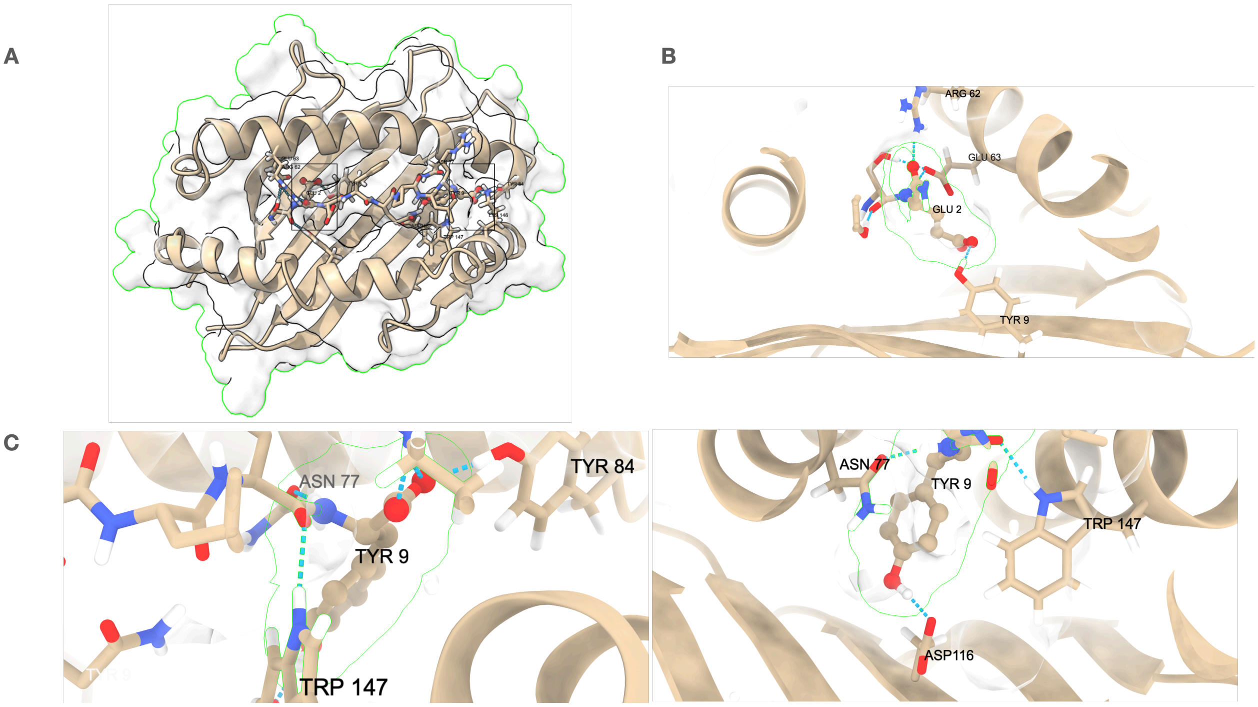


Figure 50. – 3D structural modeling of CBFA2T3-GLIS2 peptide (SEDFQPLRY) with HLA-B*44:03.

*A) represents the top view of 3D structure of the SEDFQPLRY-HLA-B*44:03 complex and displays the polar interactions (hydrogen bonds) between the peptide and the HLA molecule. B) In panel B, the structure was zoomed in on the interaction between the HLA-B*44:03 B pocket and the E2 (glutamine) second amino acid residue in SEDFQPLRY.C) Panel C shows a zoomed in view on the HLA-B*44:03 F pocket residues interacting with the Y9 (Tyrosine) ninth amino acid residue in SEDFQPLRY.*

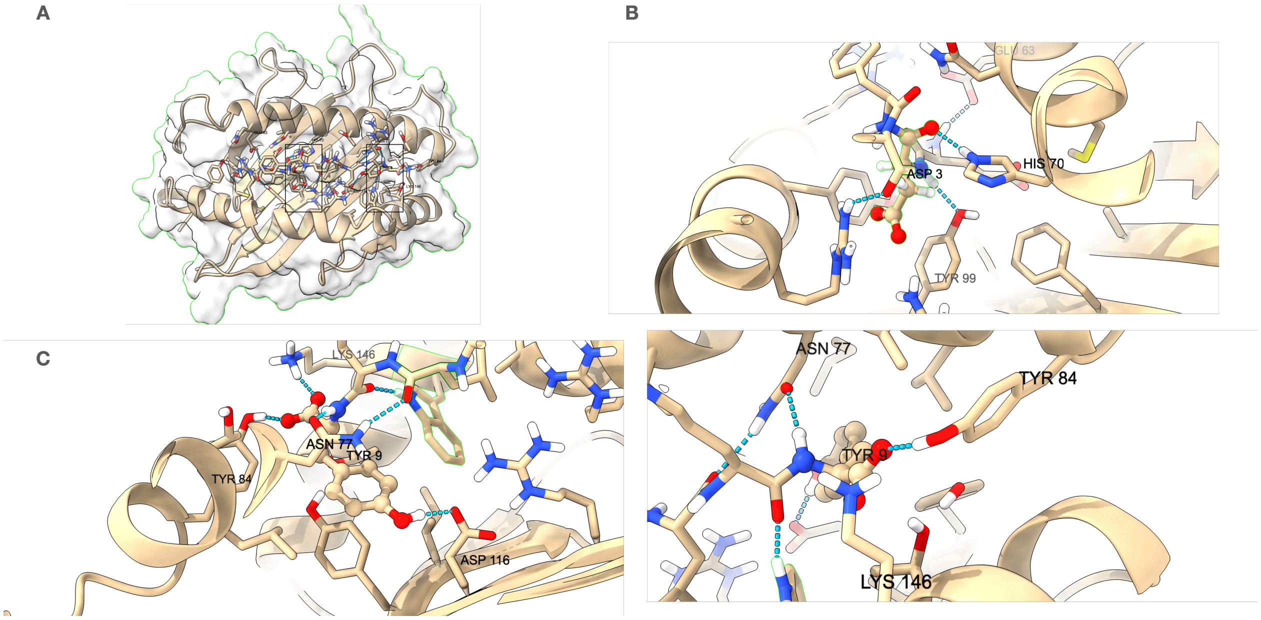


Figure 51. – 3D structural modeling of CBFA2T3-GLIS2 peptide (SEDFQPLRY) with HLA-A*01:01.

*A) represents the top view of 3D structure of the SEDFQPLRY-HLA-A*01:01 complex and displays the polar interactions (hydrogen bonds) between the peptide and the HLA molecule. B) In panel B, the structure was zoomed in on the interaction between the HLA-A*01:01 B pocket and the D3 (Aspartate) second amino acid residue in SEDFQPLRY.C) Panel C shows a zoomed in view on the HLA-A*01:01 F pocket residues interacting with the Y9 (Tyrosine) ninth amino acid residue in SEDFQPLRY.*

Chapter 5 – Discussion

5.1. The identification of hybrid peptides

5.1.1. Proteasomally-spliced peptides

The human mind treats a new idea the same way the body treats a strange protein; it rejects it. – Sir Peter Medawar

It is not without challenges that a new concept would be accepted. Proteasomally-spliced peptides are a "hot topic" in the immunopeptidomic field and an ongoing debate around the identification and proportion of these has taken a toll (Admon, 2021; Endert, 2021; Faridi et al., s. d.; P. M. Kloetzel, 2022; Mishto, 2020, 2021; Purcell, 2021).

5.1.1.1. Bio-informatic identification of spliced

The identification of proteasomally-spliced peptides should rely on ensuring that the peptide spectra are of good quality (such as through a comparison of the observed with the calculated theoretical retention time, precursor error mass distributions) (Lichti et al., 2022; Wen et al., 2020). Some other challenges have been mentioned in the published article (chapter 2) include isobaric amino acids (Isoleucine and Leucine) and modifications. Also, the ALC (average local confidence score) by PEAKS software in denovo analyses could be misleading since as pointed out by Lichti et al (2022) the scores for one fragment of the spliced has high values whereas the second has lower ones. Other issues include the fact that a cis-spliced peptide searched against the proteome database could have more than 1 protein hit, begging the question: is the unique source of a spliced peptide a determining factor? Additionally, in vitro experiments that aimed at elucidating proteasomal splicing rules had findings contradictory with validated spliced hinting at the difficulty of in-vitro experiments to completely replicate the in vivo splicing (Paes et al., 2020). That said, it is also difficult to confirm the exact source of a given peptide as these could originate also from non-coding regions or from frameshifted translations or even from a gene fusion, therefore, a complete proteogenomic approach would be needed to flag such (Cuevas et al., 2021; Erhard et al., 2020).

One of the main insights that have come from the debate include the fact that bio-informatic tools should be treated as guiding truths (Endert, 2021). Therefore, research has to be intensified on these proteasome spliced peptides to better understand them and in turn continuously improve the available tools. Briefly, no matter the algorithm, experimental validation is extremely warranted to confirm identifications; ruling out a genomic source and then validating that the proteasome is responsible for the ligation. Several approaches to validate can be used including the use of proteasome inhibitors, heavy-stable isotope labeling, T-cell assays (Dalet et al., 2010, 2011; Ebstein et al., 2016; Faridi et al., 2020; Hanada et al., 2004; Michaux et al., 2014; Platteel, Liepe, Eden, et al., 2017; Platteel, Liepe, Textoris-Taube, et al., 2017; Vigneron et al., 2004; Warren et al., 2006).

Nevertheless, while writing this manuscript, a very recent study published has investigated spliced peptides using a bioinformatic protocol that resembles RHybridFinder. They have used their method to identify proteasome spliced peptides which they were also able to confirm the source through rigorous experimentation (Kato et al., 2021).

5.1.2. Gene fusion-derived peptides

5.1.2.1. Cross-HLA binding potential of gene fusion peptides

Similarly to proteasomal spliced peptides, gene fusion peptides are also considered hybrid. Gene fusions can generate chimeric proteins which would have novel function and peptides degraded from them. Moreover, the junction region can be a source of neoantigens as has been previously demonstrated with BCR-ABL1, ETV6-RUNX1, CBFB-MYH11 (Biernacki et al., 2020a; M Bocchia et al., 1996; Yotnda, Garcia, et al., 1998; Zamora et al., 2019a).

To predict the ability of GF-Neo to bind multiple alleles whether of the same supertype or of different supertypes, we have firstly developed FCR and then used it to analyze a gene fusion database of cancer cell lines. Our prediction results showed that gene fusions can generate peptides that are predicted to bind HLA-I alleles belonging to one and more than one supertype with a large majority binding alleles belonging to one up to four supertypes (70% of gene fusions). At the extreme, we found 1 gene fusion peptide which was predicted to bind up to 25 HLA-I alleles. Of note is that approximatively 6% of fusions (n=27) were not predicted to generate any HLA-I binders.

Additionally, the analysis of gene fusions involving oncogenes and those identified in cosmic has shown that these gene fusions are predicted to generate cross-HLA-I peptides that are estimated to cover a large population worldwide.

Since we also aimed at reducing the scope of HLA-I alleles to those presented, we focused on HLA-typed cancer cell lines allowed us to see that on average 23% of fusion peptides across cell lines can bind more than 1 HLA alleles of those expressed on average. Nevertheless, it is noteworthy that out of the 100 fusions found in 62 cell lines, we were able to find 86 fusions that are predicted to generate at least 1 peptide that binds 1 HLA-I alleles. On the other hand, 12 fusions do not generate an HLA-I binder (12%).

Furthermore, in vitro validation on a small subset of clinically relevant gene fusion peptides confirmed the utility of FusionChoppeR in prioritizing gene fusion neoantigens. Notably, the in-vitro binding results demonstrated that gene fusion-derived peptides can bind different HLA-I alleles of the same and different supertypes (ZNF274-JAK2), and HLA-I A & B alleles (CBFA2T3-GLIS2, ATF7IP-PDGFRB). HLA-A & HLA-B alleles are considered functionally divergent and to have low peptide binding similarities between HLA-A alleles and HLA-B alleles, as demonstrated by Di et al. (2020).

5.1.2.2. Gene fusion peptides constitute a novel type of HLA-I binders

The ETV6-RUNX1 peptide RIAECILGM which is predicted by netMHCpan to be a weak binder to HLA-A*02:01 but a strong binder by MHCflurry, has also shown weak binding affinity in our in vitro results despite that it has been described to exert immunogenicity (Yotnda, Garcia, et al., 1998; Zamora et al., 2019a). The earlier findings by Yotnda et al (1998) that this peptide is antigenic to HLA-A*02:01 and immunogenic were disputed by Popovic et al (2011) who obtained opposing results from both binding affinity assay, natural processing and T-cell assay (Popovic et al., 2011; Yotnda, Garcia, et al., 1998) The authors claimed that the proteasome cleavage did not generate the RIAECILGM peptide and that original T cell response in Yotnda's work (1998) was due to non-specific reactivity of T cells because they use T cell lines instead of T-cell clones. However, in a seminal paper by Zamora et al. (2019) the RIAECILGM peptide was shown to trigger an immune response from patient derived PBMCs (Zamora et al., 2019a). Therefore, considering the contradictory results, further experiments aimed at confirming whether the peptide is naturally processed and presented are warranted.

Furthermore, in our analyses, the CBFA2T3-GLIS2 peptide with the sequence SEDFQPLRY which is able to bind multiple HLA-I alleles, was a very interesting finding because, 1) it is derived from a gene fusion that is associated with an unfavorable prognosis, 2) it is predicted and able to bind highly divergent alleles and 3) in our clustering analysis, the SEDFQPLRY sequence was predicted to bind these based on its sequence similarity with peptides eluted with these alleles. (Fig. 37 and Table 3). To sum up, this supports the idea that the CBFA2T3-GLIS2 fusion seems to generate a very attractive potential target for testing with immunotherapies such as PCARs.

Another interesting finding was that the gene fusion peptide-HLA IALPFKVVV-HLA-B*08:01 pair was not predicted by both algorithms. Moreover, its sequence does not conform with the HLA binding motif of HLA-B*08:01 (Fig. 38A-B). Much to our surprise, the peptide bound strongly to HLA-B*08:01 in vitro. Substitutions of the residues at the 5th and 6th positions of the peptide revealed that the amino acid at the 6th position within the peptide was responsible for this strong binding which contradictory to the usual peptides bound by the B*08:01 molecule. This meant that structurally it is possible the peptide adopts a specific a conformation to overcome the position of the residue. This has two important implications, 1) the reliance on prediction algorithms could lead to the disposal of valuable neoantigens and 2) peptides derived from fusions have distinct amino acid compositions that probably allow them to circumvent HLA-I binding motifs. Therefore, these possibly constitute a novel type of HLA-I binders.

The only peptide whose predictions and results were consistent with the in-vitro results and the literature was the REEMEVHEL peptide of the CBFB-MYH11 fusion (Biernacki et al., 2020b).

In summary, our results have shown that research is warranted to improve the predictions of fusion peptide HLA-I binders. Researchers generally tend to apply stringent thresholds in order to limit their lists to the strongest binding peptides. This could probably lead to the underestimation of the gene fusions are underestimated as neoantigen sources. Thus, further studies should focus on understanding the antigenic and immunogenic properties of these.

5.2. The Future of hybrid peptides

5.2.1. Proteasome spliced peptides: a proposition for a proof-of-concept

Despite the thoroughness of the experiments aimed at proving the existence/identification of spliced peptides, doubts are still looming in the immunopeptidomic community around the existence of these (Admon, 2021). Thus, alternative methods are warranted. Two important aspects still need to be further explored, the proof-of-concept and the evaluation of in-vivo proteasome processing vs in-vitro. Focusing on the former. Illustrated below is a strategy approach that relies on obtaining proof of splicing through the validation with traditional methods. Essentially, a spliced version of the thoroughly and widely characterized ovalbumin epitope is synthesized and then digested by the proteasome. If the proteasome does indeed splice peptides, then, the canonical ovalbumin epitope should be generated by the proteasome. Testing that the proteasome has generated the canonical peptide from its spliced version can be easily achieved with SIINFEKL-specific CD8+ T cells as well as with mass spectrometry.

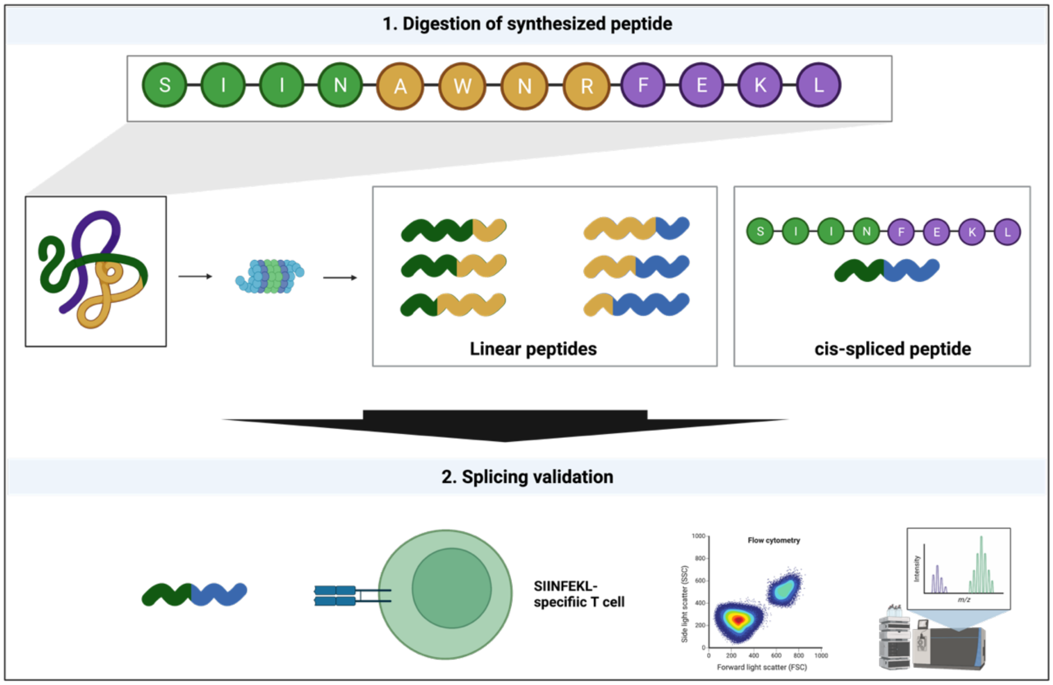


Figure 52. – Proposed proof-of-concept experiment.

Additional methodologies can be incorporated to this approach such as proteasome inhibitors and heavy-stable labeling. That is, it would also be possible to devise an experiment where the green and purple fragments have been labeled, if proteasomal splicing should occur then, the resulting peptide will be doubly labeled, as has been studied before by Berkers et al. (Berkers et al., 2015). Furthermore, the use of proteasome inhibitors would help validate that the splicing is catalyzed as has been used in the literature(Dalet et al., 2011; Ebstein et al., 2016; Hanada et al., 2004; Michaux et al., 2014; Vigneron et al., 2004).

This approach has some visible issues such as the intervening fragment to be used (here -AWNR-, based on the heavily tested gp100 spliced peptide **RTK**AWNR**QLYPEW**) its length, its amino acid composition among others. Nevertheless, the ability to demonstrate proteasome splicing through obtaining a canonical product could be a game-changer and the start of a new era for proteasome peptide splicing.

5.2.1. Gene fusion-derived peptides

5.2.1.1. The prospect of discovering cross-HLA-I binding peptides

Recently, a peptide-centric CAR (PCAR) has been developed for targeting a neuroblastoma PHOX2B-derived neoantigen (Yarmarkovich et al., 2021). They were also able to successfully test for its ability to target the peptide while being presented on HLA-A and HLA-B alleles. The authors tested the cross-reactivity of the PCAR construct, they found that it was able to specifically target the PHOX2B peptide presented on the HLA-I alleles in comparison with other peptides. Given the results shown in this manuscript it would be reasonable the prospect of applying PCARs for treating pediatric tumours. FCR was developed in order to enable researchers to apply it to fusions and facilitate the discovery and selection of neoantigens. The application of PCARs for target gene fusion peptides could open the possibility of having off-the shelf treatment for pediatric cancers.

5.2.1.2. Immunoopeptidomic studies on cancer cell lines harboring gene fusions

The analysis of the LigeA database of gene fusions in cancer cell lines demonstrates the need for mass immunopeptidomic experiments on cell lines. The neoantigenic potential of the different gene fusions (in-frame, out-of-frame) can have a great impact on the development of fusion neoantigen centric therapies, as well as on our understanding of how these hybrid peptides interact with the degradation machinery of the cell. Finally, in order to contribute our analyses to the research community, we provide a database with the FusionChoppeR tool that would allow to have access to all the results of neoantigen predictions computed on the cell lines.

Conclusion

In this project, we have emphasized the relevance of the hybrid immunopeptidome for the development of immunotherapies. Also, we have introduced two new tools that we have developed for probing these computationally.

We have also shown that gene fusions can generate peptides that bind multiple HLA-I alleles and that gene fusions can be heterogeneous therefore they generate attractive targets. We argued the reliance on predictions and demonstrated need for improving peptide-MHC-I predictions for gene fusion peptides. We have also shared a database of gene fusion derived peptides of cancer cell lines with the research community.

Additionally, we have proposed a workflow for the validation of proteasome-catalyzed spliced peptides (presented in the published manuscript) as well as a novel strategy for the validation of proteasomal splicing which could also impact future research.

Finally, the hybrid immunopeptidome's shareability among individuals could constitute a novel target for an "off-the-shelf" form of immunotherapy.

Annex

HLA-A	Supertype
A*01:01	A01
A*02:01	A02
A*02:03	A02
A*02:06	A02
A*03:01	A03
A*11:01	A03
A*23:01	A24
A*24:02	A24
A*26:01	A01
A*29:02	A01A24
A*30:01	A01A03
A*30:02	A01
A*33:01	A03
A*68:01	A03
A*68:02	A02

HLA-B	Supertype
B*07:02	B07
B*08:01	B08
B*15:01	B62
B*35:01	B07
B*40:01	B44
B*44:02	B44
B*44:03	B44
B*51:01	B07
B*53:01	B07
B*57:01	B58
B*58:01	B58

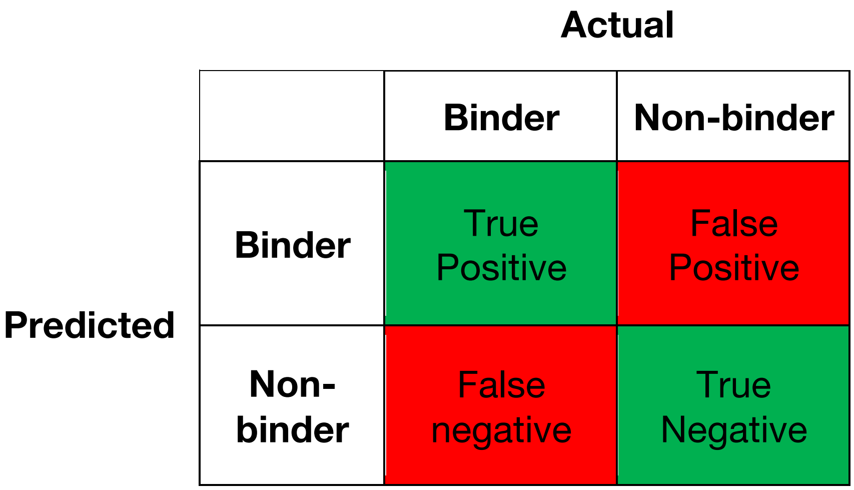
HLA-C	Supertype
C*03:02	C1
C*03:03	C1
C*03:04	C1
C*04:01	C4
C*06:02	C4
C*07:02	C1
C*08:01	C1
C*17:01	C4

Tableau 1. – (Annex) 35 most common HLA-I (A/B/C) alleles selected for the predictions and the supertype they belong to.

Fusion	**Peptide Sequence**	**HLA-I predicted to be bound**	**# of HLA-I**	**Supertypes that the HLA-I alleles belong to**	**# of Supertypes**
ATF7IP-PDGFRB	IALPFKVVV	A*02:01, A*02:06, A*68:02, B*51:01, B*58:01, _C*03:02, C*03:03, C*03:04, C*06:02, C*08:01, C*17:01_	11	A02, B07, B58, _C1, C2_	5
CBFA2T3-GLIS2	SEDFQPLRY	A*01:01, A*29:02, A*30:02, B*35:01, B*40:01, B*44:02, B*44:03, C*04:01, C*07:02, C*18:02	10	A01, A01 A24, B07, B44, _C2, C1_	6
ZNF274-JAK2	IPQDTIPVL	B*07:02, B*08:01, B*35:01, B*51:01, B*53:01, _C*03:02, C*03:03, C*03:04, C*04:01, C*07:02, C*08:01, C*17:01, C*18:02_	13	B07, B08, C1, _C2_	4
ETV6-RUNX1	RIAECILGM	A*02:01, A*02:03, A*02:06, _C*17:01_	4	A02, _C2_	2
CBFB-MYH11	REEMEVHEL	B*40:01, B*44:02, B*44:03, _C*04:01, C*18:02_	5	B44, _C2_	2

Tableau 2. – (Annex) Candidate gene fusion peptides selected for in vitro binding assay along with their predictions.

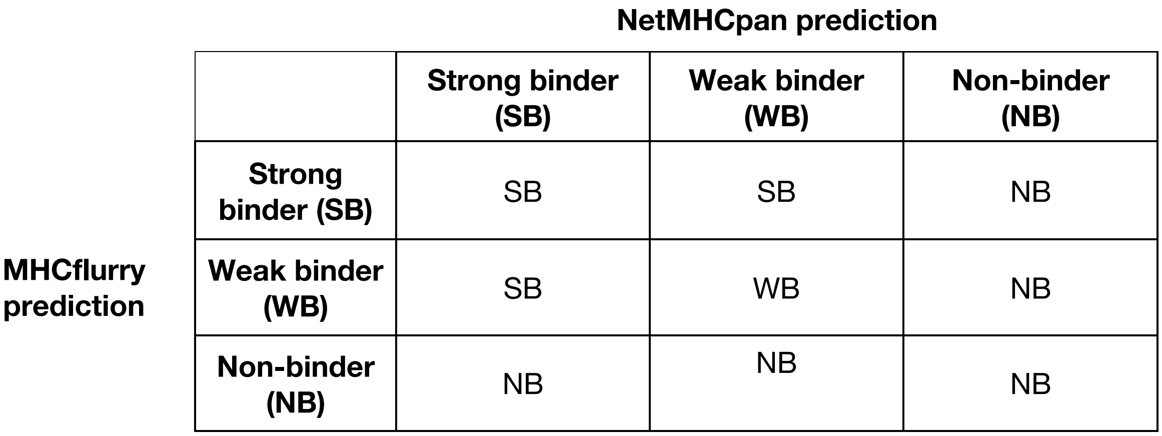
The peptides selected for in vitro binding assay. The HLA-C alleles are in italic and lower font because they were not tested.



		Actual	
		Binder	Non-binder
Predicted	Binder	True Positive	False Positive
	Non-binder	False negative	True Negative

Tableau 3. – (Annex) Confusion matrix of possible results of predicted vs. actual result of peptide-MHC -I binding.

The table presents all possible scenarios of comparisons between predictions and results. If a peptide is predicted to be binder to a MHC-I, in actual results it can either be binder in which case it would be a true positive or a non-binder in which case it would be a False positive (predicted to be binder but is not actually). On the other hand, If a peptide is predicted to be non-binder to a MHC-I, and in actual results it is a binder then this means it is a false negative however if it is non-binder then it is considered a true negative.



		NetMHCpan prediction		
		Strong binder (SB)	Weak binder (WB)	Non-binder (NB)
MHCflurry prediction	Strong binder (SB)	SB	SB	NB
	Weak binder (WB)	SB	WB	NB
	Non-binder (NB)	NB	NB	NB

Tableau 4. – (Annex) Confusion matrix of netMHCpan and MHCflurry prediction results that explains the formula Strong binder ratio in section 3.1.2.1.

The table presents all possible scenarios of comparisons between peptide-MHC-I predictions of both algorithms: netMHCpan and MHCflurry. In that formula, a peptide was considered as SB if at least one algorithm prediction deemed it as SB and the other as SB/WB. On the other hand, WB is considered so if both algorithms deem the peptide-MHC-I complex as WB. Finally, any NB led to being deemed as non-binder and therefore not considered.

Peptide	Cluster	Alleles of the peptides from the same cluster (>1%) *
IALPFKVVV	3	A0203, A0206, A6802, B0702, B0801, B5101
SEDFQPLRY	5	A0101, A2902, A3002, B3501, B4402, B4403, B5801
IPQDTIPVL	11	A0206, A6802, B0702, B0801, B3501, B4001, B5101, B5301, B5801
REEMEVHEL	21	B4001, B4402

Tableau 5. – (Annex) The alleles presented by the peptides of each cluster.

*The table shows for the peptides selected, the alleles of the peptides that within that cluster. *Only alleles of peptides representing a percentage greater than 1 % were considered. This table also shown in barplot format in Annex Figure 1.*

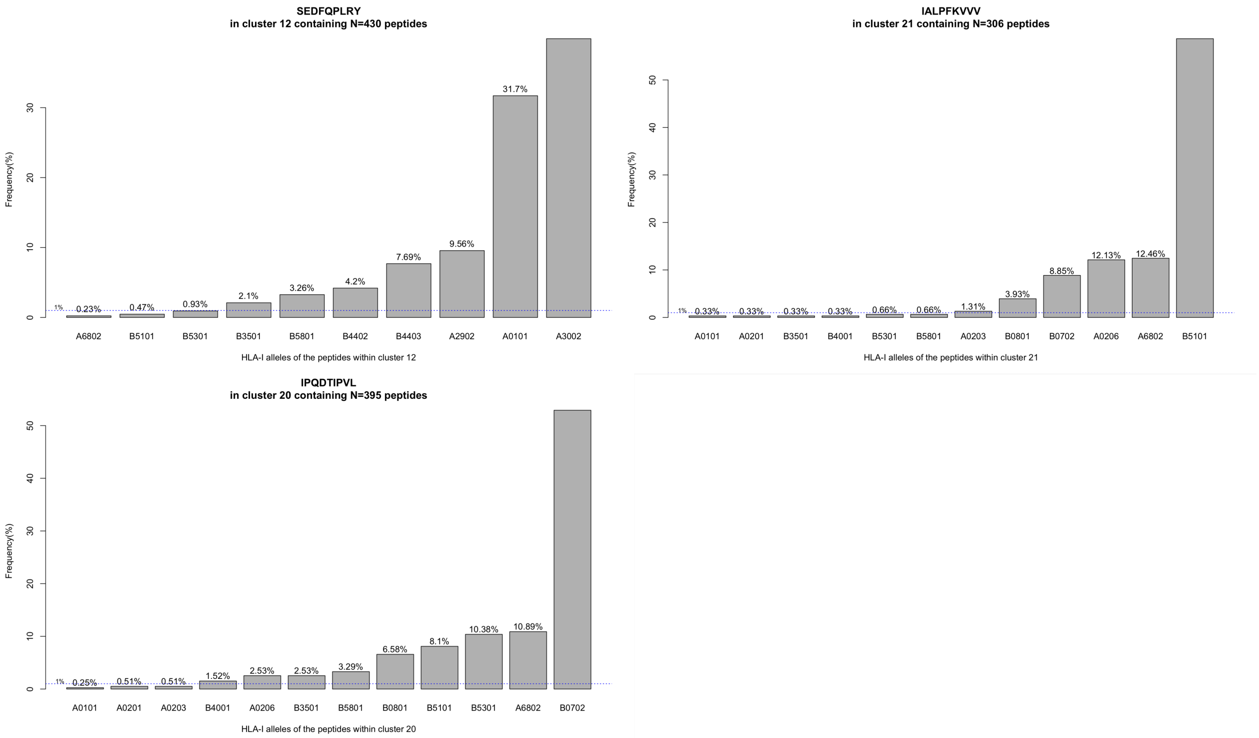


Figure 1. – (Annex) Barplots of the HLA-I allele frequencies (%) within the clusters of our candidate peptides

Each one of these barplots represents the HLA-I allele frequencies within the clusters of our candidate peptides.

References

Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C. H., Frattini, V., Lasorella, A., Iavarone, A., Inghirami, G. and Rabadan, R. (2014). Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Systems Biology*, *8*(1), 97. https://doi.org/10.1186/s12918-014-0097-z

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A. and Wu, C. J. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*, *46*(2), 315–326. https://doi.org/10.1016/j.immuni.2017.02.007

Admon, A. (2021). Are There Indeed Spliced Peptides in the Immunopeptidome? *Molecular & Cellular Proteomics*, *20*, 100099. https://doi.org/10.1016/j.mcpro.2021.100099

Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M. and Nielsen, M. (2019). NNAlign_MA; MHC Peptidome Deconvolution for Accurate MHC Binding Motif Characterization and Improved T-cell Epitope Predictions. *Molecular & Cellular Proteomics*, *18*(12), 2459–2477. https://doi.org/10.1074/mcp.tir119.001658

Barber, L. D., Castro, B. G.-, Percival, L., Li, X., Clayberger, C. and Parham, P. (1995). Overlap in the repertoires of peptides bound in vivo by a group of related class I HLA-B allotypes. *Current Biology*, *5*(2), 179–190. https://doi.org/10.1016/s0960-9822(95)00039-x

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G. and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Computational Biology*, *13*(8), e1005725. https://doi.org/10.1371/journal.pcbi.1005725

Berkers, C. R., Jong, A. de, Schuurman, K. G., Linnemann, C., Meiring, H. D., Janssen, L., Neefjes, J. J., Schumacher, T. N. M., Rodenko, B. and Ovaa, H. (2015). Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. *The Journal of Immunology*, *195*(9), 4085–4095. https://doi.org/10.4049/jimmunol.1402455

Biernacki, M. A. and Bleakley, M. (2020). Neoantigens in Hematologic Malignancies. *Frontiers in Immunology*, *11*, 121. https://doi.org/10.3389/fimmu.2020.00121

Biernacki, M. A., Foster, K. A., Woodward, K. B., Coon, M. E., Cummings, C., Cunningham, T. M., Dossa, R. G., Brault, M., Stokke, J., Olsen, T. M., Gardner, K., Estey, E., Meshinchi, S., Rongvaux, A. and Bleakley, M. (2020a). CBFB-MYH11 fusion neoantigen enables T cell recognition and killing of acute myeloid leukemia. *Journal of Clinical Investigation*, *130*(10), 5127–5141. https://doi.org/10.1172/jci137723

Biernacki, M. A., Foster, K. A., Woodward, K. B., Coon, M. E., Cummings, C., Cunningham, T. M., Dossa, R. G., Brault, M., Stokke, J., Olsen, T. M., Gardner, K., Estey, E., Meshinchi, S., Rongvaux, A. and Bleakley, M. (2020b). CBFB-MYH11 fusion neoantigen enables T cell recognition and killing of acute myeloid leukemia. *Journal of Clinical Investigation*, *130*(10), 5127–5141. https://doi.org/10.1172/jci137723

BILLINGHAM, R. E., BRENT, L. and MEDAWAR, P. B. (1953). 'Actively Acquired Tolerance' of Foreign Cells. *Nature*, *172*(4379), 603–606. https://doi.org/10.1038/172603a0

Bocchia, M, Gentili, S., Abruzzese, E., Fanelli, A., Iuliano, F., Tabilio, A., Amabile, M., Forconi, F., Gozzetti, A., Raspadori, D., Amadori, S. and Lauria, F. (2005). Effect of a p210 multipeptide vaccine associated with imatinib or interferon in patients with chronic myeloid leukaemia and persistent residual disease: a multicentre observational trial. *The Lancet*, *365*(9460), 657–662. https://doi.org/10.1016/s0140-6736(05)17945-8

Bocchia, M, Korontsvit, T., Xu, Q., Mackinnon, S., Yang, S. Y., Sette, A. and Scheinberg, D. A. (1996). Specific human cellular immunity to bcr-abl oncogene-derived peptides. *Blood*, *87*(9), 3587–92.

Bocchia, Monica, Defina, M., Aprile, L., Ippoliti, M., Crupi, R., Rondoni, M., Gozzetti, A. and Lauria, F. (2010). Complete molecular response in CML after p210 BCR–ABL1-derived peptide vaccination. *Nature Reviews Clinical Oncology*, *7*(10), 600–603. https://doi.org/10.1038/nrclinonc.2010.141

Bolouri, H., Farrar, J. E., Triche, T., Ries, R. E., Lim, E. L., Alonzo, T. A., Ma, Y., Moore, R., Mungall, A. J., Marra, M. A., Zhang, J., Ma, X., Liu, Y., Liu, Y., Auvil, J. M. G., Davidsen, T. M., Gesuwan, P., Hermida, L. C., Salhia, B., … Meshinchi, S. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nature Medicine*, *24*(1), 103–112. https://doi.org/10.1038/nm.4439

Borcard, D., Gillet, F. and Legendre, P. (2018). Numerical Ecology with R. *Use R!* https://doi.org/10.1007/978-3-319-71404-2

Buhler, S., Nunes, J. M. and Sanchez-Mazas, A. (2016). HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*, *68*(6–7), 401–416. https://doi.org/10.1007/s00251-016-0918-x

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L. D., Christie, C., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., Liang, Y., Lowe, R., Peisach, E., … Ioannidis, Y. E. (2018). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, *47*(D1), D520–D528. https://doi.org/10.1093/nar/gky949

Cardinaud, S., Consiglieri, G., Bouziat, R., Urrutia, A., Graff-Dubois, S., Fourati, S., Malet, I., Guergnon, J., Guihot, A., Katlama, C., Autran, B., Endert, P. van, Lemonnier, F. A., Appay, V., Schwartz, O., Kloetzel, P. M. and Moris, A. (2011). CTL Escape Mediated by Proteasomal Destruction of an HIV-1 Cryptic Epitope. *PLoS Pathogens*, *7*(5), e1002049. https://doi.org/10.1371/journal.ppat.1002049

Carluccio, A. R. D., Triffon, C. F. and Chen, W. (2018). Perpetual complexity: predicting human CD8+ T-cell responses to pathogenic peptides. *Immunology and Cell Biology*, *96*(4), 358–369. https://doi.org/10.1111/imcb.12019

Caron, E., Vincent, K., Fortier, M.-H., Laverdure, J.-P., Bramoullé, A., Hardy, M.-P., Voisin, G., Roux, P. P., Lemieux, S., Thibault, P. and Perreault, C. (2011). The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Molecular Systems Biology*, *7*(1), 533–533. https://doi.org/10.1038/msb.2011.68

Carreno, B. M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A. A., Ly, A., Lie, W.-R., Hildebrand, W. H., Mardis, E. R. and Linette, G. P. (2015). A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science*, *348*(6236), 803–808. https://doi.org/10.1126/science.aaa3828

Chandran, S. S., Ma, J., Klatt, M. G., Dündar, F., Bandlamudi, C., Razavi, P., Wen, H. Y., Weigelt, B., Zumbo, P., Fu, S. N., Banks, L. B., Yi, F., Vercher, E., Etxeberria, I., Bestman, W. D., Paula, A. D. C., Aricescu, I. S., Drilon, A., Betel, D., … Klebanoff, C. A. (2022). Immunogenicity and therapeutic targeting of a public neoantigen derived from mutated PIK3CA. *Nature Medicine*, *28*(5), 946–957. https://doi.org/10.1038/s41591-022-01786-3

Chang, F., Lin, F., Cao, K., Surrey, L. F., Aplenc, R., Bagatell, R., Resnick, A. C., Santi, M., Storm, P. B., Tasian, S. K., Waanders, A. J., Hunger, S. P. and Li, M. M. (2019a). Development and Clinical Validation of a Large Fusion Gene Panel for Pediatric Cancers. *The Journal of Molecular Diagnostics*, *21*(5), 873–883. https://doi.org/10.1016/j.jmoldx.2019.05.006

Chang, F., Lin, F., Cao, K., Surrey, L. F., Aplenc, R., Bagatell, R., Resnick, A. C., Santi, M., Storm, P. B., Tasian, S. K., Waanders, A. J., Hunger, S. P. and Li, M. M. (2019b). Development and Clinical Validation of a Large Fusion Gene Panel for Pediatric Cancers. *The Journal of Molecular Diagnostics*, *21*(5), 873–883. https://doi.org/10.1016/j.jmoldx.2019.05.006

Chaudhury, S., Lyskov, S. and Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, *26*(5), 689–691. https://doi.org/10.1093/bioinformatics/btq007

Chelvanayagam, G. (1996). A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics*, *45*(1), 15–26. https://doi.org/10.1007/s002510050162

Chen, C., Liu, S., Qu, R. and Li, B. (2020). Recurrent Neoantigens in Colorectal Cancer as Potential Immunotherapy Targets. *BioMed Research International*, *2020*, 2861240. https://doi.org/10.1155/2020/2861240

Choo, S. Y. (2007). The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Medical Journal*, *48*(1), 11–23. https://doi.org/10.3349/ymj.2007.48.1.11

Chowell, D., Krishna, C., Pierini, F., Makarov, V., Rizvi, N. A., Kuo, F., Morris, L. G. T., Riaz, N., Lenz, T. L. and Chan, T. A. (2019). Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nature Medicine*, *25*(11), 1715–1720. https://doi.org/10.1038/s41591-019-0639-4

Chowell, D., Morris, L. G. T., Grigg, C. M., Weber, J. K., Samstein, R. M., Makarov, V., Kuo, F., Kendall, S. M., Requena, D., Riaz, N., Greenbaum, B., Carroll, J., Garon, E., Hyman, D. M., Zehir, A., Solit, D., Berger, M., Zhou, R., Rizvi, N. A. and Chan, T. A. (2018). Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*, *359*(6375), 582–587. https://doi.org/10.1126/science.aao4572

Consortium, T. U., Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. D., Denny, P., Dogan, T., Ebenezer, T., Fan, J., … Teodoro, D. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Cuevas, M. V. R., Hardy, M.-P., Hollý, J., Bonneil, É., Durette, C., Courcelles, M., Lanoix, J., Côté, C., Staudt, L. M., Lemieux, S., Thibault, P., Perreault, C. and Yewdell, J. W. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Reports*, *34*(10), 108815. https://doi.org/10.1016/j.celrep.2021.108815

Dalet, A., Robbins, P. F., Stroobant, V., Vigneron, N., Li, Y. F., El-Gamil, M., Hanada, K., Yang, J. C., Rosenberg, S. A. and Eynde, B. J. V. den. (2011). An antigenic peptide produced by reverse splicing and double asparagine deamidation. *Proceedings of the National Academy of Sciences*, *108*(29), E323–E331. https://doi.org/10.1073/pnas.1101892108

Dalet, A., Vigneron, N., Stroobant, V., Hanada, K. and Eynde, B. J. V. den. (2010). Splicing of Distant Peptide Fragments Occurs in the Proteasome by Transpeptidation and Produces the

Spliced Antigenic Peptide Derived from Fibroblast Growth Factor-5. *The Journal of Immunology*, *184*(6), 3016–3024. https://doi.org/10.4049/jimmunol.0901277

Dawson, D. V., Ozgur, M., Sari, K., Ghanayem, M. and Kostyu, D. D. (2001). Ramifications of HLA class I polymorphism and population genetics for vaccine development. *Genetic Epidemiology*, *20*(1), 87–106. https://doi.org/10.1002/1098-2272(200101)20:1<87::aid-gepi8>3.0.co;2-r

Dersh, D., Hollý, J. and Yewdell, J. W. (2020). Author Correction: A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion. *Nature Reviews Immunology*, *20*(10), 644–644. https://doi.org/10.1038/s41577-020-00445-3

Doytchinova, I. A., Guan, P. and Flower, D. R. (2004). Identifiying Human MHC Supertypes Using Bioinformatic Methods. *The Journal of Immunology*, *172*(7), 4314–4323. https://doi.org/10.4049/jimmunol.172.7.4314

Doytchinova, I. and Flower, D. (2003). The HLA-A2-supermotif: a QSAR definition. *Organic & Biomolecular Chemistry*, *1*(15), 2648–2654. https://doi.org/10.1039/b300707c

Draenert, R., Gall, S. L., Pfafferott, K. J., Leslie, A. J., Chetty, P., Brander, C., Holmes, E. C., Chang, S.-C., Feeney, M. E., Addo, M. M., Ruiz, L., Ramduth, D., Jeena, P., Altfeld, M., Thomas, S., Tang, Y., Verrill, C. L., Dixon, C., Prado, J. G., … Goulder, P. J. R. (2004). Immune Selection for Altered Antigen Processing Leads to Cytotoxic T Lymphocyte Escape in Chronic HIV-1 Infection. *The Journal of Experimental Medicine*, *199*(7), 905–915. https://doi.org/10.1084/jem.20031982

Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P. R., Bellier, E., Bivand, R., Blanchet, F. G., Cáceres, M. D., Dufour, A.-B., Heegaard, E., Jombart, T., Munoz, F., Oksanen, J., Thioulouse, J. and Wagner, H. H. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, *82*(3), 257–275. https://doi.org/10.1890/11-1183.1

Dupain, C., Harttrampf, A. C., Urbinati, G., Geoerger, B. and Massaad-Massade, L. (2017). Relevance of Fusion Genes in Pediatric Cancers: Toward Precision Medicine. *Molecular Therapy. Nucleic Acids*, *6*, 315–326. https://doi.org/10.1016/j.omtn.2017.01.005

Ebstein, F., Textoris-Taube, K., Keller, C., Golnik, R., Vigneron, N., Eynde, B. J. V. den, Schuler-Thurner, B., Schadendorf, D., Lorenz, F. K. M., Uckert, W., Urban, S., Lehmann, A., Albrecht-Koepke, N., Janek, K., Henklein, P., Niewienda, A., Kloetzel, P. M. and Mishto, M. (2016). Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. *Scientific Reports*, *6*, 24032. https://doi.org/10.1038/srep24032

Endert, P. van. (2021). Beware the algorithm. *eLife*, *10*, e69657. https://doi.org/10.7554/elife.69657

Erhard, F., Dölken, L., Schilling, B. and Schlosser, A. (2020). Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunology Research*, *8*(8), 1018–1026. https://doi.org/10.1158/2326-6066.cir-19-0886

Eshof, B. L. van den, Medfai, L., Nolfi, E., Wawrzyniuk, M. and Sijts, A. J. A. M. (2021). The Function of Immunoproteasomes-An Immunologists' Perspective. *Cells*, *10*(12), 3360. https://doi.org/10.3390/cells10123360

Faridi, P., Dorvash, M. and Purcell, A. (s. d.). Spliced HLA bound peptides; a Black-Swan event in Immunology. https://doi.org/10.22541/au.160976778.87618567/v1

Faridi, P., Li, C., Ramarathinam, S. H., Vivian, J. P., Illing, P. T., Mifsud, N. A., Ayala, R., Song, J., Gearing, L. J., Hertzog, P. J., Ternette, N., Rossjohn, J., Croft, N. P. and Purcell, A. W. (2018). A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Science Immunology*, *3*(28), eaar3947. https://doi.org/10.1126/sciimmunol.aar3947

Faridi, P., Woods, K., Ostrouska, S., Deceneux, C., Aranha, R., Duscharla, D., Wong, S. Q., Chen, W., Ramarathinam, S. H., Sian, T. C. C. L. K., Croft, N. P., Li, C., Ayala, R., Cebon, J. S., Purcell, A. W., Schittenhelm, R. B. and Behren, A. (2020). Spliced Peptides and Cytokine-Driven Changes in the Immunopeptidome of Melanoma. *Cancer Immunology Research*, *8*(10), 1322–1334. https://doi.org/10.1158/2326-6066.cir-19-0894

Gambacorti-Passerini, C., Bertazzoli, C., Dermime, S., Scardino, A., Schendel, D. and Parmiani, G. (1997). Mapping of HLA class I binding motifs in forty-four fusion proteins involved in human cancers. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *3*(5), 675–83.

Giam, K., Ayala-Perez, R., Illing, P. T., Schittenhelm, R. B., Croft, N. P., Purcell, A. W. and Dudek, N. L. (2015). A comprehensive analysis of peptides presented by HLA-A1. *Tissue Antigens*, *85*(6), 492–496. https://doi.org/10.1111/tan.12565

Gioiosa, S., Bolis, M., Flati, T., Massini, A., Garattini, E., Chillemi, G., Fratelli, M. and Castrignanò, T. (2018). Massive NGS data analysis reveals hundreds of potential novel gene fusions in human cell lines. *GigaScience*, *7*(10), giy062. https://doi.org/10.1093/gigascience/giy062

Gruber, T. A., Larson Gedman, A., Zhang, J., Koss, C. S., Marada, S., Ta, H. Q., Chen, S.-C., Su, X., Ogden, S. K., Dang, J., Wu, G., Gupta, V., Andersson, A. K., Pounds, S., Shi, L., Easton, J., Barbato, M. I., Mulder, H. L., Manne, J., … Downing, J. R. (2012). An Inv(16)(p13.3q24.3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute Megakaryoblastic Leukemia. *Cancer Cell*, *22*(5), 683–697. https://doi.org/10.1016/j.ccr.2012.10.007

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N. and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biology*, *20*(1), 213. https://doi.org/10.1186/s13059-019-1842-9

Hamilton, S. N., Carlson, R., Hasan, H., Rassekh, S. R. and Goddard, K. (2017). Long-term Outcomes and Complications in Pediatric Ewing Sarcoma. *American Journal of Clinical Oncology*, *40*(4), 423–428. https://doi.org/10.1097/coc.0000000000000176

Hanada, K., Yewdell, J. W. and Yang, J. C. (2004). Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature*, *427*(6971), 252–256. https://doi.org/10.1038/nature02240

Harndahl, M., Rasmussen, M., Roder, G., Pedersen, I. D., Sørensen, M., Nielsen, M. and Buus, S. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *European Journal of Immunology*, *42*(6), 1405–1416. https://doi.org/10.1002/eji.201141774

Hertz, T. and Yanover, C. (2007). Identifying HLA supertypes by learning distance functions. *Bioinformatics (Oxford, England)*, *23*(2), e148-55. https://doi.org/10.1093/bioinformatics/btl324

Hewitt, E. W. (2003). The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, *110*(2), 163–169. https://doi.org/10.1046/j.1365-2567.2003.01738.x

Istrail, S., Florea, L., Halldórsson, B. V., Kohlbacher, O., Schwartz, R. S., Yap, V. B., Yewdell, J. W. and Hoffman, S. L. (2004). Comparative immunopeptidomics of humans and their pathogens. *Proceedings of the National Academy of Sciences*, *101*(36), 13268–13272. https://doi.org/10.1073/pnas.0404740101

Juji, T. (1988). HLA in Narcolepsy, 10–23. https://doi.org/10.1007/978-3-642-83387-8_2

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. *The Journal of Immunology*, *199*(9), 3360–3368. https://doi.org/10.4049/jimmunol.1700893

Kato, K., Nakatsugawa, M., Tokita, S., Hirohashi, Y., Kubo, T., Tsukahara, T., Murata, K., Chiba, H., Takahashi, H., Hirano, N., Kanaseki, T. and Torigoe, T. (2021). Characterization of Proteasome-Generated Spliced Peptides Detected by Mass Spectrometry. *Journal of immunology (Baltimore, Md. : 1950)*. https://doi.org/10.4049/jimmunol.2100717

Kaufman, J. (2018). Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. *Trends in Immunology*, *39*(5), 367–379. https://doi.org/10.1016/j.it.2018.01.001

Kim, Y., Sidney, J., Pinilla, C., Sette, A. and Peters, B. (2009). Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC bioinformatics*, *10*(1), 394. https://doi.org/10.1186/1471-2105-10-394

Kloetzel, P. M. (2022). Neo-Splicetopes in Tumor Therapy: A Lost Case? *Frontiers in Immunology*, *13*, 849863. https://doi.org/10.3389/fimmu.2022.849863

Kloetzel, P.-M. (2001). Antigen processing by the proteasome. *Nature Reviews Molecular Cell Biology*, *2*(3), 179–188. https://doi.org/10.1038/35056572

Kubiniok, P., Marcu, A., Bichmann, L., Kuchenbecker, L., Schuster, H., Hamelin, D. J., Duquette, J. D., Kovalchik, K. A., Wessling, L., Kohlbacher, O., Rammensee, H.-G., Neidert, M. C., Sirois, I. and Caron, E. (2022). Understanding the constitutive presentation of MHC class I immunopeptidomes in primary tissues. *iScience*, *25*(2), 103768. https://doi.org/10.1016/j.isci.2022.103768

Kubo, R. T., Sette, A., Grey, H. M., Appella, E., Sakaguchi, K., Zhu, N. Z., Arnott, D., Sherman, N., Shabanowitz, J. and Michel, H. (1994). Definition of specific peptide motifs for four major HLA-A alleles. *Journal of immunology (Baltimore, Md. : 1950)*, *152*(8), 3913–24.

LaHaye, S., Fitch, J. R., Voytovich, K. J., Herman, A. C., Kelly, B. J., Lammi, G. E., Arbesfeld, J. A., Wijeratne, S., Franklin, S. J., Schieffer, K. M., Bir, N., McGrath, S. D., Miller, A. R., Wetzel, A., Miller, K. E., Bedrosian, T. A., Leraas, K., Varga, E. A., Lee, K., … White, P. (2021). Discovery of clinically relevant fusions in pediatric cancer. *BMC Genomics*, *22*(1), 872. https://doi.org/10.1186/s12864-021-08094-z

Latysheva, N. S. and Babu, M. M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research*, *44*(10), 4487–4503. https://doi.org/10.1093/nar/gkw282

Legendre, P. (2014). Interpreting the replacement and richness difference components of beta diversity: Replacement and richness difference components. *Global Ecology and Biogeography*, *23*(11), 1324–1334. https://doi.org/10.1111/geb.12207

Lichti, C. F., Vigneron, N., Clauser, K. R., Eynde, B. J. V. den and Bassani-Sternberg, M. (2022). Navigating Critical Challenges Associated with Immunopeptidomics-Based Detection of Proteasomal Spliced Peptide Candidates. *Cancer Immunology Research*, *10*(3), 275–284. https://doi.org/10.1158/2326-6066.cir-21-0727

Liepe, J., Marino, F., Sidney, J., Jeko, A., Bunting, D. E., Sette, A., Kloetzel, P. M., Stumpf, M. P. H., Heck, A. J. R. and Mishto, M. (2016). A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science*, *354*(6310), 354–358. https://doi.org/10.1126/science.aaf4384

Liepe, J., Mishto, M., Textoris-Taube, K., Janek, K., Keller, C., Henklein, P., Kloetzel, P. M. and Zaikin, A. (2010). The 20S Proteasome Splicing Activity Discovered by SpliceMet. *PLoS Computational Biology*, *6*(6), e1000830. https://doi.org/10.1371/journal.pcbi.1000830

Loupe, J. M., Miller, P. J., Crabtree, J. S., Zabaleta, J. and Hollenbach, A. D. (2017). Acquisition of an oncogenic fusion protein is sufficient to globally alter the landscape of miRNA expression to inhibit myogenic differentiation. *Oncotarget*, *8*(50), 87054–87072. https://doi.org/10.18632/oncotarget.19693

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. and Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, *17*(20), 2337–2342. https://doi.org/10.1002/rcm.1196

Manczinger, M., Koncz, B., Balogh, G. M., Papp, B. T., Asztalos, L., Kemény, L., Papp, B. and Pál, C. (2021). Negative trade-off between neoantigen repertoire breadth and the specificity of HLA-I molecules shapes antitumor immunity. *Nature Cancer*, *2*(9), 950–961. https://doi.org/10.1038/s43018-021-00226-4

Masetti, R., Pigazzi, M., Togni, M., Astolfi, A., Indio, V., Manara, E., Casadio, R., Pession, A., Basso, G. and Locatelli, F. (2013). CBFA2T3-GLIS2 fusion transcript is a novel common feature in pediatric, cytogenetically normal AML, not restricted to FAB M7 subtype. *Blood*, *121*(17), 3469–3472. https://doi.org/10.1182/blood-2012-11-469825

Matsumura, M., Fremont, D. H., Peterson, P. A. and Wilson, lan A. (1992). Emerging Principles for the Recognition of Peptide Antigens by MHC Class I Molecules. *Science*, *257*(5072), 927–934. https://doi.org/10.1126/science.1323878

McCarthy, M. K. and Weinberg, J. B. (2015). The immunoproteasome and viral infection: a complex regulator of inflammation. *Frontiers in Microbiology*, *6*, 21. https://doi.org/10.3389/fmicb.2015.00021

McGranahan, N. and Swanton, C. (2019). Neoantigen quality, not quantity. *Science Translational Medicine*, *11*(506), eaax7918. https://doi.org/10.1126/scitranslmed.aax7918

Michaux, A., Larrieu, P., Stroobant, V., Fonteneau, J.-F., Jotereau, F., Eynde, B. J. V. den, Moreau-Aubry, A. and Vigneron, N. (2014). A Spliced Antigenic Peptide Comprising a Single Spliced Amino Acid Is Produced in the Proteasome by Reverse Splicing of a Longer Peptide Fragment followed by Trimming. *The Journal of Immunology*, *192*(4), 1962–1971. https://doi.org/10.4049/jimmunol.1302032

Mishto, M. (2020). What We See, What We Do Not See, and What We Do Not Want to See in HLA Class I Immunopeptidomes. *Proteomics*, *20*(15–16), 2000112. https://doi.org/10.1002/pmic.202000112

Mishto, M. (2021). Commentary: Are there indeed spliced peptides in the immunopeptidome? *Molecular & Cellular Proteomics*, *20*, 100158. https://doi.org/10.1016/j.mcpro.2021.100158

Mitelman, F., Johansson, B. and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, *7*(4), 233–245. https://doi.org/10.1038/nrc2091

Mulhern, R. K. and Butler, R. W. (2004). Neurocognitive sequelae of childhood cancers and their treatment. *Pediatric rehabilitation*, *7*(1), 1–14; discussion 15-6. https://doi.org/10.1080/13638490310001655528

Murakoshi, H., Koyanagi, M., Akahoshi, T., Chikata, T., Kuse, N., Gatanaga, H., Rowland-Jones, S. L., Oka, S. and Takiguchi, M. (2018). Impact of a single HLA-A*24:02-associated escape mutation on the detrimental effect of HLA-B*35:01 in HIV-1 control. *EBioMedicine*, *36*, 103–112. https://doi.org/10.1016/j.ebiom.2018.09.022

Murata, S., Takahama, Y., Kasahara, M. and Tanaka, K. (2018). The immunoproteasome and thymoproteasome: functions, evolution and human disease. *Nature Immunology*, *19*(9), 923–931. https://doi.org/10.1038/s41590-018-0186-z

Mylonas, R., Beer, I., Iseli, C., Chong, C., Pak, H.-S., Gfeller, D., Coukos, G., Xenarios, I., Müller, M. and Bassani-Sternberg, M. (2018). Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome*. *Molecular & Cellular Proteomics*, *17*(12), i–2357. https://doi.org/10.1074/mcp.ra118.000877

Nerli, S. and Sgourakis, N. G. (2020). Structure-Based Modeling of SARS-CoV-2 Peptide/HLA-A02 Antigens. *Frontiers in Medical Technology*, *2*, 553478. https://doi.org/10.3389/fmedt.2020.553478

Nguyen, A. T., Szeto, C. and Gras, S. (2021). The pockets guide to HLA class I molecules. *Biochemical Society Transactions*, *49*(5), 2319–2331. https://doi.org/10.1042/bst20210410

Nicorici, D., Satalan, M., Edgren, H., Kangaspeska, S., Murumagi, A., Kallioniemi, O., Virtanen, S. and Kilkku, O. (2014). FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*, 011650. https://doi.org/10.1101/011650

O'Donnell, T. J., Rubinsteyn, A. and Laserson, U. (2020). MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Systems*, *11*(1), 42-48.e7. https://doi.org/10.1016/j.cels.2020.06.010

Okuda, T., Taki, T., Nishida, K., Chinen, Y., Nagoshi, H., Sakakura, C. and Taniwaki, M. (2017). Molecular heterogeneity in the novel fusion gene APIP-FGFR2: Diversity of genomic breakpoints in gastric cancer with high-level amplifications at 11p13 and 10q26. *Oncology Letters*, *13*(1), 215–221. https://doi.org/10.3892/ol.2016.5386

Oliver, G. R., Jenkinson, G. and Klee, E. W. (2020). Computational Detection of Known Pathogenic Gene Fusions in a Normal Tissue Database and Implications for Genetic Disease Research. *Frontiers in Genetics*, *11*, 173. https://doi.org/10.3389/fgene.2020.00173

Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., … Wu, C. J. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, *547*(7662), 217–221. https://doi.org/10.1038/nature22991

Paes, W., Leonov, G., Partridge, T., Nicastri, A., Ternette, N. and Borrow, P. (2020). Elucidation of the Signatures of Proteasome-Catalyzed Peptide Splicing. *Frontiers in Immunology*, *11*, 563800. https://doi.org/10.3389/fimmu.2020.563800

Panigrahi, P., Jere, A. and Anamika, K. (2018). FusionHub: A unified web platform for annotation and visualization of gene fusion events in human cancer. *PLoS ONE*, *13*(5), e0196588. https://doi.org/10.1371/journal.pone.0196588

Parham, P. and Ohta, T. (1996). Population Biology of Antigen Presentation by MHC Class I Molecules. *Science*, *272*(5258), 67–74. https://doi.org/10.1126/science.272.5258.67

Park, I. and Terasaki, P. (2000). Origins of the first HLA specificities. *Human Immunology*, *61*(3), 185–189. https://doi.org/10.1016/s0198-8859(99)00154-8

Pearlman, A. H., Hwang, M. S., Konig, M. F., Hsiue, E. H.-C., Douglass, J., DiNapoli, S. R., Mog, B. J., Bettegowda, C., Pardoll, D. M., Gabelli, S. B., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. and Zhou, S. (2021). Targeting public neoantigens for cancer immunotherapy. *Nature Cancer*, *2*(5), 487–497. https://doi.org/10.1038/s43018-021-00210-y

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H. and Ferrin, T. E. (2020). UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Science*, *30*(1), 70–82. https://doi.org/10.1002/pro.3943

Pierini, F. and Lenz, T. L. (2018). Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection. *Molecular Biology and Evolution*, *35*(9), 2145–2158. https://doi.org/10.1093/molbev/msy116

Platteel, A. C. M., Liepe, J., Eden, W. van, Mishto, M. and Sijts, A. J. A. M. (2017). An Unexpected Major Role for Proteasome-Catalyzed Peptide Splicing in Generation of T Cell Epitopes: Is There Relevance for Vaccine Development? *Frontiers in Immunology*, *8*, 1441. https://doi.org/10.3389/fimmu.2017.01441

Platteel, A. C. M., Liepe, J., Textoris-Taube, K., Keller, C., Henklein, P., Schalkwijk, H. H., Cardoso, R., Kloetzel, P. M., Mishto, M. and Sijts, A. J. A. M. (2017). Multi-level Strategy for Identifying

Proteasome-Catalyzed Spliced Epitopes Targeted by CD8+ T Cells during Bacterial Infection. *Cell Reports*, *20*(5), 1242–1253. https://doi.org/10.1016/j.celrep.2017.07.026

Popovic, J., Li, L.-P., Kloetzel, P. M., Leisegang, M., Uckert, W. and Blankenstein, T. (2011). The only proposed T-cell epitope derived from the TEL-AML1 translocation is not naturally processed. *Blood*, *118*(4), 946–54. https://doi.org/10.1182/blood-2010-12-325035

Poran, A., Harjanto, D., Malloy, M., Arieta, C. M., Rothenberg, D. A., Lenkala, D., Buuren, M. M. van, Addona, T. A., Rooney, M. S., Srinivasan, L. and Gaynor, R. B. (2020). Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. *Genome Medicine*, *12*(1), 70. https://doi.org/10.1186/s13073-020-00767-w

Purcell, A. W. (2021). Is the Immunopeptidome Getting Darker?: A Commentary on the Discussion around Mishto et al., 2019. *Frontiers in Immunology*, *12*, 720811. https://doi.org/10.3389/fimmu.2021.720811

Pyke, R. M., Mellacheruvu, D., Dea, S., Abbott, C. W., McDaniel, L., Bhave, D. P., Zhang, S. V., Levy, E., Bartha, G., West, J., Snyder, M. P., Chen, R. O. and Boyle, S. M. (2022). A machine learning algorithm with subclonal sensitivity reveals widespread pan-cancer human leukocyte antigen loss of heterozygosity. *Nature Communications*, *13*(1), 1925. https://doi.org/10.1038/s41467-022-29203-w

Rammensee, H.-G., Friede, T. and Stevanović, S. (1995). MHC ligands and peptide motifs: first listing. *Immunogenetics*, *41*(4), 178–228. https://doi.org/10.1007/bf00172063

Reardon, B., Koşaloğlu-Yalçın, Z., Paul, S., Peters, B. and Sette, A. (2021). Allele-Specific Thresholds of Eluted Ligands for T-Cell Epitope Prediction. *Molecular & Cellular Proteomics : MCP*, *20*, 100122. https://doi.org/10.1016/j.mcpro.2021.100122

Reshmi, S. C., Harvey, R. C., Roberts, K. G., Stonerock, E., Smith, A., Jenkins, H., Chen, I.-M., Valentine, M., Liu, Y., Li, Y., Shao, Y., Easton, J., Payne-Turner, D., Gu, Z., Tran, T. H., Nguyen, J. V., Devidas, M., Dai, Y., Heerema, N. A., … Hunger, S. P. (2017). Targetable kinase gene fusions in high-risk B-ALL: a study from the Children's Oncology Group. *Blood*, *129*(25), 3352–3361. https://doi.org/10.1182/blood-2016-12-758979

Rist, M. J., Theodossis, A., Croft, N. P., Neller, M. A., Welland, A., Chen, Z., Sullivan, L. C., Burrows, J. M., Miles, J. J., Brennan, R. M., Gras, S., Khanna, R., Brooks, A. G., McCluskey, J., Purcell, A. W., Rossjohn, J. and Burrows, S. R. (2013). HLA Peptide Length Preferences Control CD8+ T Cell Responses. *The Journal of Immunology*, *191*(2), 561–571. https://doi.org/10.4049/jimmunol.1300292

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P. and Marsh, S. G. E. (2019). IPD-IMGT/HLA Database. *Nucleic acids research*, *48*(D1), D948–D955. https://doi.org/10.1093/nar/gkz950

Rock, K. L., Reits, E. and Neefjes, J. (2016). Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends in Immunology*, *37*(11), 724–737. https://doi.org/10.1016/j.it.2016.08.010

Rolfs, Z., Solntsev, S. K., Shortreed, M. R., Frey, B. L. and Smith, L. M. (2018). Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion. *Journal of Proteome Research*, *18*(1), 349–358. https://doi.org/10.1021/acs.jproteome.8b00651

Roukos, V. and Misteli, T. (2014). The biogenesis of chromosome translocations. *Nature Cell Biology*, *16*(4), 293–300. https://doi.org/10.1038/ncb2941

Saab, F., Hamelin, D. J., Ma, Q., Kovalchik, K. A., Sirois, I., Faridi, P., Li, C., Purcell, A. W., Kubiniok, P. and Caron, E. (2021). RHybridFinder: An R package to process immunopeptidomic data for putative hybrid peptide discovery. *STAR protocols*, *2*(4), 100875. https://doi.org/10.1016/j.xpro.2021.100875

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., Bukur, V., Tadmor, A. D., Luxemburger, U., Schrörs, B., Omokoko, T., Vormehr, M., Albrecht, C., Paruzynski, A., Kuhn, A. N., Buck, J., Heesch, S., Schreeb, K. H., Müller, F., … Türeci, Ö. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, *547*(7662), 222–226. https://doi.org/10.1038/nature23003

Sahu, I. and Glickman, M. H. (2021). Proteasome in action: substrate degradation by the 26S proteasome. *Biochemical Society Transactions*, *49*(2), 629–644. https://doi.org/10.1042/bst20200382

Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., Justesen, S., Stevens, J., Lane, W. J., Eisenhaure, T., … Keskin, D. B. (2020a). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature biotechnology*, *38*(2), 199–209. https://doi.org/10.1038/s41587-019-0322-9

Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., Justesen, S., Stevens, J., Lane, W. J., Eisenhaure, T., … Keskin, D. B. (2020b). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature biotechnology*, *38*(2), 199–209. https://doi.org/10.1038/s41587-019-0322-9

Scholtalbers, J., Boegel, S., Bukur, T., Byl, M., Goerges, S., Sorn, P., Loewer, M., Sahin, U. and Castle, J. C. (2015). TCLP: an online cancer cell line catalogue integrating HLA type, predicted neo-epitopes, virus and gene expression. *Genome Medicine*, *7*(1), 118. https://doi.org/10.1186/s13073-015-0240-5

Sette, A. and Sidney, J. (1999). Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, *50*(3–4), 201–212. https://doi.org/10.1007/s002510050594

Sette, Alessandro, Sidney, J., Guercio, M.-F. del, Southwood, S., Ruppert, J., Dahlberg, C., Grey, H. M. and Kubo, R. T. (1994). Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Molecular Immunology*, *31*(11), 813–822. https://doi.org/10.1016/0161-5890(94)90019-1

Sidney, J., Peters, B., Frahm, N., Brander, C. and Sette, A. (2008). HLA class I supertypes: a revised and updated classification. *BMC Immunology*, *9*(1), 1–1. https://doi.org/10.1186/1471-2172-9-1

Sidney, J., Southwood, S., Moore, C., Oseroff, C., Pinilla, C., Grey, H. M. and Sette, A. (2013). Measurement of MHC/Peptide Interactions by Gel Filtration or Monoclonal Antibody Capture. *Current Protocols in Immunology*, *100*(1), 18.3.1-18.3.36. https://doi.org/10.1002/0471142735.im1803s100

Smith, J. L., Ries, R. E., Hylkema, T., Alonzo, T. A., Gerbing, R. B., Santaguida, M. T., Brodersen, L. E., Pardo, L., Cummings, C. L., Loeb, K. R., Le, Q., Imren, S., Leonti, A. R., Gamis, A. S., Aplenc, R., Kolb, E. A., Farrar, J. E., Triche, T. J., Nguyen, C., … Meshinchi, S. (2020). Comprehensive Transcriptome Profiling of Cryptic CBFA2T3–GLIS2 Fusion–Positive AML Defines Novel Therapeutic Options: A COG and TARGET Pediatric AML Study. *Clinical Cancer Research*, *26*(3), 726–737. https://doi.org/10.1158/1078-0432.ccr-19-1800

Solberg, O. D., Mack, S. J., Lancaster, A. K., Single, R. M., Tsai, Y., Sanchez-Mazas, A. and Thomson, G. (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Human Immunology*, *69*(7), 443–464. https://doi.org/10.1016/j.humimm.2008.05.001

Specht, G., Roetschke, H. P., Mansurkhodzhaev, A., Henklein, P., Textoris-Taube, K., Urlaub, H., Mishto, M. and Liepe, J. (2020). Large database for the analysis and prediction of spliced and non-spliced peptide generation by proteasomes. *Scientific Data*, *7*(1), 146. https://doi.org/10.1038/s41597-020-0487-6

Tanaka, K. (2009). The proteasome: Overview of structure and functions. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, *85*(1), 12–36. https://doi.org/10.2183/pjab.85.12

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(Database issue), D941–D947. https://doi.org/10.1093/nar/gky1015

Thibaudeau, T. A. and Smith, D. M. (2019). A Practical Review of Proteasome Pharmacology. *Pharmacological Reviews*, *71*(2), 170–197. https://doi.org/10.1124/pr.117.015370

Trolle, T., McMurtrey, C. P., Sidney, J., Bardet, W., Osborn, S. C., Kaever, T., Sette, A., Hildebrand, W. H., Nielsen, M. and Peters, B. (2016). The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference. *The Journal of Immunology*, *196*(4), 1480–1487. https://doi.org/10.4049/jimmunol.1501721

Venema, W. J., Hiddingh, S., Boer, J. H. de, Claas, F. H. J., Mulder, A., Hollander, A. I. den, Stratikos, E., Sarkizova, S., Veken, L. T. van der, Janssen, G. M. C., Veelen, P. A. van and Kuiper, J. J. W. (2021). ERAP2 Increases the Abundance of a Peptide Submotif Highly Selective for the Birdshot Uveitis-Associated HLA-A29. *Frontiers in Immunology*, *12*, 634441. https://doi.org/10.3389/fimmu.2021.634441

Vigneron, N., Stroobant, V., Chapiro, J., Ooms, A., Degiovanni, G., Morel, S., Bruggen, P. van der, Boon, T. and Eynde, B. J. V. den. (2004). An Antigenic Peptide Produced by Peptide Splicing in the Proteasome. *Science*, *304*(5670), 587–590. https://doi.org/10.1126/science.1095522

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. and Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Warren, E. H., Vigneron, N. J., Gavin, M. A., Coulie, P. G., Stroobant, V., Dalet, A., Tykodi, S. S., Xuereb, S. M., Mito, J. K., Riddell, S. R. and Eynde, B. J. V. den. (2006). An Antigen Produced by Splicing of Noncontiguous Peptides in the Reverse Order. *Science*, *313*(5792), 1444–1447. https://doi.org/10.1126/science.1130660

Weingarten-Gabbay, S., Klaeger, S., Sarkizova, S., Pearlman, L. R., Chen, D.-Y., Gallagher, K. M. E., Bauer, M. R., Taylor, H. B., Dunn, W. A., Tarr, C., Sidney, J., Rachimi, S., Conway, H. L., Katsis, K., Wang, Y., Leistritz-Edwards, D., Durkin, M. R., Tomkins-Tinch, C. H., Finkel, Y., … Sabeti, P. C. (2021). Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. *Cell*, *184*(15), 3962-3980.e17. https://doi.org/10.1016/j.cell.2021.05.046

Wen, B., Li, K., Zhang, Y. and Zhang, B. (2020). Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature Communications*, *11*(1), 1759. https://doi.org/10.1038/s41467-020-15456-w

Yang, W., Lee, K.-W., Srivastava, R. M., Kuo, F., Krishna, C., Chowell, D., Makarov, V., Hoen, D., Dalin, M. G., Wexler, L., Ghossein, R., Katabi, N., Nadeem, Z., Cohen, M. A., Tian, S. K., Robine, N., Arora, K., Geiger, H., Agius, P., … Morris, L. G. T. (2019). Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nature Medicine*, *25*(5), 767–775. https://doi.org/10.1038/s41591-019-0434-2

Yarmarkovich, M., Marshall, Q. F., Warrington, J. M., Premaratne, R., Farrel, A., Groff, D., Li, W., Marco, M. di, Runbeck, E., Truong, H., Toor, J. S., Tripathi, S., Nguyen, S., Shen, H., Noel, T., Church, N. L., Weiner, A., Kendsersky, N., Martinez, D., … Maris, J. M. (2021). Cross-HLA targeting of intracellular oncoproteins with peptide-centric CARs. *Nature*, *599*(7885), 477–484. https://doi.org/10.1038/s41586-021-04061-6

Yotnda, P., Firat, H., Garcia-Pons, F., Garcia, Z., Gourru, G., Vernant, J. P., Lemonnier, F. A., Leblond, V. and Langlade-Demoyen, P. (1998). Cytotoxic T cell response against the chimeric p210 BCR-ABL protein in patients with chronic myelogenous leukemia. *Journal of Clinical Investigation*, *101*(10), 2290–2296. https://doi.org/10.1172/jci488

Yotnda, P., Garcia, F., Peuchmaur, M., Grandchamp, B., Duval, M., Lemonnier, F., Vilmer, E. and Langlade-Demoyen, P. (1998). Cytotoxic T cell response against the chimeric ETV6-AML1 protein in childhood acute lymphoblastic leukemia. *Journal of Clinical Investigation*, *102*(2), 455–462. https://doi.org/10.1172/jci3126

Zamora, A. E., Crawford, J. C., Allen, E. K., Guo, X. J., Bakke, J., Carter, R. A., Abdelsamed, H., Moustaki, A., Li, Y., Chang, T.-C., Awad, W., Dallas, M. H., Mullighan, C. G., Downing, J. R., Geiger, T. L., Chen, T., Green, D. R., Youngblood, B. A., Zhang, J. and Thomas, P. G. (2019a). Pediatric patients with acute lymphoblastic leukemia generate abundant and functional neoantigen-specific CD8+ T cell responses. *Science Translational Medicine*, *11*(498), eaat8549. https://doi.org/10.1126/scitranslmed.aat8549

Zamora, A. E., Crawford, J. C., Allen, E. K., Guo, X. J., Bakke, J., Carter, R. A., Abdelsamed, H., Moustaki, A., Li, Y., Chang, T.-C., Awad, W., Dallas, M. H., Mullighan, C. G., Downing, J. R., Geiger, T. L., Chen, T., Green, D. R., Youngblood, B. A., Zhang, J. and Thomas, P. G. (2019b). Pediatric patients with acute lymphoblastic leukemia generate abundant and functional neoantigen-specific CD8+ T cell responses. *Science Translational Medicine*, *11*(498), eaat8549. https://doi.org/10.1126/scitranslmed.aat8549

Zhang, J., Mardis, E. R. and Maher, C. A. (2017). INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*, *33*(4), 555–557. https://doi.org/10.1093/bioinformatics/btw674

Zhao, W. and Sher, X. (2018). Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLoS Computational Biology*, *14*(11), e1006457. https://doi.org/10.1371/journal.pcbi.1006457