Université de Montréal

Expanding the immune self: Impact of non-canonical translation on the repertoire of MHC I-associated peptides

par
Céline M. Laumont

Programme de Biologie Moléculaire, option Biologie des Systèmes
Faculté de Médecine

Thèse présentée à la Faculté de Médecine
en vue de l'obtention du grade de Philosophiae Doctor
en Biologie Moléculaire
option Biologie des Systèmes

Août, 2018

Résumé

Les molécules du complexe majeur d'histocompatibilité de classe I (MHC I) sont des glycoprotéines de surface exprimées par la majorité des cellules nucléés de notre organisme. Ces molécules servent à exposer une vue intégrative de l'état interne de nos cellules (soi immunitaire) *via* la présentation de courts peptides (MAPs) générés lors de la dégradation des protéines cytosoliques par le protéasome. Le répertoire des MAPs de chaque cellule, véritable carte d'identité peptidique, est constamment passé en revue par nos lymphocytes T CD8+, cellules centrales du système immunitaire, afin de débusquer et d'éliminer toute cellule anormale, e.g., celles présentant des MAPs d'origine virale ou tumorale (TSAs).

Au vu du nombre grandissant d'articles démontrant que des ARNs autres que les ARNs messagers peuvent être traduits, nous avons décidé d'évaluer l'impact de ces mécanismes de traduction non-canonique sur le répertoire des MAPs présentés par des cellules B. En développant une approche protéogénomique, i.e., combinant spectrométrie de masse et séquençage d'ARN à haut-débit, nous avons pu démontrer qu'environ 10 % des MAPs présentés par nos cellules B dérivent d'évènements de traduction non-canonique incluant (i) la traduction d'ARNs messagers dans un cadre de lecture alternatif ou (ii) la traduction de régions ou ARNs supposés non-codants. L'analyse subséquente des caractéristiques de ces MAPs dits « cryptiques » suggère que leur biogenèse diffère de celle des MAPs conventionnels, les MAPs cryptiques étant principalement encodés par des ARNs instables produisant de courtes protéines dont la dégradation ne semble pas exiger l'intervention du protéasome.

Sachant que la déméthylation globale du génome des cellules cancéreuses permet l'expression d'un plus grand bassin d'ARNs non-codants, nous avons supposé que ces cellules pourraient présenter de nombreux MAPs (et TSAs) cryptiques. En adaptant notre approche protéogénomique, nous avons pu analyser le répertoire des MAPs de cellules cancéreuses, incluant celui de deux lignées tumorales de souris (EL4 et CT26) et sept échantillons primaires humains (quatre leucémies aigues lymphoblastiques B et trois biopsies de cancer du poumon). Cette analyse nous a

permis de découvrir qu'environ 90% des TSAs sont des TSAs cryptiques. Ayant observé que la plupart de ces TSAs dérivent de séquences normales dont l'expression est restreinte aux cellules tumorales, comme les retroéléments endogènes, il est plausible que ces TSAs soient partagés par plusieurs patients. Enfin, nos études chez la souris nous ont permis de démontrer qu'au moins deux facteurs influencent positivement le potentiel protectif d'un TSA *in vivo* : l'expression de cet antigène par les cellules cancéreuses et la fréquence des lymphocytes T capables de le reconnaître.

En conclusion, le recours à la protéogénomique pour analyser les MAPs présentés par les cellules normales et cancéreuses nous a permis de démontrer que les MAPs cryptiques contribuent significativement au bassin de peptides constituant le soi immunitaire et qu'ils permettent aux lymphocytes T CD8+ d'effectuer une surveillance immunitaire plus efficace.

Mots-clés : Complexe majeur d'histocompatibilité de classe I, traduction non-canonique, antigènes spécifiques aux tumeurs, lymphocytes T, immunothérapie du cancer, protéogénomique, spectrométrie de masse, séquençage de nouvelle génération

Abstract

On their surface, nucleated cells present major histocompatibility complex class I (MHC I) molecules in complex with short peptides, that we will refer to as MHC I-associated peptides (MAPs). These MAPs derive from the degradation of cytosolic proteins by the proteasome and provide an integrative view of the inner state of cells to CD8+ T cells, which can, in turn, eliminate abnormal cells, e.g., those presenting viral MAPs or tumor-specific antigens (TSAs).

With the growing body of evidence suggesting that translation does occur outside of protein-coding transcripts, we tried to evaluate the impact of non-canonical translation on the repertoire of MAPs. Combining RNA-sequencing and mass spectrometry to analyze the MAP repertoire of B-lymphoblastoid cell lines, we uncovered that ~ 10 % of the MAP repertoire derives from such non-canonical translation events, including (i) the out-of-frame translation of protein-coding transcripts or (ii) the translation of non-coding regions (UTRs, introns, etc.) or transcripts (antisense, pseudogene, etc.). Interestingly, our data suggest that the biogenesis of cryptic and conventional MAPs differs, as cryptic MAPs derive from unstable transcripts generating short proteins that might be degraded in a proteasome-independent fashion.

Because the global DNA hypomethylation observed in cancer cells tend to de-repress non-coding transcripts, we developed another proteogenomic approach to probe the cryptic MAP repertoire of two murine cancer cell lines (EL4 and CT26) and seven humor primary tumor samples (four B-lineage acute lymphoblastic leukemias and three lung tumor biopsies). This second analysis revealed that ~ 90% of TSAs are cryptic TSAs. Interestingly, most of those TSAs derived from cancer-restricted yet non-mutated sequences, such as endogenous retroelements, thereby suggesting that such TSAs could be shared between patients. Lastly, our validation study in mice demonstrated that at least two parameters can influence the *in vivo* protective effect of TSAs, namely TSA expression in cancer cells and the frequency of TSA-specific T cells.

Altogether, our proteogenomic studies on the MAP repertoire of normal and cancer cells demonstrate that cryptic MAPs significantly expand the immune self and, consequently, the scope of CD8+ T cell immunosurveillance.

Keywords : Major histocompatibility complex class I, non-canonical translation, tumor-specific antigens, T lymphocytes, cancer immunotherapy, proteogenomics, mass spectrometry, next-generation sequencing

Table of contents

List of Tables

CHAPTER 2

CHAPTER 4

List of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

APPENDIX I

APPENDIX II

List of Abbreviations

0-9

7-AAD	7-aminoactinomycin D

A

aa	Amino acid
aeTSA	Aberrantly expressed TSA
APC	Antigen-presenting cell
AS	Ankylosing spondylitis

B

β_2m	β_2-microglobulin
B-ALL	B-lineage acute lymphoblastic leukemia
B-LCL	B-lymphoblastoid cell line
BCR	B-cell receptor
bp	Base pair

C

CAR	Chimeric antigen receptor
CALR	Calreticulin
CDAM	Cell death-associated molecule
cT	Cumulative numbers of transcripts
cTEC	Cortical thymic epithelial cell

D

DAMP	Danger-associated molecular pattern
DB	Database
DC	Dendritic cell
DP	Double-positive thymocyte
DRiP	Defective ribosomal product

E

eIF	Eukaryotic initiation factor
ER	Endoplasmic reticulum
ERAP1/2	ER aminopeptidase 1 and 2
ERE	Endogenous retroelement

F-G-H

FDR	False discovery rate
FPKM	Fragments per kilobase of transcript per million mapped reads
H	Hyperploid
HLA	Human leucocyte antigen
HPLC	High-performance liquid chromatography

I-J-K-L

IC	Immunocompetent
ID	Immunodeficient
IFN-γ	Interferon gamma
IL	Interleukin
IP	Immunoprecipitation
LC-MS/MS	Liquid chromatography-MS/MS

M

m/z	Mass-to-charge ratio
MAE	Mild acid elution
Met-tRNA$_i^{Met}$	Methionine-loaded initiator tRNA
MHC	Major histocompatibility complex
MHC I	Major histocompatibility complex class I
MHC II	Major histocompatibility complex class II
MIP or MAP	MHC I-associated peptide
mRNA	Messenger RNA
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
mTEC^{hi}	MHC II^{high} medullary thymic epithelial cell
mTSA	Mutated TSA
MTX	Mitoxantrone

N

NMD	Nonsense mediated decay
nS	Number of analyzed samples
ns-SNP	Non-synonymous single-nucleotide polymorphism
NSG	NOD Cg-Prkdc^{scid}Il2rg^{tm1Wjl}/SzJ mice
nt	Nucleotide

O-P

ORF	Open reading frame
P	Parental
PAMP	Pathogen-associated molecular pattern
PCR	Peptide-coding region
PLC	Peptide-loading complex
MSPRF	Programmed ribosomal frameshifting
PRM-MS	Parallel reaction monitoring MS
PRR	Pattern-recognition receptor

Q-R-S

RDP	Rapidly degraded protein
Ribo-Seq	Ribosome profiling
RNA-Seq	RNA-sequencing
rphm	Reads detected per 10^8 reads sequenced
S-value	Seen-value
S/MRM-MS	Selected/Multiple reaction monitoring MS
SMII	Self MHC I immunopeptidome
SP	Single-positive thymocyte

T-U-V

TAA	Tumor-associated antigen
TAP	Transporter associated with antigen processing
TCR	T-cell receptor
TIL	Tumor-infiltrating lymphocyte
tpm	Transcripts per million
TRA	Tissue-restricted antigen
TSA	Tumor-specific antigen
Tsn	Tapasin
UEA1	Ulex europaeus lectin 1
UTR	untranslated region

W-X-Y-Z

WT	Wildtype

In loving memory of Victor Manuel Garza Ibarra

(October 19, 1988 – April 12, 2018)

To Write About You After You Are Gone Is

To find hollyhocks boasting blooms
of pinks and whites and burgundies,
arranging them in a bundle,
watching them explode the dining
room table with perfumed color,
and then—to wait until they have
wilted, died, crumpled, and been cleared,
to try and explain to someone
all that this empty vase once held.

Sarah Kay

“In the beginner's mind there are many possibilities;
in the expert's mind there are few.”

In Zen Mind, Beginner’s Mind
by Shunryu Suzuki

Acknowledgements

Because most of the people with who I had the pleasure to work with are francophones, this section is the sole section of my thesis that will be written in French. But really, the main message is THANK YOU!

À l'aube d'écrire cette thèse, entre excitation et angoisse de la page blanche, je ne peux m'empêcher de penser que, peu importe le résultat, ces six années de doctorat auront été un voyage à nul autre pareil : traversée d'un océan d'incertitude(s) ponctuée de joies dont l'intensité n'aura eu d'égale que leur rareté. Et, si la passion m'a aidée à garder le cap, envers et contre tout, sachez que je suis venue à bout de cette thèse grâce à vous, vos sourires, vos encouragements et vos bottages de fesses (parfois nécessaires) lors de mes périodes de déni majeur ! Maintenant que le voyage s'achève (et quel voyage !), il ne me reste plus qu'à vous remercier et je suis curieusement anxieuse à l'idée d'écrire ces quelques lignes… la peur de ne pas trouver les mots ou d'oublier quelques noms sans doute ! Mais, c'est comme la thèse, il faut bien se lancer !

Tout d'abord, j'aimerais remercier **Claude Perreault**, mon directeur de recherche. Je me revois encore débarquer dans votre bureau pour mon entrevue fin janvier 2011, motivée, certes, mais sans expérience réelle, stressée et forcément un peu embrouillée. Je ne sais pas ce qui a fait pencher la balance en ma faveur (mon dossier académique ou ma lecture de 'La Mort' par Vladimir Jankélévitch ?), mais je ne vous remercierai jamais assez d'avoir misé sur moi ! Merci ! Merci d'avoir pris le temps de me former, de m'avoir poussée quand il le fallait et surtout de m'avoir laissé suffisamment de latitude pour que je puisse trouver ma place. Merci aussi pour vos sages conseils et autres citations éclairantes, toujours entre rigueur et créativité avec une touche d'humour. Merci enfin pour toutes nos discussions non-scientifiques mais néanmoins passionnantes, du théâtre à la littérature en passant par la politique et la religion. Vous êtes définitivement un mentor à nul autre pareil et je me sais chanceuse d'avoir pu, pendant toutes ces années, apprendre auprès de vous.

Il m'est impossible de remercier Claude sans prendre quelques lignes pour remercier **Pierre Thibault** et **Sébastien Lemieux**, co-directeurs officieux de cette thèse. Pierre, merci de m'avoir initiée à la spectrométrie de masse sans laquelle des projets comme les nôtres ne pourraient exister. Merci aussi de m'avoir appris l'affirmation de soi, j'en manquais cruellement au départ et n'aurais pu y arriver sans elle. Seb, merci pour ta disponibilité ainsi que la simplicité et la clarté de tes explications sur des notions de stats ou d'info parfois obscures pour moi. Merci aussi de m'avoir initiée, dans la vie comme en science, à l'art du vide, mon équilibre mental et mes lecteurs t'en sont extrêmement reconnaissants !

Enfin, un gros merci à tous mes membres de comités et de jury d'avoir pris le temps de m'écouter, de me conseiller et de m'épauler sur les différents projets que j'ai menés lors de cette thèse : **Jonathan L. Bramson**, **François-Michel Boisvert**, **Damien D'Amours**, **Philippe P. Roux**, **Daniel Sinnett** et **Jacques Thibodeau**. Merci aussi à la **Fondation Cole** de m'avoir soutenue financièrement pendant les deux dernières années de cette thèse. Ma toute première 'vraie' bourse ! Je me revois hurler de joie dans le bureau en l'apprenant ! Votre confiance aura été un moteur pour moi et faire partie de la famille des boursiers Cole un véritable honneur.

Évidemment, pour entreprendre une thèse, il ne suffit pas d'avoir de bons mentors et du soutien financier, il faut aussi de bons compagnons de route, de ceux qui sèchent vos larmes quand tout s'écroule et festoient avec vous quand tout s'emballe, de ceux qui vous poussent et vous défient, de ceux qui vous aiment en somme. Sur ce plan, je peux dire que j'ai été gâtée car la liste est longue !

Un gros merci à **tous les membres du labo Perreault** pour nos discussions animées autour du café matinal, pour les lunchs 'en famille', entre fous-rires et coups de gueule (faut bien relâcher la pression !!!) et bien sûr pour votre indéfectible soutien. Un merci tout particulier à **Caroline Côté** de m'avoir pris sous son aile quand je suis arrivée, toute perdue, de Paris. Mon Caro, je peux te dire que mon intégration au Québec aurait été beaucoup plus difficile sans ton écoute, tes encouragements et ton

humour ! Merci à **Diana Paola Granados**, d'avoir pris autant de temps pour me former au cours de mon stage de bac et de ma maîtrise. Ma petite Didi, nos nombreuses discussions et tes explications m'ont toujours beaucoup appris et, vraiment, c'est un peu grâce à toi si j'en suis là aujourd'hui. Merci à **Tariq Daouda** de m'avoir challengée voire déstabilisée, sans toi je ne me serai jamais mise à la bioinformatique et un projet comme le projet TSA n'aurait jamais pu voir le jour ! Merci à **Krystel Vincent** pour… ben pour tout quoi !! Mon KV, je sais qu'on se connaissait à peine au début du projet TSA et, si on s'est d'abord découverte comme collègues, tu es désormais bien plus que ça. Shit, meuf, we did it !! Tu réalises ?! Toi pis moi, Docteur… Allo làààààààà (bon moi pas encore mais bientôt si tout va bien ! ah ah ah). Ça mérite bien un de mes emblématiques YOLOOOOOOO !!!! Je t'aime fort Partner, à la vie à la mort, toi-même tu sais !!! Enfin, merci aux autres membres du labo d'avoir partagé des petits bouts de vie ici et là, en particulier **Anca Apavaloaei** (for all our seriously funny discussions !), **Marie-Pierre Hardy** (MPPPP, qu'est-ce que je ferais sans tes robes hein ?!), **Leslie Hesnard** (yoga time anytime ma Leslou chou !), **Jean-David Larouche** (tu goooooosses mais je t'aime pareil ma Douille !), **Charles St-Pierre** (la preuve par mille que les opposés peuvent s'entendre hein mon CSP !) et **Assya Trofimov** (toi et ton internet parallèle, vous me faites mourir !).

Mille mercis à **tous les membres des plateformes de spectrométrie de masse et de bioinformatique de l'IRIC** pour leur écoute et leurs conseils. Sans vous pour me guider, la biologiste que je suis aurait été rapidement dépassée par ces projets multidisciplinaires ! Un gros merci à **Éric Bonneil**. Ma vieille, on aura tout traversé, du rire aux larmes avec des 'Tiens-moi ça !' par-ci par-là mais qu'est-ce qu'on se sera amuse !!! Alors, du fond du cœur, merci, merci de m'avoir expliqué si clairement la mass spec', d'avoir toujours pris le temps de réfléchir à mes requêtes plus bizarres les unes que les autres (j'y peux rien, c'est ça travailler avec des bases de données géantes !!) et de m'avoir remonté le moral quand je n'y croyais plus. Merci aussi à **Patrick Gendron** et **Jean-Philippe Laverdure**. Mon Pat, mon JPJP, merci d'avoir pris le temps de former le bébé bioinfo que je suis et d'avoir pardonné mes erreurs de débutante (genre quand Pat se fait réveiller à 3h du mat' parce que quelqu'un, c'est-à-

dire moi, a rempli proteo1… Oupssssss !). Je sais que je vous ai demandé beaucoup et vous n'avez jamais failli. Toujours présents et avec le sourire en plus ! Vous êtes hoooooot les boys !!!! Enfin, un merci tout particulier à **Éric Audemard** pour nos discussions, toujours animées et passionnantes, et ta franchise sans égale. Sans toi, pas de k-mers certain ! Merci aussi d'avoir cru en mes capacités de 'codage' et, par ricochet, de m'y avoir fait croire. Sans ta confiance, je n'aurais jamais osé travailler sur un projet de l'ampleur du projet TSA. T'es le meilleur !

Finalement, merci à **tous mes amis** d'avoir illuminé ma vie montréalaise. **Aux amis des premiers pas** : Ale, Alex, Caro, Christian, Colunga, Danny (and Emily now !), DryDryy, Sammy, Roberto, Rodrigo, Victor, Will et Yassou. C'est à travers votre joie de vivre, votre insouciance et votre légèreté d'expat' que j'ai pu découvrir et aimer Montréal, entre folles soirées, voyages impromptus, projets photos/peintures et discussions en tout genre. C'est aussi vous qui, en soignant mes petits bobos de tous les jours, m'avez permis de garder le cap et de devenir la fille indépendante que je suis aujourd'hui. Merci d'avoir été mon premier cocon à 6000 km de la maison.

Aux amis d'hier, d'aujourd'hui et de demain : Alex (et Vi !), Alexou (et Padam !), Chachou, DB, Dave (et Jack !), Hibouuu, Jalilouille, Lolito (et BBB !), Marjoooo (et Kev !), Maxou, Mimi (et la famille !), Mister Y (#no-us #6959), (flying) Piiiiiiit, Romarion, toute la famille du Studio de Yoga Wanderlust, Thomas Milan et Vince. C'est don' ben l'fun la vie avec vous !!!! On en fait des affaires folles ! La route des vins à vélo (et le parc des îles-de-Boucherville aussi… ou pas !), du kayak au Laaaaaac, des randos en hurlant 'Ouhouuuuuu !!!!' pour éloigner les ours, une rando sur 8 jours en autonomie complète (sans en avoir jamais fait avant… *challenge accepted !*), des tours en CESSNA Skyhawk II au-dessus de Montréal (si on a peur des gros avions, pourquoi ne pas embarquer dans un mini 4 places ? #logiqueImparable), de l'escalade à Kamouraska (malgré la peur du vide… *if you can't beat fear, just do it scared !*), des midterms à chaque semaine (parce que 'Gracias a la vidaaaa !!'), des handstands plus ou moins stables, de la vulgarisation scientifique en s'amusant, des combats de boxe sur fond de rap, des chants de Noweellll sur la route des vacances, une dégustation de gin qui tourne au carnage, des karaokés de feu au

2P, des # partout tout le temps parce qu'on est des malades, une coloc inoubliable à l'odeur parfois douteuse (la faute à qui hein ma C.A.D.A. d'amour ?!), un petit tour aux urgences en pleine nuit juste pour dire, des restau de folie et des brunchs aussi (entre Apérol Spritz et Mimosas !!!), des spectacles qu'on veut revoir à l'infini ('Cravate ou nœud papillon ?'), un accident de vélo avec plus de peur que de mal (mais un peu de mal quand même !), des courses dans le cimetière (le Ziiiiiiiiig de la mort et JP II forever !!), du yoga, du yoga et toujours du yoga puis de belles casquettes ! Bref, merci pour toutes ces belles aventures ! J'espère que vous me pardonnerez la liste non-exhaustive et n'oubliez pas que je vous aime ++ les choubidous ! Je ne l'aurais jamais fait sans vous !

Merci enfin **à la famille et aux amis de toujours** : Papi, Papa, Maman et ma Dinde puis mon Alexia, mon Spaghetti (avec la Maoua et Aya bien sûr !), Minou Senior (sans oublier Gégé et Martine !), ma Féfé, mon Chaton, mon Gros Chat, ma Choooooose d'amour et mon Bébé chou. Malgré la distance vous êtes toujours là avec moi, accompagnant chacun de mes gestes, chacune de mes pensées. Tout plein d'amour !

CHAPTER 1

1 Introduction

1.1 The immune system: a fine-tuned self/non-self discrimination strategy

To discriminate self from non-self is a fundamental requirement for life. Indeed, throughout the entire evolutionary tree, from Bacteria to Eukaryota, we can find a plethora of self/non-self discrimination strategies that are primarily used (i) to avoid self-mating, thereby decreasing the probability of an inbreeding depression to occur, and (ii) to identify and eliminate pathogens, thereby increasing the organism's life-span[1]. Although immune defense mechanisms do exist in prokaryotes, with, for instance, the now famous CRISPR-Cas system[2,3], the immune system of jawed vertebrates (or gnathostomes) is probably the most complex system of all. It is composed of two parts, *the innate and the adaptive immune system*, that work in concert to orchestrate immune responses against invaders (i.e., pathogens such as bacteria, viruses, parasites), while limiting damages to self-tissues.

1.1.1 The innate immune system: one (receptor) for all (specificities)

Once one of our body's epithelial barrier has been breached by a microorganism, components of the innate immune system will act rapidly, i.e., within minutes after the infection, and locally, i.e., near the entry point, to counter and (hopefully) eliminate the threat. To discriminate 'infectious non-self from non-infectious self', as stated by Charles A. Janeway Jr. in 1992[4], innate immune cells express various families of pattern-recognition receptors (PRRs) able to recognize a wide range of pathogen-associated molecular patterns (PAMPs)[5]. Such PAMPs include, but are not restricted to, complex polysaccharides, glycolipids and nucleic acids shared by multiple pathogens[6,7]. Recognition of any of those PAMPs by a PRR is sufficient to unleash the host's innate immune response, which in turn promotes a faster elimination of this non-self threat through (i) the destruction of pathogens by phagocytosis for macrophages and dendritic cells (DCs) or degranulation for eosinophils, (ii) the release of pro-

inflammatory cytokines and the remodelling of the extracellular matrix to facilitate the recruitment of additional immune cells, such as neutrophils, and (iii) the pathogen's opsonisation, or even lysis, by the complement pathway.

The use of germline-encoded receptors and effectors by the innate immune system has three related implications for innate immune responses. First, they are *non-specific* and *stereotyped*[8,9]. Indeed, because the same PRR can recognize multiple pathogens through a common PAMP, such immune responses are necessarily PRR-specific rather than pathogen-specific and this regardless of the number of times the pathogen was encountered. Second, they are *unlikely to target self structures*[10]. Indeed, according to Darwin's theory of evolution, any allele encoding a self-specific PRR should be selected against because of its deleterious consequences for the host, thereby limiting the risk of PRR-triggered autoimmunity. Finally, they are *likely to be overcome by pathogens* (at some point in evolution). Indeed, any microbe presenting with PRR-insensitive molecular patterns or ways to counter downstream effectors[11,12] are likely to be selected for, as these confer selective advantages to overcome innate immune responses. Thus, although relying on a perfect self/non-self discrimination strategy, the limited versatility of the innate immune system might render it powerless against some pathogens so that it will require, at times, the help of the adaptive immune system to fully eliminate non-self threats.

1.1.2 The adaptive immune system: to each (receptor) its own (specificity)

About ~ 500 $\times$ 10^6 years ago, the adaptive (or acquired) immune system appeared independently in both agnathans and gnathostomes[13]. For jawed vertebrates, this system constitutes a second line of defense against pathogens as it requires priming by innate immune cells, such as macrophages and DCs, to be unleashed[14]. Now, even if adaptive immune responses take longer to mount than innate ones (~ a few days vs. minutes), they are *highly specific* to the pathogen that induced them and provide *long-lasting* protection against it, through the generation of immune memory cells, which react faster and stronger in case of pathogen re-exposure[15,16].

The core cellular components of the adaptive immune system are *B and T lymphocytes*, which develop in the bone marrow and the thymus, respectively[17,18]. B cells carry out antibody-mediated responses to neutralize and favor the phagocytosis of non-self extracellular antigens[19], while T cells carry out cell-mediated responses to eliminate infected cells[20]. Rather than sensing pathogens with a set of germline-encoded, and therefore predefined, PRRs like innate immune cells, lymphocytes make use of clonally distributed antigen-binding receptors, namely the B-cell receptor (BCR) for B cells and the T-cell receptor (TCR) for T cells. These receptors are generated, in each developing lymphocyte, by a highly regulated somatic recombination process, termed V(D)J recombination, in which the stochastic cleavage and subsequent joining of defined, yet non-contiguous, portions of the BCR- or TCR-encoding genes can generate an extremely diverse repertoire of antigen-binding receptors (~ 10^{18} BCRs and $10^{15\text{-}20}$ TCRs vs. < 1,000 for PRRs)[21-24]. In fact, this theoretical BCR/TCR repertoire is so diverse that it vastly exceeds the number of cells available in our body (~ 3.72×10^{13})[25], thereby implying that the limiting factor is the number of lymphocytes we possess, rather than the number of receptors we can generate[26]! Although each lymphocyte expresses only *one* acquired antigen-binding receptor, i.e., a BCR or a TCR, with a given antigenic specificity, the overall lymphocyte population possesses an extremely *diverse* repertoire of BCRs/TCRs that can recognize a very wide repertoire of antigens. As people say, ‘there is strength in numbers’!

The fact that such receptors are acquired rather than inherited and selected for over evolutionary time can be seen as a double-edged sword. On one hand, the somatic recombination process can be seen as an advantage because it confers an extreme versatility, and therefore flexibility, to the adaptive immune system. Indeed, having such a diverse repertoire of antigen-binding receptors increases the likelihood of at least one lymphocyte recognizing an antigen from any, i.e., even the most recent, pathogens. On the other hand, it also implies that some of those antigen-binding receptors might recognize self-structures, thereby triggering, often harmful, autoimmune reactions.

Altogether, this highlights that, while the adaptive immune system is a powerful system to overcome constantly evolving pathogens, it also requires mechanisms to establish immunological tolerance towards self-structures, i.e., self-tolerance.

1.2 T lymphocytes: detecting non-self threats by cultivating the art of self-referential indifference

The establishment of self-tolerance, that is the non-responsiveness of B or T lymphocytes towards self-structures, occurs primarily during their development in primary lymphoid organs (i.e., the bone marrow or the thymus) and is further maintained in periphery. Although required for the proper functioning of both lymphocyte lineages[27], we will only review mechanisms involved in T-cell tolerance after briefly describing how T cells discriminate self from non-self.

1.2.1 The nature of self for T cells

Based on the composition of their TCR, circulating T lymphocytes belong either to (i) the rare innate-like $\gamma\delta$ subgroup[28], mostly involved in epithelial and mucosal immune responses[29], or to (ii) the frequent and better characterized $\alpha\beta$ subgroup. $\alpha\beta$ T cells, referred to as T cells or T lymphocytes in this thesis, represent ~ 90% of circulating T cells and are eminently *self-referential*[30]. Indeed, they rely on TCR-dependent interactions with self-molecules, namely peptides associated with molecules of the major histocompatibility complex (MHC), to survive in periphery[31] and eliminate non-self threats[32].

T-cell subsets are defined based on their role in immune responses (conventional vs. regulatory), their stage of differentiation (naive vs. effector vs. memory) as well as their expression of the CD4 or CD8 co-stimulatory molecule, which defines their antigenic specificity. As such, CD4+ T cells recognize peptide/MHC class II (MHC II) complexes presented mainly by professional antigen-presenting cells (APCs, e.g., macrophages, DCs and B cells), while CD8+ T cells recognize peptide/MHC class I (MHC I) complexes presented by most nucleated cells. Following TCR-dependent antigen recognition and clonal expansion, both T-cell lineages will differentiate into effector T cells to carry out T helper responses for CD4+ T cells and cytolytic responses for CD8+ T cells. Later, ~ 95% of those antigen-experienced clones undergo apoptosis, while the remaining ones persist as memory T cells in periphery.

Because self-peptide/MHC complexes are required for both T-cell survival and T-cell immune responses in periphery, circulating T cells must bear a well-calibrated TCR, that is a TCR combining (i) an acceptable level of self-reactivity, thereby promoting survival without triggering autoimmunity[33] with (ii) an exquisite sensitivity to the broadest repertoire of non-self MAPs, thereby allowing for a greater versatility[34].

1.2.2 The establishment of self-tolerance

The selection and, later on, the maintenance of a self-tolerant/non-self-reactive T-cell repertoire requires mechanisms (i) to eliminate autoreactive lymphocytes during their maturation in the thymus, i.e., *central tolerance*[35] and (ii) to control autoreactive lymphocytes that might have nonetheless escaped in periphery, i.e., *peripheral tolerance*[36]. These processes are complementary and failures in either one of them inevitably results in autoimmune disorders such as the "autoimmune polyglandular syndrome type 1" for central tolerance[37,38] or the "immune dysfunction, polyendocrinopathy, enteropathy, X-linked" syndrome for peripheral tolerance[39].

1.2.2.1 Central tolerance, the deadly odyssey of thymocytes through the thymus

Naive T cells are continually produced by the thymus through the recruitment and maturation of bone-marrow-derived early thymic progenitors[40], though this production decreases with age[41]. Throughout their thymic odyssey, thymocytes first rearrange their α/β TCR-encoding genes[42] and then undergo a drastic two-step selection process, namely *the positive and negative selection* (**Figure 1.1**). At this stage, TCRs' affinity for self-peptide/MHC complexes appears to be one, if not the most, crucial parameter dictating cell fate, that is survival or death (**Figure 1.2**). Overall, less than 3% of thymocytes survive their 'thymic test' and are therefore exported in periphery, as mature naive T cells, to fulfill their immune functions[43].

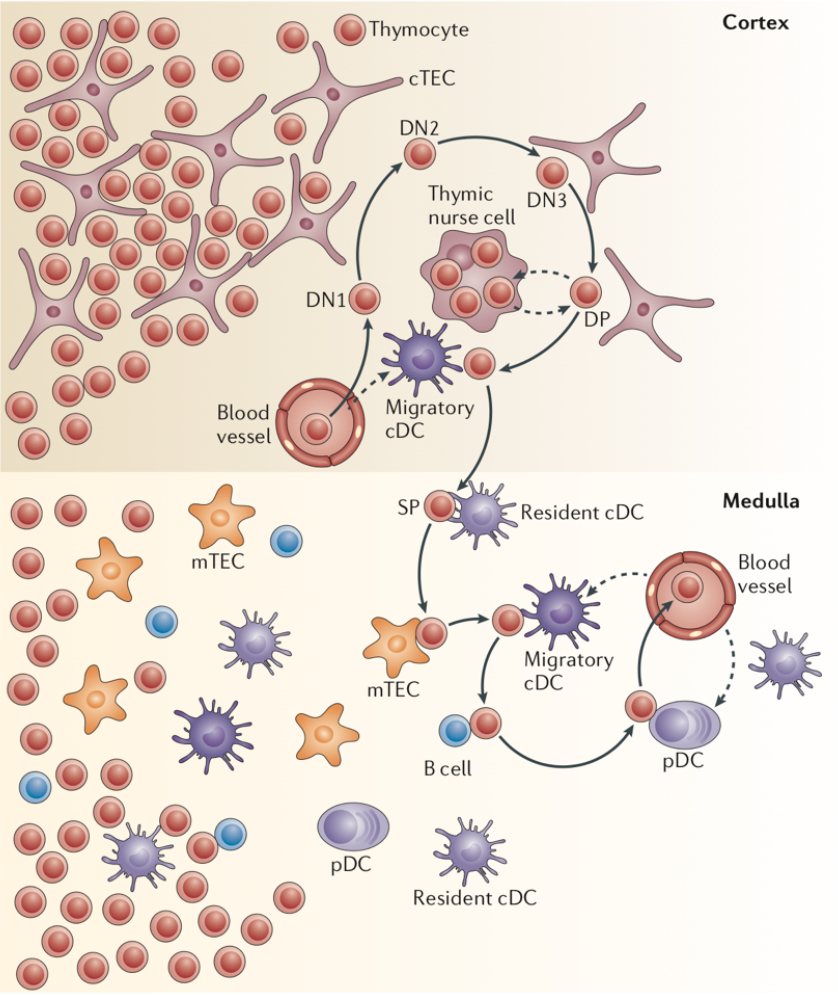


Figure 1.1 | T cell development in the thymus. CD4-CD8- double-negative (DN) thymocytes enter the thymus through blood vessels located at the cortico-medullary junction. Through their thymic migration, DNs pass by various stages of maturation, called DN1 to DN4, before acquiring a TCR and finally becoming CD4+CD8+ double-positive (DP) thymocytes. At this stage, DPs wander through the outer cortex scanning peptide/MHC complexes presented by cortical thymic epithelial cells (cTECs), involved in positive selection, and cortical conventional dendritic cells (cDCs), likely leading to negative selection. Positively selected DPs must then commit to the CD4 or CD8 lineage before migrating through the thymic medulla and undergo negative selection. During this migration lasting 4 to 5 days, CD4+ or CD8+ single-positive (SP) thymocytes scan self-peptide/MHC complexes presented by hundreds of antigen-presenting cells including cDCs, plasmacytoid DCs (pDCs), medullary thymic epithelial cells (mTECs) and B cells. Solid and dashed arrows represent main migratory pathways involved in thymocyte selection and other relevant migratory pathways, respectively. Adapted with permission from Springer Nature: Nature Reviews Immunology (Klein L, Kyewski B, Allen PM, and Hogquist KA)[44] © 2014

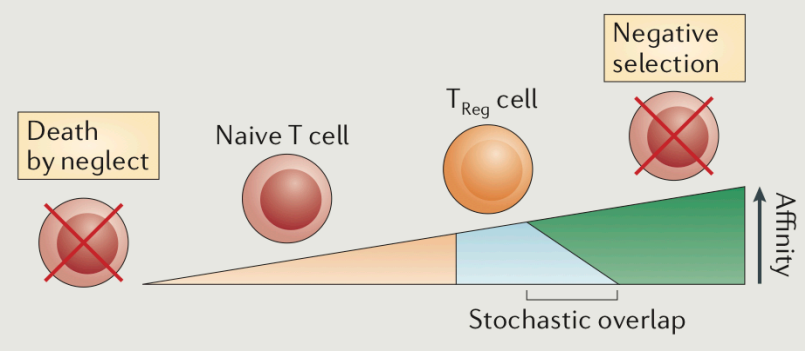


Figure 1.2 | The affinity model of thymocyte selection. In this model, the strength of the interaction between the TCR and self-peptide/MHC complexes dictates cell fate. As such, thymocytes can either die by neglect (none), give rise to naive and regulatory T cells (T_{Reg}, weak to intermediate) or undergo negative selection (strong). Adapted with permission from Springer Nature: Nature Reviews Immunology (Klein L, Kyewski B, Allen PM, and Hogquist KA)[44] © 2014

Positive selection (**Figure 1.1, top part**) takes place in the thymic cortex, where CD4+CD8+ double-positive (DP) thymocytes encounter cortical thymic epithelial cells (cTECs), that act as APCs in this process[45-47]. There, DPs able to establish weak TCR-dependent interactions with self-peptide/MHC complexes presented on cTECs receive a *positive* signal to survive and further differentiate into CD4+CD8- or CD4-CD8+ single-positive (SP) thymocytes, while DPs unable to make contact die by neglect[48] (**Figure 1.2**). Interestingly, this positive selection process is done on a very peculiar set of self-peptides, as their generation relies on cTEC-specific antigen processing pathways involving the thymoproteasome (MHC I)[49] or the cathepsin L1 and thymus-specific serine protease (MHC II)[50,51]. Although the role of such cTEC-specific self-peptide/MHC complexes remains a matter of debate[52], it is clear that positive selection enriches the T-cell repertoire for functional thymocytes, i.e., thymocytes able to interact with self-peptides, and, as a corollary, for highly self-reactive thymocytes.

Deletion of highly self-reactive cells requires SP thymocytes to pass through the thymic medulla, where they undergo *negative selection* (**Figure 1.2, bottom part**). Mature (MHC II^{high}) medullary thymic epithelial cells ($mTEC^{hi}$) are the key, though not the sole[53,54], APCs involved in this selection process. Indeed, $mTEC^{hi}$ express two transcription factors, AIRE (the famous autoimmune regulator)[55] and the recently

discovered Fezf2[56], that trigger the promiscuous expression of, otherwise, tissue-restricted transcripts, such as insulin[57]. Self-peptide/MHC complexes generated by those tissue-restricted transcripts are called tissue-restricted antigens (TRAs) and mimic self-peptides that might be presented by peripheral cells[57,58]. Similarly to positive selection, TCRs' affinity for those complexes greatly influences the fate of SP thymocytes with (i) high-affinity interactions acting as a *negative* signal leading to clonal deletion, i.e., apoptosis of highly autoreactive thymocytes, (ii) intermediate-affinity interactions leading to clonal diversion, i.e., differentiation into regulatory T cells (T_{Regs}) and (iii) low-affinity interactions leading to survival and differentiation into canonical naive T cells[48] (**Figure 1.2**). Due to the mosaic expression of TRA-source transcripts in $mTEC^{hi}$, each being found in ~ 1–3% of $mTEC^{hi}$[59,60], SP thymocytes are required to scan ~ 200–500 $mTEC^{hi}$ to be fully tolerized prior to their export in periphery[44].

Altogether, one can say that the establishment of central tolerance in the thymus is the perfect exemplification of the ancient Delphic maxim 'Know thyself' mentioned by Socrates in Plato's Phaedrus. As opposed to Socrates that did not understand himself enough to bother study other things, developing T cells that did acquire an extensive self-knowledge through the, more or less subsequent, positive and negative thymic selection[44] can definitely focus on other things, that is to eliminate non-self threats, e.g., infected or cancerous cells, by finding the few non-self-peptide/MHC complexes hidden in a sea of self-complexes.

1.2.2.2 Peripheral tolerance, maintaining the *status quo*

As drastic as it might seem, central tolerance alone does not fully prevent the export of self-reactive T cells in periphery. Indeed, it has been hypothesized that such deletion might create 'holes' in the peripheral T-cell repertoire that could be exploited by pathogens[61-63]. Thus, several mechanisms, collectively referred to as peripheral tolerance mechanisms, do exist to keep autoreactive T cells in check.

Ignorance of self-peptide/MHC complexes *by naive T cells* is perhaps the most straightforward way to prevent the unwanted firing of autoreactive T cells. Under steady-state conditions, the probability of T-cell attacks against self-tissues is reduced

by simply excluding T cells from peripheral tissues, which might contain cells presenting TRAs at a density sufficient to trigger T-cell activation[64,65]. As a surrogate, naive T cells continuously circulate between the blood and the lymph, passing through secondary lymphoid organs, namely the spleen and lymph nodes, where they scan the surface of DCs in search for non-self-peptide/MHC complexes. Indeed, in this setting, DCs are the professional APCs that can freely circulate through peripheral tissues to sample their respective microenvironment and migrate back to draining secondary lymphoid organs to present the result of their wandering adventures, in the form of MHC-associated peptides, to T cells.

In the event of T-cell activation by self-peptide/MHC complexes, other mechanisms can help shutdown unwanted adaptive immune responses. First, it was shown that *the phenotype of DCs can control T-cell responsiveness*. Indeed, tissue-resident DCs, that mature and migrate towards lymph nodes upon infection, present with an immunogenic phenotype that can facilitate T-cell activation and differentiation into effector T cells. To do so, these immunogenic DCs have been shown to combine the presentation of non-self-peptide/MHC complexes with the production of pro-inflammatory cytokines and the upregulation of co-stimulatory molecules[36]. On the contrary, more immature forms of DCs are known to display a tolerogenic phenotype[66], which can prevent T-cell activation through peripheral clonal deletion[67,68], clonal anergy[69] or induction of T_{Regs}[70]. Besides DCs' phenotype, *immune suppression by peripherally-induced and thymus-derived T_{Regs}* is an efficient mechanism to quench autoimmune reactions. Indeed, these regulatory cells, that likely require a lower activation threshold than naive T cells[71], can shutdown ongoing immune responses through (i) secretion of immunosuppressive cytokines (IL-10, TGF-β)[72,73], (ii) lysis of autoreactive T cells[74], (iii) cytokine deprivation[75] or metabolic disruption[76-78] leading to responder T-cell apoptosis and (iv) prolonged interaction with DCs[79,80] or inhibition of their maturation[81,82] to limit T-cell priming. Finally, *TCR-signalling dampening by co-inhibitory molecules* represent the last safeguard to prevent autoimmune reactions. Following their activation, self-reactive T cells present co-inhibitory molecules, such as

CTLA-4 or PD-1, which can, upon binding to their ligand on target cells, shutdown T-cell responses[83,84].

Altogether, central and peripheral tolerance mechanisms ensure that the adaptive immune system will only fire when required through the selection and the maintenance of a T-cell repertoire cultivating the art of self-referential indifference. Having said that, one can understand that, besides the theoretical number of non-self-peptide/MHC complexes that can be recognized by T cells[85], the clearance of non-self threats from our organism will greatly depend on the capacity of infected cells to process and present those non-self-peptides.

1.3 The MHC I antigen presentation pathway: generating diversity through a common framework

Peptide/MHC complexes are central for T-cell activation, since their TCR-dependent recognition by T cells is one of the crucial cues required to launch an adaptive immune response. With such a central role in adaptive immunity, the collection, or repertoire, of MHC-associated peptides presented by each cell must faithfully represent what is happening, at any given time, both in the extracellular and intracellular milieu. To generate such comprehensive representation of their outer space and inner state, professional APCs, as well as peripheral cells, rely on two distinct antigen presentation pathways: *the MHC II and MHC I antigen presentation pathways*, which both ensure the continuous production of peptides for loading onto the relevant MHC molecules prior to their review by CD4+ or CD8+ T cells.

The *MHC II antigen presentation pathway*, which will not be detailed here, is essential for the degradation, processing and presentation of peptides derived from endocytosed *extracellular proteins* on MHC II molecules[86], though intracellular proteins degraded through autophagy can also enter this pathway[87]. Presentation of MHC II-associated peptides to CD4+ T cells is primarily done by professional APCs, which constitutively express MHC II-coding genes, other peripheral cells requiring IFN-γ exposure to do so[88]. Nonetheless, the repertoire of MHC II-associated peptides can be seen as a 'cellular digest' of the extracellular milieu, which is expected to contain self-proteins under steady state conditions and a mix of self/non-self proteins upon infection.

The *MHC I antigen presentation pathway* is essential for the production and presentation of peptides derived from the degradation of *intracellular proteins*, though alternative pathways allow for the presentation of peptides derived from extracellular proteins. In this section, we will first review the central role of MHC I molecules in shaping our immune self, i.e., selecting a wide variety of MHC I-associated peptides (MAPs), and further highlight some of the key steps involved in their biogenesis.

1.3.1 MHC I molecules: from polymorphic genes to molecules with specific peptide-binding properties

With few exceptions, classical MHC I molecules are presented by all nucleated cells. These molecules, also referred to as human leucocyte antigens (HLA) in humans, are heterodimers composed of an invariant soluble light chain, β_2-microglobulin (β_2m), and a polymorphic transmembrane heavy chain. The MHC I heavy chain comes in three flavors (HLA-A, -B or -C in humans and H-2-D, -K and -L in mice), which all have a similar tertiary structure composed of three extracellular domains, called α_1, α_2 and α_3, a transmembrane segment and a short cytoplasmic tail. The membrane-proximal α_3 domain is required for T-cell recognition, as it interacts with the CD8 co-stimulatory molecule of T cells[89], while the more distal α_1 and α_2 domains fold to form a closed peptide-binding groove able to accommodate 8–11 amino acid-long peptides for presentation at the cell surface[90] (**Figure 1.3**).

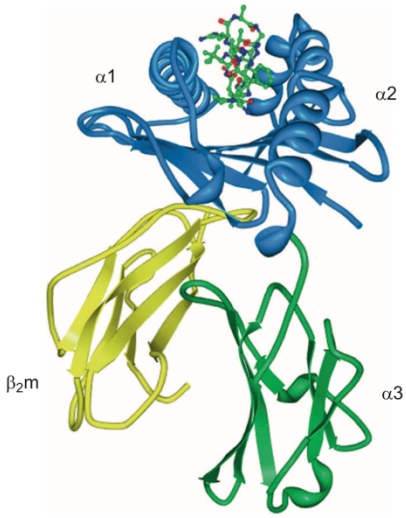


Figure 1.3 | Tertiary structure of HLA-B*07:02 in complex with a pp65 peptide. The α_1 and α_2 domains are shown in blue and the conserved α_3 and β_2m in green and yellow, respectively. The pp65 peptide is depicted in a ball- and-stick model. Adapted with permission from Springer Nature: Cellular & Molecular Immunology (Halenius A, Gerke C and Hengel H)[91] © 2014

Despite this common structure, MHC I heavy chains are encoded by the most polymorphic genes of the human genome, with a total of 4,340, 5,212 and 3,930 alleles reported for *HLA-A*, *-B* and *-C*, respectively (IPD-IMGT/HLA database last consulted on August 10th 2018)[92]. Implications of this unprecedented allelic variation at the MHC I loci are essential for self/non-self discrimination by CD8+ T cells. First, it increases the likelihood that *each individual possesses its own and unique MHC I haplotype*, that is a combination of up to six co-dominantly expressed MHC I alleles. Second, because most of these polymorphisms affect amino acids located within the peptide-binding groove of MHC I molecules[93,94], *differences in MHC I haplotype drastically affect the MAP repertoire* of cells[95,96]. Indeed, besides hydrogen bonds, peptide binding to MHC I molecules requires interactions between binding pockets, found within the peptide-binding groove, and side chains of specific residues along the peptide sequence. Depending on the location and nature of the amino acid substitutions induced by those polymorphisms, the number and properties of these binding pockets changes, thereby creating peptide-binding motifs specific to each MHC I molecule[97,98]. Often, the primary and secondary anchors, i.e., residues imposed by those binding motifs, are located at the N- and C-terminus (position 2 and Ω) of the peptide sequence[99], though some molecules also require a central anchor (position 5 or 6)[100,101]. While non-anchor residues can accommodate a great variety of amino acids, anchor residues mostly accommodate hydrophobic, and to a lesser extent, basic residues[102]. Thus, although restrictive, these binding motifs still allow MHC I molecules to present very different MAP repertoires, some being narrower than others[103]. By combining up to six MAP repertoires, one for each of the MHC I molecule encoded by their haplotype, individuals can present an even more diverse set of MAPs, thereby increasing their chance to present MAPs derived from viral or mutated proteins. Accordingly, experimental evidence suggest that individuals heterozygous for all MHC I genes are better equipped to fight pathogens than homozygous ones[104-106], though, taken alone, MHC I molecules can have opposite effects on one's susceptibility to infectious diseases[107,108] or autoimmune disorders[109,110].

Altogether, one can say that *the MHC I haplotype defines the immune self* by selecting what will be presented (or not) on cells, thereby creating a *diverse MAP repertoire unique to each individual*. Nonetheless, the diversity of MAPs within each one of those repertoires must be influenced by the diversity of peptides that MHC I molecules can sample, that is to say by the diversity of peptides the MHC I antigen presentation pathway can output.

1.3.2 To present your inner self, use the classical pathway!

In most nucleated cells, the pool of MAP precursor peptides sampled by empty MHC I molecules are produced by the *classical* MHC I antigen presentation pathway. In this two-in-one pathway, the first arm produces peptide-receptive empty MHC I molecules in the endoplasmic reticulum (ER, **section 1.3.2.1**), while the second arm produces a wide variety of MAP precursor peptides for loading onto those empty molecules prior to their export at the plasma membrane (**section 1.3.2.2**).

1.3.2.1 Production of peptide-receptive empty MHC I molecules

Assembly of peptide-receptive empty MHC I molecules takes place in the ER through an MHC-adapted version of the ER glycoprotein quality control cycle, where binding by high-affinity peptides, rather than just proper folding, triggers the export of MHC I molecules towards the plasma membrane (**Figure 1.4**).

Briefly, nascent MHC I heavy chains are co-translationally translocated into the ER and monoglucosylated to allow for the recruitment of calnexin[111], an ER-resident transmembrane chaperone, and ERp57[112], its associated thiol oxidoreductase. This interaction, which prevents aggregation of free heavy chains[113], also assists their proper oxidative folding with the sequential formation of two disulfide bonds within the α_3 and α_2 domains[114], respectively. Such pre-folded MHC I heavy chains can now non-covalently associate with β_2m, thereby stabilizing the α_2 disulfide bond, though the overall complex remains unstable[115]. Finally, calnexin is replaced by calreticulin, a soluble ER-resident chaperone, which can in turn facilitate the recruitment of the peptide-loading complex (PLC). This multiprotein complex is formed by the association

of an MHC I heavy chain/β_2m heterodimer with calreticulin, ERp57, the heterodimeric transporter associated with antigen processing (TAP), composed of TAP1 and TAP2, and tapasin (Tsn), a transmembrane MHC I-specific chaperone[116]. Upon interaction with components of the PLC, especially the Tsn/ERp57 heterodimer, peptide-receptive empty MHC I molecules are stabilized in the ER and can now await their loading with a MAP prior to their export at the plasma membrane.

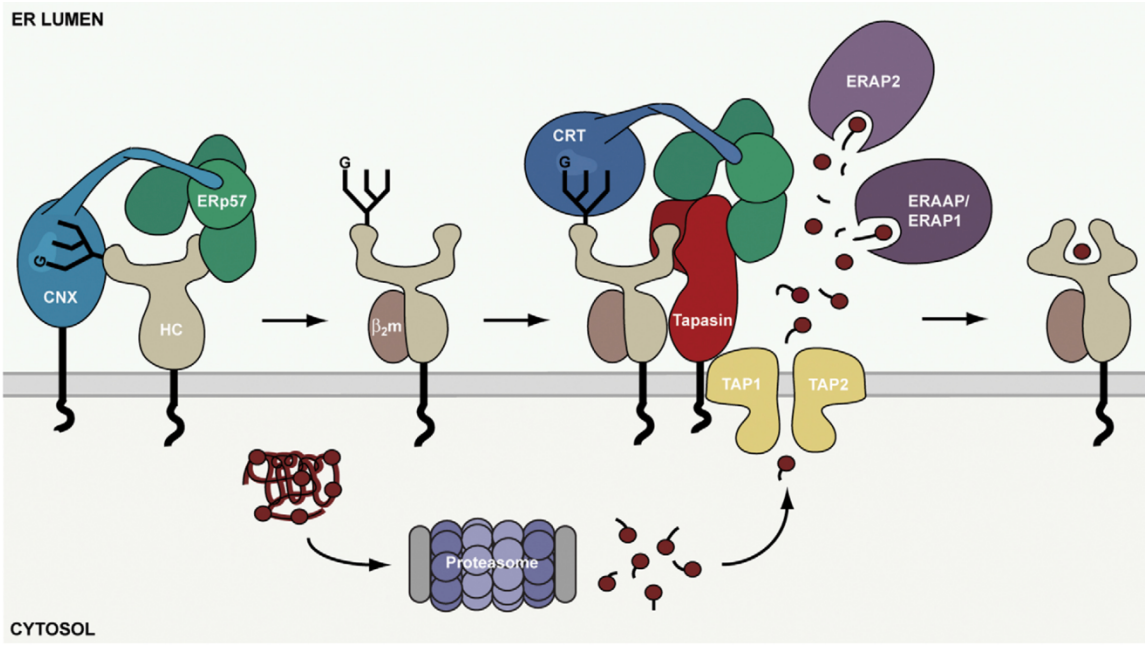


Figure 1.4 | MHC class I assembly in the ER. The MHC I heavy chain (HC), assisted by calnexin (CNX) and ERp57, undergoes oxidative folding prior to its association with β_2m. Then, the empty and unstable HC/β_2m heterodimer interacts with components of the peptide-loading complex (tapasin, ERp57, calreticulin (CRT), and TAP) to await for peptide binding in the ER. Adapted with permission from Elsevier: Current Opinion in Cell Biology (Wearsch PA, Cresswell P)[117] © 2008

1.3.2.2 From the cytosol to the plasma membrane: dangerous adventures in a proteolytic country!

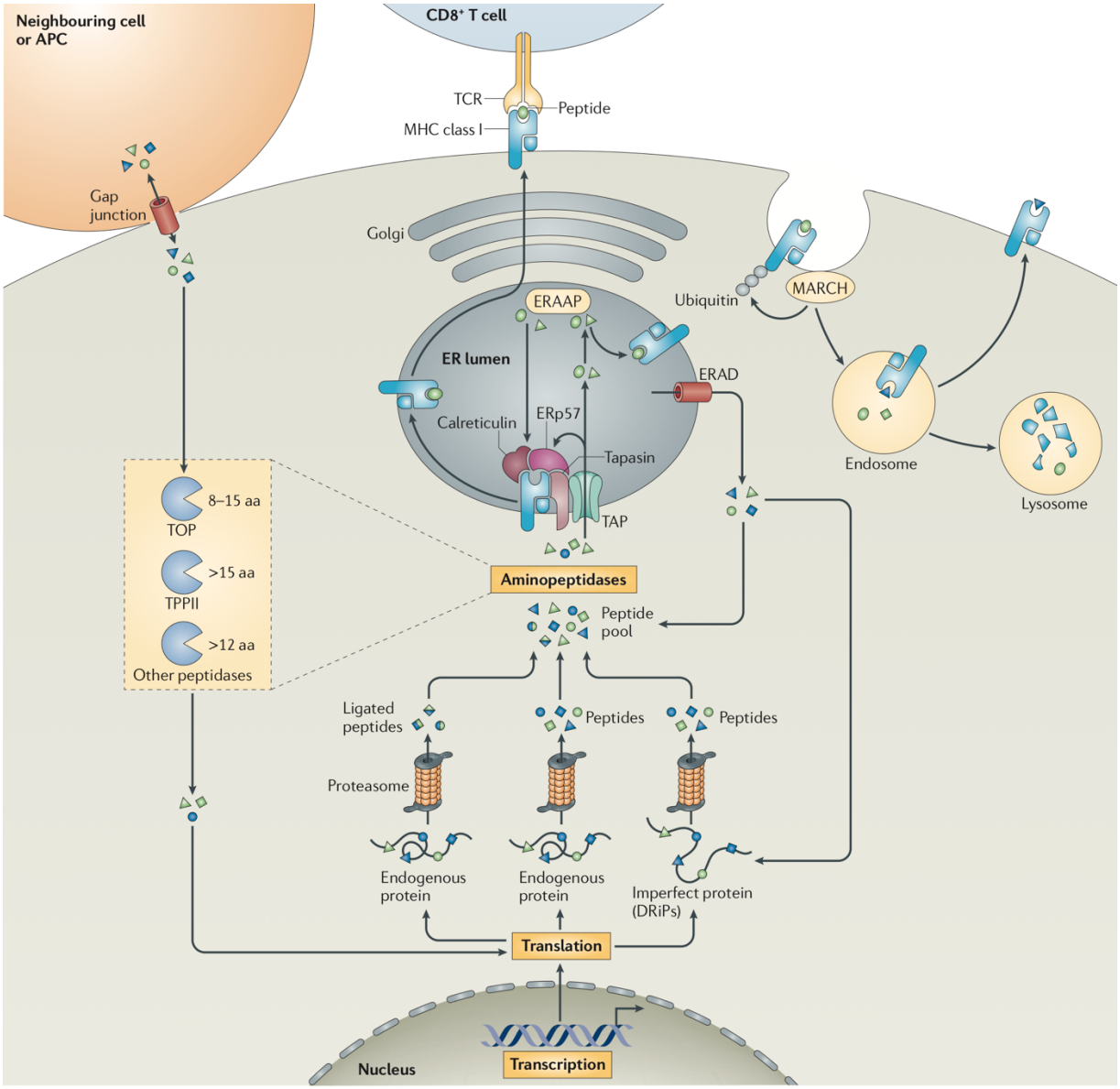


Figure 1.5 | Schematic of the MHC I antigen presentation pathway. Production of peptides for binding to MHC I molecules starts in the cytosol with the degradation of endogenous proteins and defective ribosomal products (DRiPs) by the proteasome. Although cytosolic aminopeptidases (TOP: thimet oligopeptidase, TPPII: tripeptidyl peptidase II) degrade most of those peptides, a few of them are translocated in the ER by the transporter associated with antigen presentation (TAP). These peptides, which might require additional trimming by the ER aminopeptidase associated with antigen processing (ERAAP), are then loaded onto MHC I molecules, with or without the help of the peptide-loading complex (composed of TAP, Tapasin, ERp57, Calreticulin), before being exported at the plasma membrane and reviewed by CD8+

T cells. Unbound peptides and empty MHC I molecules are re-exported to the cytosol by the ER-associated protein degradation (ERAD) system. Interestingly, peptides from neighboring cells, translocated in the cytosol by gap junctions, can also be processed and presented. Lastly, ubiquitination of MHC I molecules targets them to lysosomes, where they can either be degraded or recycled to expose endosomal peptides. Adapted with permission from Springer Nature: Nature Reviews Immunology (Neefjes J, Jongsma MLM, Paul P and Bakke O)[118] © 2011

1.3.2.2.1 Book 1: Using cytosolic proteins to make peptides

Cytosolic proteins enter the classical MHC I antigen presentation pathway through the ubiquitin-proteasome system (**Figure 1.5, bottom**). This requires the constitutive proteasome[119,120], a huge barrel-shaped multisubunit proteolytic complex composed of a 20S catalytic core and one or two regulatory caps[121], that recognizes and subsequently degrades polyubiquitinated proteins. Thanks to its wide spectrum of cleavage preferences, conferred by its three active subunits (β_1, β_2 and β_5), the constitutive proteasome can degrade almost any protein into 4–20 amino acid-long peptides[122], among which N-terminal extensions of MAPs can be found[123]. Most of these peptides are further degraded by cytosolic peptidases, often aminopeptidases, which play a crucial role in both MAP generation[124,125] and destruction[126-128]. Indeed, if some proteolytic activity is required to recycle amino acids and guarantee cells' proper functioning, too much of it can also destruct MAP precursor peptides, thereby decreasing the diversity of MAPs cells can present[126].

Other proteases than the constitutive proteasome have been involved in the classical MHC I antigen presentation pathway. Such proteases include the immunoproteasome, the thymoproteasome and a handful of cytosolic proteases. First, the immunoproteasome is a form of proteasome constitutively expressed by immune cells, though inducible by pro-inflammatory cytokines in non-immune cells[129,130]. Because it differs from the constitutive proteasome by its three active subunits (β_{1i}, β_{2i} and β_{5i}), the immunoproteasome's cleavage preferences are slightly different and favor the generation of peptides bearing more hydrophobic C-termini[131,132]. Consequently, its expression can change the MAP repertoire of cells[133,134] and further alter the ability of

CD8+ T cells to clear pathogens[135,136]. Second, the thymoproteasome is a cTEC-restricted form of proteasome that differ from the immunoproteasome by one of its three active subunit (β_{5t}). Although its role remains unclear[52], the change in cleavage preferences induced by the incorporation of β_{5t} have been shown to alter the MAP repertoire[49] so that it becomes optimal to support positive selection in the thymus[137,138]. Lastly, some *cytosolic proteases* have been shown to generate MAPs in a proteasome-independent fashion[139], and, despite the scarcity of those examples, still enlarge the MAP repertoire of cells by processing proteins that might be poorly degraded by proteasomes[140-143]. Thus, degradation of proteins by the ubiquitous constitutive proteasome and/or alternative proteases ensure that MHC I molecules can sample a very diverse repertoire of MAP precursor peptides[144], i.e., derived from any regions of any proteins.

1.3.2.2.2 Book 2: Finding the perfect match in the ER to get to plasma membrane

In the cytosol, balance between peptide production and degradation is clearly shifted towards degradation, ~ 99% of all produced peptides being degraded[127]. Thus, translocation of MAP precursor peptides in the ER, within seconds after their production, is a key step for them to reach the cell surface (**Figure 1.5, center**). TAP, the ER-resident transporter, continuously performs this pumping task and favors the import of 9–16 amino acid-long peptides in the ER[145]. There, peptides requiring additional N-terminal trimming (i.e., with length > 8 amino acids) are shortened by ER aminopeptidase 1 and 2 (ERAP1/2) to reach a length comprised between 8 and 11 amino acids[146-149], optimal for binding to MHC I molecules. This processing is essential for MAP generation, as the absence of ERAP1/2 has been shown to reduce presentation of MHC I molecules[150,151] and change the composition of the MAP repertoire[152,153].

Now that all the required components are in the ER, peptides can be loaded onto peptide-receptive MHC I molecules (for which we reviewed the folding in **section 1.3.2.1**, **Figure 1.5, center**). Most of these empty MHC I molecules, though not all[154-156], are stabilized by the PLC and rely on the Tsn/ERp57 heterodimer[157] to ensure their

quick loading with high-affinity peptides selected from this ER pool (**Figure 1.4**)[158-161]. During this process, called peptide proofreading or editing, only high-affinity peptides can compete with Tsn to close the peptide-binding groove of the MHC I molecule, subsequently triggering its release from the PLC[162]. Besides Tsn, the recently discovered TAP-binding protein-related[163] has been shown to act as an MHC I-specific chaperone and peptide proofreader, though in a PLC-independent fashion[164]. Following peptide loading and/or dissociation from the PLC, MHC I molecules are further exported from the ER through the Golgi, likely with the help of cargo receptors, to reach the plasma membrane[165,166] (**Figure 1.5, top**).

1.3.2.2.3 Book 3: All good things must come to an end… recycling and degradation of MHC I molecules

Once at the plasma membrane, MHC I molecules can fulfill their role, that is to present their peptide to CD8+ T cells. But, just like any other protein complex, peptide/MHC I complexes are not eternal and mechanisms exist to ensure their proper recycling or degradation (**Figure 1.5, top right**).

In cells, the constitutive recycling of plasma membrane plays an important role in the internalization of MHC I molecules[167], though the observation that free MHC I heavy chains have shorter half-life than fully folded peptide/MHC I complexes suggests the existence of MHC I-specific internalization mechanisms[168-170]. In line with this, ubiquitination of MHC I molecules on their cytoplasmic tail by proteins of the MARCH family has been shown to promote MHC I molecule internalization[171] as well as their degradation in endolysosomes[172]. Besides degradation, some internalized (and correctly folded) MHC I molecules can also be recycled through a multi-step process that implies (i) peptide stripping from the molecule by the acidic environment of endosomes, (ii) reloading of the same molecule with endosomal degradation products and finally (iii) rerouting of this (now loaded) MHC I molecule towards the plasma membrane[173-175].

In summary, the classical MHC I antigen presentation pathway is used by *all* nucleated cells to get their inside out. Interestingly, because MAPs generated through

this pathway necessarily derive from the degradation of endogenous proteins, this system also ensures that CD8+ T cells will only attack abnormal cells, rather than otherwise healthy cells presenting MAPs derived from proteins acquired from neighboring cells.

1.3.3 To represent others, use cross-presentation or cross-dressing!

Besides presenting MAPs derived from proteins they synthesize (like most of our cells), professional APCs, such as DCs and macrophages, have been shown to present MAPs derived from exogenous proteins, i.e., proteins synthesized by other cells[176-178]. As counterintuitive as it might seem, having cells able to cross-present is actually a must-have for the adaptive immune system! Indeed, as mentioned in **section 1.2.2.2**, CD8+ T cells are mainly located in secondary lymphoid organs rather than spread throughout peripheral tissues in search of abnormal cells[179]. Thus, they rely on professional APCs to do the scanning job, that is to sample the microenvironment of peripheral tissues and bring back suspicious antigens that might activate them[180-182].

For cross-presentation, professional APCs must first internalize exogenous proteins. Common sources for such exogenous proteins are bacteria[183,184], parasites[185] and autologous dying cells[186], though proteins (or peptides) can also be acquired from live cells[187,188] through transfer across gap junctions[189] (**Figure 1.5, top left**) or internalization of exosomes[190]. Among the plethora of known protein internalization mechanisms, phagocytosis, micropinocytosis and receptor-mediated endocytosis of intact exogenous proteins seems to be more efficient than pinocytosis in this context[178,191,192]. The subsequent processing of these exogenous proteins into peptides relies on two main pathways, the phagosome-to-lysosome pathway and the vacuolar pathway[193]. In the phagosome-to-lysosome pathway, internalized proteins can (i) be transferred in the cytosol to be fully processed by the classical MHC I antigen presentation pathway (reviewed in **section 1.3.2**) or (ii) be transferred in the cytosol for degradation, pumped back (as peptides) in the endocytic compartment for further processing and final loading onto MHC I molecules. In the vacuolar pathway, the whole processing of internalized proteins into peptides as well as peptide loading onto MHC I

molecules is done within endocytic compartments, using lysosomal proteases such as cathepsin S. Thus, professional APCs using the phagosome-to-lysosome pathway are expected to generate the same set of MAPs than infected cells, while those using the vacuolar pathway might not, another way to enlarge the MAP repertoire.

Lastly, more recent evidence have shown that, besides cross-presentation, DCs can also cross-dress, i.e., steal pre-formed peptide/MHC I complexes from other cells, to activate CD8+ T cells[194]. To do so, DCs use trogocytosis, meaning that they steal plasma membrane fragments from their target cells through the establishment of intimate cell-to-cell contact[195].

In conclusion, the MHC I antigen presentation pathway can be seen as a common framework used by all cells to present their inner state, or (in few cases) the inner state of other cells, to CD8+ T cells. Importantly, the fact that this pathway relies on ‘generalist’ proteases and peptidases ensures that theoretically any protein can give rise to MAPs, thereby increasing cells’ chances to present their abnormalities.

1.4 Origin of MHC I-associated peptides: what to present to be representative?

The estimation that cells can present ~ 2% of their proteome to CD8+ T cells emphasize the fact that the MAP repertoire is a very compressed representation of the inner state of cells[196]. However, compression of this input, i.e., the proteome, into this very narrow output, i.e., the MAP repertoire, does not seem to result in a significant loss of data. Indeed, the MAP repertoire can quickly and efficiently capture changes induced by cell type[133], drug treatment[197], infections[198] or even neoplastic transformation[199] (all extensively reviewed in the article presented in **Appendix I**). Even the ploidy status of cells can affect their MAP repertoire, as demonstrated in **Appendix II**. Altogether those observations demonstrate than the MAP repertoire is highly plastic and suggest that several factors are at play when it comes to choose the proteins that will enter the MHC I antigen presentation pathway.

1.4.1 Rapidly degraded proteins, DRiPs and retirees

The observation that CD8+ T cells can eliminate infected cells as quickly as one hour post-infection[200] strongly suggests that changes in the MAP repertoire occur faster than changes in the proteome, the median protein half-life being 46h[201,202]. To reconcile this apparent paradox, Yewdell *et al.* hypothesized that MAPs primarily derive from rapidly degraded proteins (RDPs), which represent ~ 25 to 30% of all synthesized proteins[203,204]. In line with this, large-scale studies of the MAP repertoire revealed that MAPs tend to derive from efficiently translated transcripts that generate proteins bearing degradation-prone features, such as an increased density of ubiquitination sites as well as more disordered regions[205]. Besides proteins with a high turnover and excess subunits produced for protein complexes[90], RDPs also include defective ribosomal products (DRiPs), which are truncated or misfolded proteins produced by translational errors[206]. Because these DRiPs are expected to enter almost immediately the MHC I antigen presentation pathway, they are exceedingly hard to capture and their generation remains ill-defined. Nonetheless, translation-coupled mRNA destabilization appears instrumental for DRiP, and consequently MAP, generation. In line with this,

MAP source transcripts have been shown to bear more microRNA response elements than non-source transcripts[96], with the underlying idea that microRNA targeting might promote ribosome drop-off and therefore favor the formation of uncomplete protein products, that is to say DRiPs[207]. Similarly, truncated proteins produced by the translation of transcripts bearing pre-termination codon, and that are subsequently degraded by the nonsense mediated decay pathway, have been shown to efficiently generate MAPs[208].

Although the use of RDPs and DRiPs as MAP source proteins likely explain the incredible plasticity of the MAP repertoire, full-length proteins, i.e. retirees, also contribute to the generation of MAPs[209]. This is best exemplified by the detection of MAPs bearing post-translational modifications, such as phosphorylation[210,211], methylation[212], glycosylation[213], and so on[214]. Indeed, if some post-translational modifications occur spontaneously, most of them require precise interactions between the relevant enzyme and the correctly folded substrate. Thus, modified MAPs, and consequently some unmodified MAPs, likely derive from the degradation retirees rather than RDPs or DRiPs.

1.4.2 Conventional and cryptic proteins

Regardless of whether they derive from RDPs, DRiPs or retirees[209], MAPs have always been assumed to derive solely from the processing of conventional proteins, that is the in-frame translation of annotated protein-coding transcripts. Indeed, the assumption that these MAPs, referred to as *conventional MAPs*, are the sole MAPs does make sense, as no protein production clearly means no MAP! However, recent studies have highlighted that the translational activity of cells is far more complex than previously anticipated.

In 2012, the ENCODE consortium demonstrated that the transcriptome of cells contains far more tramscripts than just protein-coding transcripts. Actually, using RNA-sequencing (RNA-Seq), they showed that up to 75% of human genome can be transcribed, while protein-coding transcripts have been estimated to represent only 2% of this very same genome[215]. Moreover, although most of these additional transcripts

were considered to be non-coding, several ribosome profiling studies, performed in mice and humans, revealed that pervasive translation does occur outside of protein-coding transcripts[216,217]. Recurrent detection of these non-canonical translation events does not seem to result from experimental artefacts, as several mass spectrometry studies detected some of those atypical proteins, often referred to as cryptic proteins[218-220]. Interestingly, these cryptic proteins can derive from the translation of open reading frames (ORFs) located in (i) non-coding transcripts, (ii) non-coding regions within protein-coding transcripts, such as introns and 5'/3' untranslated regions (UTRs), as well as (iii) the out-of-frame translation of protein-coding exons[221]. As opposed to sequences coding for conventional proteins, classically defined as conserved sequences of at least 300 nucleotide-long delimited by an in-frame start (AUG) and stop codon[222], these cryptic ORFs are short (< 300 nucleotides)[218], often use near-cognate AUGs as start codons (such as CUG)[223] and show a conserved translational activity rather than a conserved nucleotide sequence[224]. Although the function and contribution of cryptic proteins to the cellular proteome remain largely unknown[221,222,225,226], these proteins being quiet difficult to capture by mass spectrometry[227], it is clear that their discovery significantly enlarges the coding capabilities of the genome and consequently drastically complexifies the proteome.

As detailed in **section 1.3.2**, the MHC I antigen presentation pathway relies on 'generalist' proteases, thus one might expect that cryptic proteins are processed by this pathway, just like conventional proteins. However, due to the very few efforts that have been made to comprehensively annotate cryptic ORFs[228], only a series of indirect evidence suggest the existence of cryptic MAPs. First, in 1989, Boon *et al.* observed that transfecting cells with tumor DNA devoid of transcriptional and translational regulatory elements did not prevent their recognition by CD8+ T cells, thereby suggesting that immunogenic MAPs were produced from those DNA fragments[229,230]. Second, Shastri *et al.* transfected murine cells with constructs encoding the SIINFEKL model peptide with various initiation contexts and demonstrated that translation at several near-cognate AUGs was sufficient to produce SIINFEKL at levels detectable by T cells[231,232]. Moreover, taking advantage of a mouse model engineered to

ubiquitously express a conventional and a cryptic MAP, the same group demonstrated that, in this context, both types of MAPs were able to induce central tolerance and prime CD8+ T cells *in vivo*[233]. Lastly, and maybe in a more physiological context, several research groups reported the existence of immunogenic MAPs derived from the translation of UTRs[234,235], introns[236-238], out-of-frame translation of coding exons[239,240] and so on[241]. Although rare, those reports clearly highlight that the processing of cryptic proteins by the MHC I antigen presentation pathway does produce cryptic MAPs and that such peptides are generated in quantities that are sufficient to support anti-tumor or autoimmune responses[242,243].

1.4.3 The complex case of tumor cells

When compared to normal cells, tumor cells have been shown (i) to re-express or overexpress transcripts and (ii) to acquire mutations that might confer a selective advantage (drivers) or not (passengers). As rare as they might be, those qualitative and quantitative changes, which propagate from the transcriptome to the proteome, are expected to trigger the processing of abnormal MAPs by the MHC I antigen presentation pathway. These MAPs, which are referred to as tumor antigens[244], can be further classified as *tumor-associated antigens* (TAAs) or *tumor-specific antigens* (TSAs)[245].

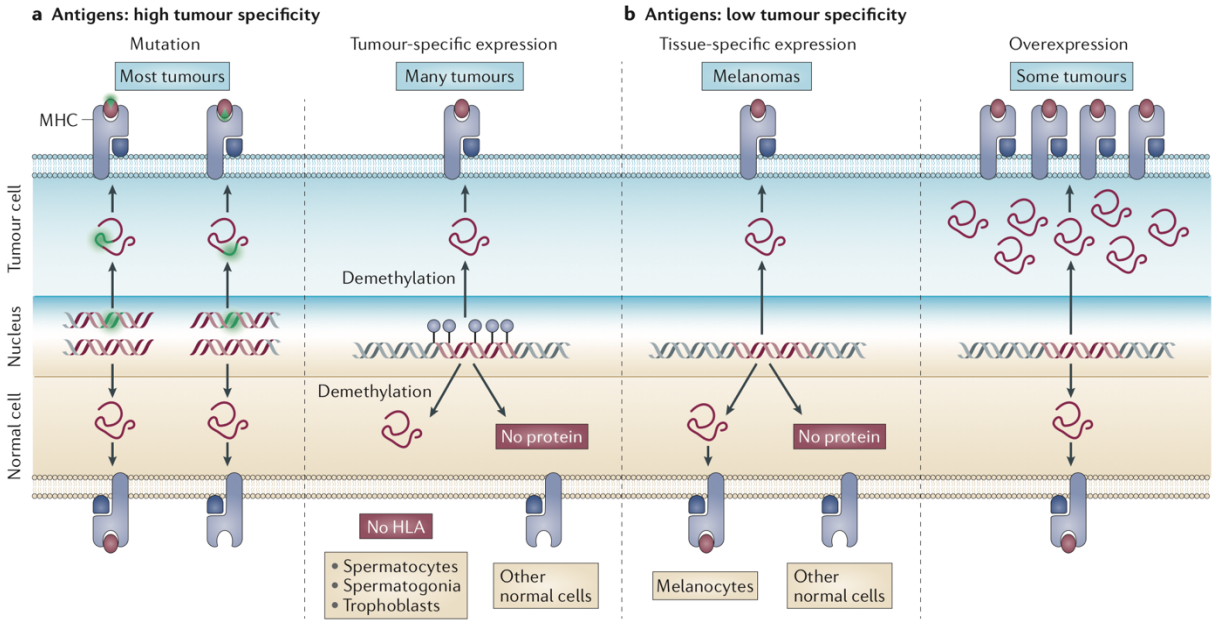


Figure 1.6 | Classes of tumour antigens recognized by T lymphocytes. (a) Mechanisms producing tumour-specific antigens (TSAs), i.e mTSAs and aeTSAs (which include cancer-testis antigens). **(b)** Mechanisms producing tumor-associated antigens (TAAs), i.e. differentiation and overexpressed. Adapted with permission from Springer Nature: Nature Reviews Cancer (Coulie PG, Van den Eynde BJ, van der Bruggen P and Boon T)[244] © 2014

1.4.3.1 Tumor-associated antigens

TAAs derive from the translation and further processing of *normal sequences* that (i) have a tissue-specific expression or that (ii) are overexpressed in cancer cells when compared to normal cells (**Figure 1.6b**). Although considered as tumor antigens, tissue-specific TAAs, often referred to as differentiation antigen, do not reflect changes induced by the neoplastic transformation but rather its tissue origin. Among others, genes with a melanocyte-restricted expression, such as Melan-A (MART-1) or gp100, have been shown to generate several of those differentiation antigens[246-248]. *A contrario*, overexpressed TAAs, which do not necessarily derive from overexpressed transcripts[249,250], are really the products of transcriptional and/or translational changes induced by the neoplastic transformation. Thus, some cancer genes, such as ERBB2 in epithelial tumors[251] or WT1 in leukaemia[252], have been shown to generate such TAAs.

Few naturally occurring CD8+ T cells responses against those TAAs have been reported up to now. Indeed, among CD8+ tumor-infiltrating lymphocytes (TILs), which are expected to be enriched for anti-tumor T cells, less than 1% of them recognize such TAAs in melanoma patients[253]. This poor immunogenicity is likely due to that fact that TAAs are *de facto* normal MAPs. Indeed, as reviewed in **section 1.2.2**, self-reactive CD8+ T cells, including TAA-reactive CD8+ T cells, have been deleted during the establishment of central tolerance[35], while peripheral tolerance mechanisms ensure the repression of the few CD8+ clones that might have escaped in periphery. Thus, recognition of TAAs is probably mediated by rare and low-avidity CD8+ T cells, thereby explaining their lack of immunogenicity and their inability to promote efficient anti-tumor responses *in vivo* (see article presented in **Appendix II** for examples of poorly immunogenic TAAs that we identified on the murine EL4 cell line).

1.4.3.2 Tumor-specific antigens

TSAs derive from the translation and further processing of (i) *tumor-specific sequences* bearing any type of non-synonymous tumor-specific mutations (mutated TSAs or mTSAs) and (ii) *normal sequences with a tumor-restricted expression* though some might be expressed in MHC I-negative tissues (aberrantly expressed TSAs or aeTSAs, **Figure 1.6a**).

mTSAs reflect the mutational profile of cancer cells in their MAP repertoire. For now, they are thought to be the best anti-tumor antigens thanks to their exquisite tumor-specificity and their neo-self origin. Accordingly, mutational load of tumors correlates with markers of CD8+ T cells infiltration and patients' survival[254-256]. For now, attempts to identify mTSAs have mainly focused on those derived from coding exons bearing single-base mutations[199,257-259] and, sometimes, insertion/deletion[260], while several other types of mutations remain poorly explored[261-263]. Although incomplete, those studies revealed that, despite their tumor-specificity and immunogenicity, *mTSAs are rare*. For instance, in melanoma, which is one of the most mutated tumor type, an average of two mTSAs per sample have been reported[199], leading some researcher to conclude that identification of mTSAs in lowly mutated tumor types is simply utopic[264].

Moreover, *mTSAs are often 'private' antigens*. Comparing the mutational profile of two cancer patients, even with those with the same cancer, always reveals great heterogeneity, with only a few drivers being shared by a certain percentage of the cohort. Accordingly, the mTSA landscape is expected to be as divergent[199,265], if not more divergent, as only few MAPs have been shown to overlap such driver mutations[266,267].

aeTSAs can be seen as perfect overexpressed TAAs because they derive from normal sequence (i.e., not mutated) that have no expression in normal tissues. For now, people have only focused on the identification of a specific type on aeTSAs, called cancer-testis antigens, as they are expressed in MHC I-negative tissues which include testis. These cancer-testis antigens appear to mediate tumor regression *in vivo*, as their expression in various tumor types correlates with markers of CD8+ and CD4+ T cell infiltration[255]. However, their identification remains extremely challenging, as demonstrating an absence of expression is nearly impossible. In line with this, Rooney *et al.* evaluated the peripheral expression of 270 cancer-testis antigens previously reported with a panel of tissues sequenced by the GTEx consortium: only 60 of those had a testis-restricted expression[268]. Nonetheless, these unmutated antigens remain highly interesting because, besides their high immunogenic potential, they also have the potential to be shared by several tumor types.

Altogether, it is clear that the MAP repertoire is able to integrate various type of input, such as the mutational, transcriptional, translational and even signalling profile of cells, to generate a very comprehensive view of the inner state of cells. Besides being comprehensive, this cell surface representation is also highly plastic, as changes in any of those input are likely to quickly modify the MAP repertoire. By scanning this representation, CD8+ T cells can therefore quickly identify and eliminate non-self or neo-self threats, that is to say infected cells presenting viral MAPs or cancer cells presenting tumor antigens. Lastly, identification of such tumor antigens, especially m/aeTSAs, should help move forward the field of cancer immunotherapy by providing targets for anticancer vaccines.

1.5 Studying the MAP repertoire: how to and what for?

As small as it might be compared to the proteome, one should not underestimate the complexity of the MAP repertoire. Indeed, besides its extreme plasticity, the MAP repertoire has been estimated to contain, on a per cell basis, 10,000 to 500,000 MHC I molecules[118], that present hundreds to thousands of different MAPs[199], including mutated and post-translationally modified MAPs, at copy number ranging from a few to a few thousands[269-271] (and unpublished data from our laboratory). Because the composition of the MAP repertoire is crucial for adaptive immunity, immunologists always invested a lot of time and efforts to better define and better understand rules governing its biogenesis.

1.5.1 Prehistoric era: dissecting the MAP repertoire… one peptide at a time!

At first, studies of the MAP repertoire were not really studies of the MAP repertoire but rather extensive studies of a few peptides within this repertoire. These studies can be roughly classified according to the type of MAP they study, namely model peptides or naturally processed MAPs.

1.5.1.1 Model peptides

Model peptides are well-characterized MAPs for which we know the sequence and possess several tools, such as stable antigen-specific CD8+ T cell clones[272] or antibodies[273], to track their processing and cell surface presentation by the relevant MHC I molecule.

Among others, researchers often use SIINFEKL, a peptide derived from the chicken ovalbumin and identified in the late 80's [274,275] that strongly binds to the murine MHC I molecule called H-2-K^b. Typical experimental designs involve the transfection of cells with constructs containing the SIINFEKL-coding region in various context followed by antigen presentation quantitation to help unravel the basic rules governing MAP generation[123,147,152]. Besides that, the use of such model peptides was instrumental to prove that non-canonical translation could generate MAPs. For instance, transfecting

cells with a library of constructs containing the SIINFEKL-coding region in various initiation context and correlate it to its cell surface presentation was crucial to establish the importance of non-canonical translation initiation as a relevant source of MAPs[231,232]. Similarly, the observation that the intronic vs. exonic location of the SIINFEKL-coding region within the β-globin gene did not alter SIINFEKL presentation in H-2-K^b-expressing HEK293 cells established that cryptic proteins derived from the nuclear translation of pre-spliced mRNAs do generate MAPs[276]. Lastly, gentamicine-induced stop codon readthrough was shown to induce translation of the SIINFEKL-coding region at levels sufficient to trigger T-cell recognition[235].

Though SIINFEKL and other model peptides should be awarded an honorary Nobel Prize for their outstanding contribution to the understanding of the ins and outs of MHC I antigen presentation pathway, one must be cautious when generalizing the conclusions obtained by those studies, as one model peptide is only one model peptide (and it might be an outlier!).

1.5.1.2 Screening for tumor antigens

Sometimes, it is not so much about gaining some mechanistic insights than determining the actual sequence of naturally processed MAPs. Tumor antigens were among the first MAPs to ignite the interest of immunologists, as their recognition by CD8+ T cells appeared to be crucial to promote tumor clearance by the adaptive immune system (see **section 1.4.3**). Thus, in hopes of developing immunotherapies targeting such tumor antigens, researchers invested a lot of resources to precisely determine their sequence.

At that time (meaning the late 80's/early 90's), the sole way forward was to use stable CD8+ T cell clones isolated from patients. Classically, reactivity of such T cell clones were screened by exposing them to antigen-negative autologous tumor cells transfected with DNA fragments isolated from antigen-positive autologous tumor cells. Upon retrieval of DNA fragments able to trigger the lysis of antigen-negative tumor cells, the same screening procedures were applied to identify the minimal DNA fragment required for T-cell lysis (by transfecting cells with various truncations of the

initial DNA fragments)[277,278]. Lastly, the exact sequence of the immunogenic epitope was determined using synthetic peptides potentially produced by the identified DNA fragments and evaluating their ability to trigger lysis by the relevant T-cell clone when exogenously loaded on target cells[279,280].

Although those studies were essential to discover the various type of tumor antigens, they are likely to be fraught with false positives[281], as screening peptide libraries does not ensure processing of the identified epitope by autologous tumor cells, while T-cell cross-reactivity suggests a one-to-many rather than a one-to-one correspondence between the T-cell space and the peptide space[85]. Besides, they involve a very tedious, and thus low-throughput, experimental design where the identification of tumor antigens must be done one antigen at a time, one tumor at a time and only for those with available stable T-cell clones.

Altogether, those pioneering studies provided the scientific community with a better understanding of the MHC I antigen presentation pathway and an extensive characterization of some tumor antigens. Although tools developed in this era are still widely used to precisely dissect the MHC I antigen presentation pathway, they would never allow for the monitoring and identification of thousands of MAPs at once. Thus, only significant technological advances could allow our field to move from these insightful, yet low-throughput, studies to systems-level analysis of the MAP repertoire.

1.5.2 Modern history: mass spectrometry-based studies of the MAP repertoire

Concomitant to the discovery of the first human tumor antigens, work performed in the laboratories of Hans-Georg Rammensee and Donald F. Hunt paved the way to systems-level analysis of the MAP repertoire. First, in several landmark papers, they developed tools to directly isolate MAPs from cells[282], to fractionate those isolated MAPs using reverse-phase high-performance liquid chromatography (HPLC)[275] and to sequence them, using automated Edman degradation[97] and, soon after, mass spectrometry (MS)[269]. And that's when the revolution started!

1.5.2.1 Direct isolation of MAPs from cells

But first things first! To sequence MAPs, you first need to isolate them. Among the plethora of strategies initially used to isolate MAPs from cells[283,284], only two are still widely used and each present their own set of strengths and weaknesses.

1.5.2.1.1 Mild acid elution (MAE)

To isolate MAPs by MAE, single-cell suspensions are shortly incubated in a citrate-phosphate buffer (pH = 3.3) to destabilize peptide/MHC I complexes exposed at the plasma membrane, while limiting cell death by lysis. Indeed, in this acidic pH, β_2m quickly dissociates from peptide/MHC I complexes, thereby releasing MAPs from the peptide-binding groove of the MHC I heavy chains. Collection of the acidic MAP extract is then followed by a desalting step, high salt concentrations being incompatible with MS analysis, and a molecular weight cut-off to remove full-length proteins, that is to say β_2m and other contaminants that might have been eluted from the cell surface.

Efficiency of the MAE procedure can be easily quantified by flow cytometry. Indeed, because free MHC I heavy chains are known to be unstable (**section 1.3.2**), they are quickly internalized by cells upon MAE. Thus, comparing the cell surface presentation of MHC I molecules in treated vs. non-treated cells should roughly estimate the percentage of MHC I molecules for which MAP were recovered by the acidic treatment. Lastly, estimation of cell death is also a must, as this is likely to reflect the extent to which the sample was contaminated by intracellular proteins and, more importantly, peptides.

1.5.2.1.2 Immunoaffinity purification

Isolation of MAPs by immunoaffinity purification can be done (i) on several fresh or frozen sample types, including dissociated cells and tumor biopsies, and (ii) for any MHC I molecules (and, let's face it, MHC II molecules) for which specific antibodies are readily available. Usually, the sample is first lysed with a non-denaturing detergent and peptide/MHC I complexes are extracted from this mixture using an immunoaffinity support on which pan-MHC I- or allotype-specific monoclonal antibodies are

immobilized. Following the required washing steps, an acidic buffer is then used to elute peptide/MHC I complexes from the support and further dissociate MAPs from their complex. As for the MAE procedure, the eluate is then filtered to remove any large proteins, i.e., mainly MHC I heavy chains and β_2m.

1.5.2.1.3 MAE? Immunoaffinity purification? Which one to choose?!

Although choosing the right isolation method depends on the ultimate goals of your study, it is clear that immunoaffinity purification presents several advantages over MAE[285,286] as it:

1. accommodates a greater variability of sample types, that is fresh or frozen single-cell suspension, tissues or tumor biopsies vs. fresh single-cell suspension for MAE[250].
2. shows a greater specificity than MAE (especially when using low amount of input sample), with contaminants representing less than 10% of MAPs analyzed by MS vs. up to 60% for MAE[250].
3. shows a greater MAP recovery than MAE, while still capturing most MAPs identified by MAE (as opposed to what was previously thought[287]).

Of note, both techniques show similar reproducibility between biological replicates and poorly accommodate really low input sample (using less 100 millions cells is a bit unrealistic). Lastly, it is important to note that immunoaffinity purification is a destructive technique, while MAE is not. Thus, temporal profiling of the MAP repertoire on a given cell population can only be performed by MAE.

1.5.2.1.4 Last steps prior to MS analysis

Prior to injection in the mass spectrometer, MAP extracts obtained by MAE or immunoaffinity purification can be processed:

1. to specifically enrich for MAPs bearing post-translational modifications, such as phosphorylation[288], which are often difficult to detect because of their low abundance in the sample.

2. to fractionate the sample, based on any relevant biochemical property of MAPs such as their charge[289,290], in order to analyze less complex peptides mixtures, called fractions, by MS and likely increase the number of MAPs subsequently identified.

Whether these optional steps are performed or not, MAPs extracts are then concentrated and further injected in the mass spectrometer.

1.5.2.2 MS analysis of the MAP repertoire, from spectrum to peptide

Tandem mass-spectrometry (MS/MS) is a fast and accurate technique that can analyze complex mixtures of peptides to retrieve information about their abundance and sequence. Such analysis is done on a mass spectrometer, instrument that can precisely measure the intensity and mass-to-charge ratio (m/z) of ions. Because of that, peptides must be ionized, often by electrospray ionization, prior to their injection in the mass spectrometer. Though the architecture of mass spectrometers can vary, instruments quickly cycle through the following tasks in order to acquire a large number of spectra[285,291]:

1. Selection of an ionized peptide, called a precursor ion, from the pool of injected ions by the first mass analyzer (quadrupole) in order to determine its intensity and m/z, that is its MS spectrum.
2. Fragmentation of this precursor ion, by collision with an inert gas, in order to obtain fragments ions, each partially covering the sequence of the initial peptide.
3. Separation and detection of the intensity and m/z of all those fragment ions in the second mass analyzer (quadrupole or orbitrap) to obtain an MS/MS spectrum.

For a given precursor ion, it Is possible to estimate its abundance, using the intensity of its MS spectrum, and its sequence, using its MS/MS spectrum. Indeed, fragments ions usually differ from one another by one amino acid. Thus, computing the mass difference between each two neighboring peaks on an MS/MS spectrum outputs a sequence of values corresponding to the mass of known amino acids and allow to

reconstruct the peptide sequence. Because current mass spectrometers output hundreds of thousands of MS and MS/MS spectra, specific software tools have been developed to estimate peptide abundance and/or sequence faster than manual annotation.

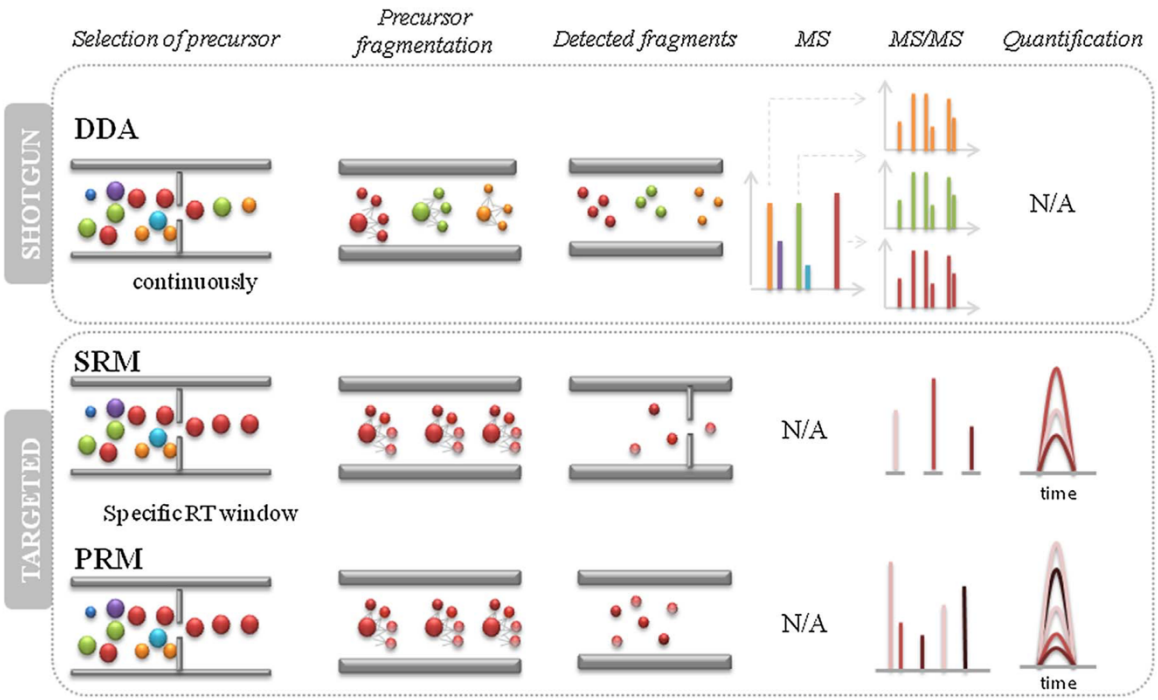


Figure 1.7 | MS acquisition modes in shotgun and targeted proteomics. In shotgun proteomics, the mass spectrometer, often operated in data dependent acquisition (DDA), continuously acquires spectra by selecting the most intense precursor ions, fragmenting them and measuring the m/z. In targeted proteomics, the mass spectrometer only selects precursor ions of interest for fragmentation. In selected reaction monitoring (SRM), you only measure the m/z of user-defined fragment ions, while in parallel reaction monitoring (PRM) you measure them all. RT: retention time. Adapted with permission from Elsevier: Journal of Chromatography B (Osinalde N, Aloria K, Omaetxebarria MJ and Kratchmarova I)[292] © 2017

1.5.2.2.1 Targeted MS for precise peptide quantitation

Selected/Multiple reaction monitoring (S/MRM)-MS and parallel reaction monitoring (PRM)-MS are two targeted MS techniques that can be used to obtain precise qualitative (presence or absence) and quantitative information for a predefined set of MAPs (**Figure 1.7, bottom**). Indeed, rather than trying to analyze everything in the sample, the mass spectrometer iteratively focuses on the extensive analysis of

specific precursor ions, that is to say MAPs in our case. These targeted MS approaches are more sensitive than shotgun MS approaches (detailed in the next section) and can monitor ~ 100–1,000 precursor ions within an experiment[293].

S/MRM-MS experiments (**Figure 1.7, bottom**) are performed on a triple quadrupole mass spectrometers and make use of a transition list, that is to say a list containing m/z of precursor/fragment ions pairs to be selected for and measured by the mass spectrometer. Upon sample ionization and entry in the mass spectrometer, one of the listed precursor ion is selected by the first quadrupole and fragmented in the second quadrupole, used as a collision cell. Then, resulting fragment ions are filtered by the third quadrupole, based on the transition list, and are further transmitted to the detector for acquisition of the partial MS/MS spectrum.

PRM-MS experiments (**Figure 1.7, bottom**) are performed on a quadrupole-Orbitrap hybrid mass spectrometer and use a precursor (rather than a transition) list, which only contains m/z of precursor ions relevant to the analysis. Upon sample ionization and entry in the mass spectrometer, one of the relevant precursor ion is selected on the first quadrupole, accumulated in the C-trap prior to fragmentation in the collision cell. Resulting fragment ions are then transferred back to the C-trap and further injected in the high-resolution Orbitrap mass analyzer for acquisition of the MS/MS spectrum.

In both cases, acquired data are then processed by specific software tools, such as Skyline[294,295], to confirm the identity (and thus the presence) of the targeted peptides and, if need be, derive quantitative information. Such quantitation can be (i) relative, meaning that the intensity of a given MAP is compared across several conditions, or (ii) absolute, meaning the intensity of a given MAP is converted into peptide copy number/cell. The implementation of absolute quantification by targeted MS is a bit more tricky than the one of relative quantification. Indeed, absolute quantification requires the concomitant analysis of endogenous and matched stable-isotope labelled synthetic peptides (usually ^{13}C- or ^{15}N-analogs), for which known quantities must be spiked in the analyzed sample prior to injection in the mass spectrometer[296]. Although S/MRM-MS and PRM-MS show more or less similar performances[297,298], PRM-MS is often

easier to implement, as labor-intensive optimization steps required for pre-selection of optimal transitions (which can significantly affect quantification) do not need to be performed[299,300].

1.5.2.2.2 Shotgun MS for peptide discovery

As opposed to targeted MS approaches which require prior knowledge of peptides to be analyzed, shotgun MS approaches just try to sequence everything there is to sequenced in the sample of interest (**Figure 1.7, top)**. These studies are often performed on high-resolution fast-scanning Orbitrap instruments, which continuously select the 10–20 most intense precursor ions for fragmentation and further separation/detection of the resulting fragment ions. Throughout the years, several fragmentation methods have been tested for the analysis of the MAP repertoire. Interestingly, electron-transfer/higher-energy collision dissociation has been shown to significantly improve the detection of MAPs when compared to collision-induced dissociation or higher-energy collisional dissociation[301].

1.5.2.2.2.1 Principles of database search engines

To determine the respective peptide sequence of the hundreds of thousands spectra acquired in shotgun MS experiments, researchers rely on database (DB)-search engines, such as Mascot[302] or Peaks[303].

Their principle is pretty simple (**Figure 1.8**): the user uploads a database, called the target database, which contain protein sequences relevant to the analyzed sample, e.g., the human proteome for the analysis human MAPs. By reversing all protein sequences in this target database, the software creates a decoy database, i.e., a database of proteins that do not exist, required to estimate the false discovery rate (FDR). From there, the software artificially chops target and decoy proteins into all possible peptides and generates their respective theoretical MS/MS spectrum. By scoring comparisons between theoretical and acquired MS/MS based on their resemblance, the software can assign a peptide sequence to each observed MS/MS spectra and the user can export a list of identifications at a specific FDR (usually 1 or 5%). This FDR is simply estimated by computing the following ratio: $FDR =$

$\left(\frac{N_{decoy}}{N_{target}}\right) \times 100$, with N_{decoy} and N_{target} the number of identifications made in the decoy and target database, respectively[304].

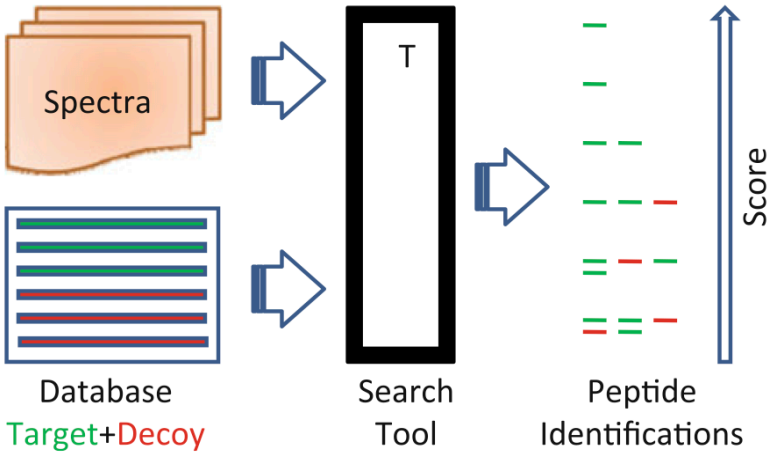


Figure 1.8 | Principle of MS/MS database searches using the target-decoy approach. To report a list of scored identifications, a search tool (T) uses a list of all acquired MS/MS spectra and a protein database containing both real (target, in green) and fake sequences (decoy, in red). The proportion of target vs. decoy identifications is then used to estimate T's error rate. Panel (**a**) adapted with permission from Springer Nature: Journal of The American Association for Mass Spectrometry (Gupta N, Bandeira N, Keich U and Pevzner PA)[305] © 2011

1.5.2.2.2.2 The advent of proteogenomics

Though this approach is reliable and can identify thousands of MAPs throughout multiple studies[90,96,133,250], one should not forget that, with this approach, you can only find peptides derived from proteins listed in your database! Indeed, it is widely known that only ~50% of acquired MS/MS spectra can be assigned to a peptide sequence. Because researchers mainly use the reference proteome (i.e., a generic proteome containing only conventional proteins) as a database, some of those unmatched spectra are likely to derive from atypical MAPs encompassing (i) polymorphic MAPs, such as mTSAs, and (ii) cryptic MAPs. This realization pushed forward the development of proteogenomic approaches, which combine MS and high-throughput DNA/RNA-sequencing technologies to build customized databases, thereby allowing the identification of such atypical MAPs (**Figure 1.9**).

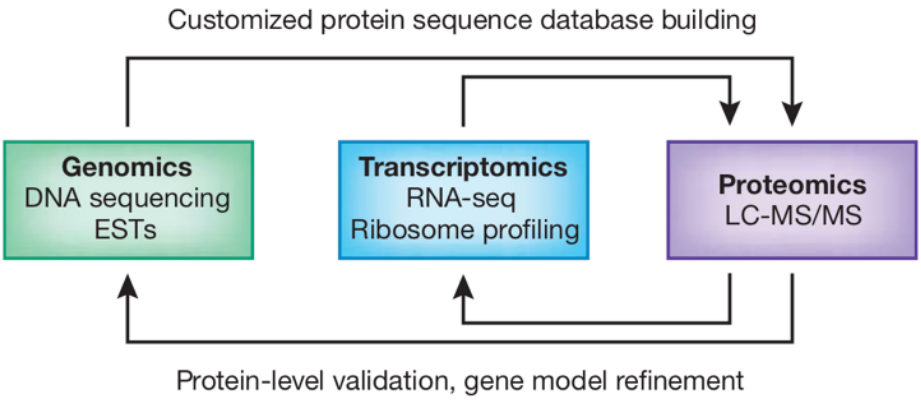


Figure 1.9 | The concept of proteogenomics. Building a customized protein sequence database from genomic and/or transcriptomic data facilitates the identification of novel proteins or MAPs, in our case. In turn, we can improve traditional proteomic analyses by generating more complete versions of protein sequence databases such as UniProt. Adapted with permission from Springer Nature: Nature Methods (Nesvizhskii AI)[306] © 2014

For now, proteogenomic approaches have mainly been used to identify mTSAs, probably because of their clinical potential. In parallel to the MS analysis of MAPs isolated from a tumor sample, the exome and/or transcriptome of the same tumor and its normal-matched counterpart are sequenced. Following the classical processing of these sequencing data, high-quality cancer-specific mutations are identified and further inserted in the reference genome to generate a customized genome. From there, annotations of protein-coding transcripts are then used to guide the *in silico* translation of these genomic sequences. The resulting customized proteins are then concatenated into a fasta file that can be submitted to the DB-search engine in place of the classical reference proteome. Although successful[199,257,307,308], this approach only allows for the identification of conventional mTSAs and can still be considered as restrictive. But, this is not what matters anymore… what really matters is the fact that researchers are aware that customized database can help characterizing the MAP repertoire! With this powerful idea in mind, sky is the limit! And MS-based studied, such as the one that confirmed the existence of MAPs derived from non-contiguous genomic that can be spliced within the proteasome (proteasome-spliced peptides), is a good example of it[309].

MS is clearly a powerful tool to study all dimensions of the MAP repertoire. Indeed, as opposed to reductionist approaches studying one peptide at a time, MS provides a plethora of information on thousands of MAPs within a few hours! These information, which include MAP sequence, abundance, post-translational modification(s) as well as variation of those parameters over time or experimental conditions, should help systems immunologists understanding the rules governing MAPs biogenesis[205]. Lastly, with the advent of proteogenomics, capturing the entirety of the MAP repertoire does not sound like a dream anymore.

1.5.3 Science fiction: predicting the MAP repertoire

Most data gathered by MS-based studies of the MAP repertoire are readily available in databases such as the Immune epitope database[310], which contains about ~300,000 MAPs, or the newly created systMHC portal[311]. With such large datasets, as well as complementary datasets, several research groups have developed machine learning-based tools trying to predict every step of the MHC I antigen presentation pathway, including (i) proteasomal cleavage[312], (ii) translocation by TAP[313] and (iii) MHC I binding[314-317] or the overall pathway[318,319]. These algorithms, which take user-defined protein or peptide sequences as input, can only predict the sequential 'physical' steps of the MHC I antigen presentation pathway and do not consider the cellular context, a parameter known to drastically affect the composition of the MAP repertoire.

Nonetheless, algorithms predicting peptide binding to MHC I molecules are often used to predict TSAs from tumor sequencing data, as MS requires infinitely more material that sequencing to start with[258,265,320-322]. Briefly, the mutational profile of tumor cells, often restricted to single-base mutations, and annotations of protein-coding transcripts are used to *in silico* generate mutation-overlapping peptides. Criteria used to prioritize the resulting TSA candidates often include high transcript expression and good predicted binding affinity for one the relevant MHC I molecules expressed by the patient (inferred from sequencing data using tools such as OptiType[323]). Although really fast and sample-economic, these prediction-based TSA identification workflow are

thought to be fraught with ~90% of false positives, likely because peptide binding does mean processing[264,324].

Thus, for now, more work is needed to understand the link between the MAP repertoire and its -ome precursors, i.e., the degradome, the proteome, the translatome, the transcriptome and the genome, before dreaming of algorithms able to predict it in all its dimensions, including post-translational modifications and copy number!

1.6 Objectives

According to the dogma, the human genome contains coding and non-coding regions, which are thought to represent 2% and 98% of it, respectively. However, strong evidence have established that cells can transcribe up to 75% of their genome and that pervasive translation does occur outside of protein-coding transcripts. Proteogenomics, which allowed for the detection of numerous cryptic proteins, i.e., derived from the non-coding genome, brought the definite proof that non-canonical translation does complexify the cellular proteome. Despite this blurred line between what should be considered coding and non-coding in our genome, systems-level studies of the MAP repertoire have only focused on the identification of conventional MAPs, that is MAPs derived from the canonical translation of protein-coding transcripts, with only few reductionist studies reporting the existence of cryptic MAPs.

1.6.1 General objective

The general objective of this thesis was *to provide the first systems-level analysis of the cryptic MAP repertoire in normal and cancer cells*.

1.6.2 Specific objectives

Objective 1: To evaluate the breadth of the cryptic MAP repertoire in normal cells (**Chapter 2**)

Objective 2: To provide a better understanding of cryptic MAPs biogenesis (**Chapter 2 and 3**)

Objective 3: To demonstrate that cryptic MAPs are valuable target for T-cell based cancer immunotherapy (**Chapter 3 and 4**)

1.7 References

1. Boehm, T., *Quality control in self/nonself discrimination.* Cell, 2006. **125**(5): p. 845-58.
2. Marraffini, L.A. and E.J. Sontheimer, *Self versus non-self discrimination during CRISPR RNA-directed immunity.* Nature, 2010. **463**(7280): p. 568-71.
3. Barrangou, R. and L.A. Marraffini, *CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity.* Mol Cell, 2014. **54**(2): p. 234-44.
4. Janeway, C.A., Jr., *The immune system evolved to discriminate infectious nonself from noninfectious self.* Immunol Today, 1992. **13**(1): p. 11-6.
5. Akira, S., S. Uematsu, and O. Takeuchi, *Pathogen recognition and innate immunity.* Cell, 2006. **124**(4): p. 783-801.
6. Ahmad-Nejad, P., et al., *Bacterial CpG-DNA and lipopolysaccharides activate Toll-like receptors at distinct cellular compartments.* Eur J Immunol, 2002. **32**(7): p. 1958-68.
7. Heil, F., et al., *Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8.* Science, 2004. **303**(5663): p. 1526-9.
8. Kumar, H., T. Kawai, and S. Akira, *Pathogen recognition in the innate immune response.* Biochem J, 2009. **420**(1): p. 1-16.
9. Boldrick, J.C., et al., *Stereotyped and specific gene expression programs in human innate immune responses to bacteria.* Proc Natl Acad Sci U S A, 2002. **99**(2): p. 972-7.
10. Barton, G.M., J.C. Kagan, and R. Medzhitov, *Intracellular localization of Toll-like receptor 9 prevents recognition of self DNA but facilitates access to viral DNA.* Nat Immunol, 2006. **7**(1): p. 49-56.
11. Lee, H.R., et al., *Viral Inhibition of PRR-Mediated Innate Immune Response: Learning from KSHV Evasion Strategies.* Mol Cells, 2016. **39**(11): p. 777-782.
12. Beachboard, D.C. and S.M. Horner, *Innate immune evasion strategies of DNA and RNA viruses.* Curr Opin Microbiol, 2016. **32**: p. 113-119.
13. Cooper, M.D. and M.N. Alder, *The evolution of adaptive immune systems.* Cell, 2006. **124**(4): p. 815-22.

14. Iwasaki, A. and R. Medzhitov, *Control of adaptive immunity by the innate immune system.* Nat Immunol, 2015. **16**(4): p. 343-53.
15. Farber, D.L., et al., *Immunological memory: lessons from the past and a look to the future.* Nat Rev Immunol, 2016. **16**(2): p. 124-8.
16. Gebhardt, T., et al., *Memory T cells in nonlymphoid tissue that provide enhanced local immunity during infection with herpes simplex virus.* Nat Immunol, 2009. **10**(5): p. 524-30.
17. Cooper, M.D., R.D. Peterson, and R.A. Good, *Delineation of the Thymic and Bursal Lymphoid Systems in the Chicken.* Nature, 1965. **205**: p. 143-6.
18. Cooper, M.D., *The early history of B cells.* Nat Rev Immunol, 2015. **15**(3): p. 191-7.
19. Nutt, S.L., et al., *The generation of antibody-secreting plasma cells.* Nat Rev Immunol, 2015. **15**(3): p. 160-71.
20. Kumar, B.V., T.J. Connors, and D.L. Farber, *Human T Cell Development, Localization, and Function throughout Life.* Immunity, 2018. **48**(2): p. 202-213.
21. Schatz, D.G. and Y. Ji, *Recombination centres and the orchestration of V(D)J recombination.* Nat Rev Immunol, 2011. **11**(4): p. 251-63.
22. Arstila, T.P., et al., *A direct estimate of the human alphabeta T cell receptor diversity.* Science, 1999. **286**(5441): p. 958-61.
23. Elhanati, Y., et al., *Inferring processes underlying B-cell repertoire diversity.* Philos Trans R Soc Lond B Biol Sci, 2015. **370**(1676).
24. Laydon, D.J., C.R. Bangham, and B. Asquith, *Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach.* Philos Trans R Soc Lond B Biol Sci, 2015. **370**(1675).
25. Bianconi, E., et al., *An estimation of the number of cells in the human body.* Ann Hum Biol, 2013. **40**(6): p. 463-71.
26. Nikolich-Zugich, J., M.K. Slifka, and I. Messaoudi, *The many important facets of T-cell repertoire diversity.* Nat Rev Immunol, 2004. **4**(2): p. 123-32.
27. Nemazee, D., *Mechanisms of central tolerance for B cells.* Nat Rev Immunol, 2017. **17**(5): p. 281-294.

28. Vantourout, P. and A. Hayday, *Six-of-the-best: unique contributions of gammadelta T cells to immunology.* Nat Rev Immunol, 2013. **13**(2): p. 88-100.
29. Nielsen, M.M., D.A. Witherden, and W.L. Havran, *gammadelta T cells in homeostasis and host defence of epithelial barrier tissues.* Nat Rev Immunol, 2017. **17**(12): p. 733-745.
30. Davis, M.M., et al., *T cells as a self-referential, sensory organ.* Annu Rev Immunol, 2007. **25**: p. 681-95.
31. Holler, P.D., L.K. Chlewicki, and D.M. Kranz, *TCRs with high affinity for foreign pMHC show self-reactivity.* Nat Immunol, 2003. **4**(1): p. 55-62.
32. Townsend, A.R., et al., *The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides.* Cell, 1986. **44**(6): p. 959-68.
33. Surh, C.D. and J. Sprent, *Homeostasis of naive and memory T cells.* Immunity, 2008. **29**(6): p. 848-62.
34. Stefanova, I., J.R. Dorfman, and R.N. Germain, *Self-recognition promotes the foreign antigen sensitivity of naive T lymphocytes.* Nature, 2002. **420**(6914): p. 429-34.
35. Anderson, M.S. and M.A. Su, *AIRE expands: new roles in immune tolerance and beyond.* Nat Rev Immunol, 2016. **16**(4): p. 247-58.
36. Mueller, D.L., *Mechanisms maintaining peripheral tolerance.* Nat Immunol, 2010. **11**(1): p. 21-7.
37. Nagamine, K., et al., *Positional cloning of the APECED gene.* Nat Genet, 1997. **17**(4): p. 393-8.
38. Finnish-German, A.C., *An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains.* Nat Genet, 1997. **17**(4): p. 399-403.
39. Bennett, C.L., et al., *The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3.* Nat Genet, 2001. **27**(1): p. 20-1.

40. Abramson, S., R.G. Miller, and R.A. Phillips, *The identification in adult bone marrow of pluripotent and restricted stem cells of the myeloid and lymphoid systems.* J Exp Med, 1977. **145**(6): p. 1567-79.
41. Murray, J.M., et al., *Naive T cells are maintained by thymic output in early ages but by proliferation without phenotypic change after age twenty.* Immunol Cell Biol, 2003. **81**(6): p. 487-95.
42. von Boehmer, H., *The developmental biology of T lymphocytes.* Annu Rev Immunol, 1988. **6**: p. 309-26.
43. Krueger, A., N. Zietara, and M. Lyszkiewicz, *T Cell Development by the Numbers.* Trends Immunol, 2017. **38**(2): p. 128-139.
44. Klein, L., et al., *Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see).* Nat Rev Immunol, 2014. **14**(6): p. 377-91.
45. Cosgrove, D., et al., *The thymic compartment responsible for positive selection of CD4+ T cells.* Int Immunol, 1992. **4**(6): p. 707-10.
46. Laufer, T.M., et al., *Unopposed positive selection and autoreactivity in mice expressing class II MHC only on thymic cortex.* Nature, 1996. **383**(6595): p. 81-5.
47. Capone, M., et al., *Dissociation of thymic positive and negative selection in transgenic mice expressing major histocompatibility complex class I molecules exclusively on thymic cortical epithelial cells.* Blood, 2001. **97**(5): p. 1336-42.
48. Hogquist, K.A., T.A. Baldwin, and S.C. Jameson, *Central tolerance: learning self-control in the thymus.* Nat Rev Immunol, 2005. **5**(10): p. 772-82.
49. Murata, S., et al., *Regulation of CD8+ T cell development by thymus-specific proteasomes.* Science, 2007. **316**(5829): p. 1349-53.
50. Nakagawa, T., et al., *Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus.* Science, 1998. **280**(5362): p. 450-3.
51. Gommeaux, J., et al., *Thymus-specific serine protease regulates positive selection of a subset of CD4+ thymocytes.* Eur J Immunol, 2009. **39**(4): p. 956-64.
52. Murata, S., et al., *The immunoproteasome and thymoproteasome: functions, evolution and human disease.* Nat Immunol, 2018.

53. Koble, C. and B. Kyewski, *The thymic medulla: a unique microenvironment for intercellular self-antigen transfer.* J Exp Med, 2009. **206**(7): p. 1505-13.
54. Hubert, F.X., et al., *Aire regulates the transfer of antigen from mTECs to dendritic cells for induction of thymic tolerance.* Blood, 2011. **118**(9): p. 2462-72.
55. Anderson, M.S., et al., *Projection of an immunological self shadow within the thymus by the aire protein.* Science, 2002. **298**(5597): p. 1395-401.
56. Takaba, H., et al., *Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance.* Cell, 2015. **163**(4): p. 975-87.
57. Fan, Y., et al., *Thymus-specific deletion of insulin induces autoimmune diabetes.* EMBO J, 2009. **28**(18): p. 2812-24.
58. DeVoss, J., et al., *Spontaneous autoimmunity prevented by thymic expression of a single self-antigen.* J Exp Med, 2006. **203**(12): p. 2727-35.
59. Pinto, S., et al., *Overlapping gene coexpression patterns in human medullary thymic epithelial cells generate self-antigen diversity.* Proc Natl Acad Sci U S A, 2013. **110**(37): p. E3497-505.
60. Brennecke, P., et al., *Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells.* Nat Immunol, 2015. **16**(9): p. 933-41.
61. Yu, W., et al., *Clonal Deletion Prunes but Does Not Eliminate Self-Specific alphabeta CD8(+) T Lymphocytes.* Immunity, 2015. **42**(5): p. 929-41.
62. Mandl, J.N., et al., *T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens.* Immunity, 2013. **38**(2): p. 263-274.
63. Fulton, R.B., et al., *The TCR's sensitivity to self peptide-MHC dictates the ability of naive CD8(+) T cells to respond to foreign antigens.* Nat Immunol, 2015. **16**(1): p. 107-17.
64. Bullock, T.N., T.A. Colella, and V.H. Engelhard, *The density of peptides displayed by dendritic cells affects immune responses to human tyrosinase and gp100 in HLA-A2 transgenic mice.* J Immunol, 2000. **164**(5): p. 2354-61.

65. Bullock, T.N., D.W. Mullins, and V.H. Engelhard, *Antigen density presented by dendritic cells in vivo differentially affects the number and avidity of primary, memory, and recall CD8+ T cells.* J Immunol, 2003. **170**(4): p. 1822-9.
66. Domogalla, M.P., et al., *Tolerance through Education: How Tolerogenic Dendritic Cells Shape Immunity.* Front Immunol, 2017. **8**: p. 1764.
67. Kurts, C., et al., *Class I-restricted cross-presentation of exogenous self-antigens leads to deletion of autoreactive CD8(+) T cells.* J Exp Med, 1997. **186**(2): p. 239-45.
68. Kurts, C., et al., *The peripheral deletion of autoreactive CD8+ T cells induced by cross-presentation of self-antigens involves signaling through CD95 (Fas, Apo-1).* J Exp Med, 1998. **188**(2): p. 415-20.
69. Steinbrink, K., et al., *CD4(+) and CD8(+) anergic T cells induced by interleukin-10-treated human dendritic cells display antigen-specific suppressor activity.* Blood, 2002. **99**(7): p. 2468-76.
70. Kretschmer, K., et al., *Inducing and expanding regulatory T cell populations by foreign antigen.* Nat Immunol, 2005. **6**(12): p. 1219-27.
71. Takahashi, T., et al., *Immunologic self-tolerance maintained by CD25+CD4+ naturally anergic and suppressive T cells: induction of autoimmune disease by breaking their anergic/suppressive state.* Int Immunol, 1998. **10**(12): p. 1969-80.
72. Chaudhry, A., et al., *Interleukin-10 signaling in regulatory T cells is required for suppression of Th17 cell-mediated inflammation.* Immunity, 2011. **34**(4): p. 566-78.
73. Read, S., V. Malmstrom, and F. Powrie, *Cytotoxic T lymphocyte-associated antigen 4 plays an essential role in the function of CD25(+)CD4(+) regulatory cells that control intestinal inflammation.* J Exp Med, 2000. **192**(2): p. 295-302.
74. Grossman, W.J., et al., *Human T regulatory cells can use the perforin pathway to cause autologous target cell death.* Immunity, 2004. **21**(4): p. 589-601.
75. Pandiyan, P., et al., *CD4+CD25+Foxp3+ regulatory T cells induce cytokine deprivation-mediated apoptosis of effector CD4+ T cells.* Nat Immunol, 2007. **8**(12): p. 1353-62.

76. Grohmann, U., et al., *CTLA-4-Ig regulates tryptophan catabolism in vivo.* Nat Immunol, 2002. **3**(11): p. 1097-101.
77. Fallarino, F., et al., *T cell apoptosis by tryptophan catabolism.* Cell Death Differ, 2002. **9**(10): p. 1069-77.
78. Deaglio, S., et al., *Adenosine generation catalyzed by CD39 and CD73 expressed on regulatory T cells mediates immune suppression.* J Exp Med, 2007. **204**(6): p. 1257-65.
79. Qureshi, O.S., et al., *Trans-endocytosis of CD80 and CD86: a molecular basis for the cell-extrinsic function of CTLA-4.* Science, 2011. **332**(6029): p. 600-3.
80. Matheu, M.P., et al., *Imaging regulatory T cell dynamics and CTLA4-mediated suppression of T cell priming.* Nat Commun, 2015. **6**: p. 6219.
81. Misra, N., et al., *Cutting edge: human CD4+CD25+ T cells restrain the maturation and antigen-presenting function of dendritic cells.* J Immunol, 2004. **172**(8): p. 4676-80.
82. Onishi, Y., et al., *Foxp3+ natural regulatory T cells preferentially form aggregates on dendritic cells in vitro and actively inhibit their maturation.* Proc Natl Acad Sci U S A, 2008. **105**(29): p. 10113-8.
83. Waterhouse, P., et al., *Lymphoproliferative disorders with early lethality in mice deficient in Ctla-4.* Science, 1995. **270**(5238): p. 985-8.
84. Sage, P.T., et al., *Dendritic Cell PD-L1 Limits Autoimmunity and Follicular T Cell Differentiation and Function.* J Immunol, 2018. **200**(8): p. 2592-2602.
85. Sewell, A.K., *Why must T cells be cross-reactive?* Nat Rev Immunol, 2012. **12**(9): p. 669-77.
86. Roche, P.A. and K. Furuta, *The ins and outs of MHC class II-mediated antigen processing and presentation.* Nat Rev Immunol, 2015. **15**(4): p. 203-16.
87. Munz, C., *Autophagy Beyond Intracellular MHC Class II Antigen Presentation.* Trends Immunol, 2016. **37**(11): p. 755-763.
88. Steimle, V., et al., *Regulation of MHC class II expression by interferon-gamma mediated by the transactivator gene CIITA.* Science, 1994. **265**(5168): p. 106-9.
89. Gao, G.F., et al., *Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2.* Nature, 1997. **387**(6633): p. 630-4.

90. Bassani-Sternberg, M., et al., *Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.* Mol Cell Proteomics, 2015. **14**(3): p. 658-73.
91. Halenius, A., C. Gerke, and H. Hengel, *Classical and non-classical MHC I molecule manipulation by human cytomegalovirus: so many targets-but how many arrows in the quiver?* Cell Mol Immunol, 2015. **12**(2): p. 139-53.
92. Robinson, J., et al., *The IPD and IMGT/HLA database: allele variant databases.* Nucleic Acids Res, 2015. **43**(Database issue): p. D423-31.
93. Hughes, A.L. and M. Nei, *Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection.* Nature, 1988. **335**(6186): p. 167-70.
94. Buhler, S., J.M. Nunes, and A. Sanchez-Mazas, *HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection.* Immunogenetics, 2016. **68**(6-7): p. 401-416.
95. van Deutekom, H.W. and C. Kesmir, *Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most?* Immunogenetics, 2015. **67**(8): p. 425-36.
96. Granados, D.P., et al., *MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements.* Blood, 2012. **119**(26): p. e181-191.
97. Falk, K., et al., *Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.* Nature, 1991. **351**(6324): p. 290-6.
98. Rammensee, H., et al., *SYFPEITHI: database for MHC ligands and peptide motifs.* Immunogenetics, 1999. **50**(3-4): p. 213-9.
99. Guo, H.C., et al., *Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle.* Nature, 1992. **360**(6402): p. 364-6.
100. Mitaksov, V. and D.H. Fremont, *Structural definition of the H-2Kd peptide-binding motif.* J Biol Chem, 2006. **281**(15): p. 10618-25.
101. Sidney, J., et al., *Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries.* Immunome Res, 2008. **4**: p. 2.

102. Madden, D.R., D.N. Garboczi, and D.C. Wiley, *The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2.* Cell, 1993. **75**(4): p. 693-708.
103. Paul, S., et al., *HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity.* J Immunol, 2013. **191**(12): p. 5831-9.
104. Doherty, P.C. and R.M. Zinkernagel, *Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex.* Nature, 1975. **256**(5512): p. 50-2.
105. McClelland, E.E., D.J. Penn, and W.K. Potts, *Major histocompatibility complex heterozygote superiority during coinfection.* Infect Immun, 2003. **71**(4): p. 2079-86.
106. Penn, D.J., K. Damjanovich, and W.K. Potts, *MHC heterozygosity confers a selective advantage against multiple-strain infections.* Proc Natl Acad Sci U S A, 2002. **99**(17): p. 11260-4.
107. Hill, A.V., et al., *Molecular analysis of the association of HLA-B53 and resistance to severe malaria.* Nature, 1992. **360**(6403): p. 434-9.
108. Goulder, P.J. and B.D. Walker, *HIV and HLA class I: an evolving relationship.* Immunity, 2012. **37**(3): p. 426-40.
109. Evans, D.M., et al., *Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility.* Nat Genet, 2011. **43**(8): p. 761-7.
110. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biol, 2017. **18**(1): p. 76.
111. Jackson, M.R., et al., *Regulation of MHC class I transport by the molecular chaperone, calnexin (p88, IP90).* Science, 1994. **263**(5145): p. 384-7.
112. Lindquist, J.A., et al., *ER-60, a chaperone with thiol-dependent reductase activity involved in MHC class I assembly.* EMBO J, 1998. **17**(8): p. 2186-95.
113. Vassilakos, A., et al., *The molecular chaperone calnexin facilitates folding and assembly of class I histocompatibility molecules.* EMBO J, 1996. **15**(7): p. 1495-506.

114. Farmery, M.R., et al., *The role of ERp57 in disulfide bond formation during the assembly of major histocompatibility complex class I in a synchronized semipermeabilized cell translation system.* J Biol Chem, 2000. **275**(20): p. 14933-8.
115. Powis, S.J., et al., *Restoration of antigen presentation to the mutant cell line RMA-S by an MHC-linked transporter.* Nature, 1991. **354**(6354): p. 528-31.
116. Blees, A., et al., *Structure of the human MHC-I peptide-loading complex.* Nature, 2017. **551**(7681): p. 525-528.
117. Wearsch, P.A. and P. Cresswell, *The quality control of MHC class I peptide loading.* Curr Opin Cell Biol, 2008. **20**(6): p. 624-31.
118. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nat Rev Immunol, 2011. **11**(12): p. 823-36.
119. Rock, K.L., et al., *Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules.* Cell, 1994. **78**(5): p. 761-71.
120. Caron, E., et al., *The structure and location of SIMP/STT3B account for its prominent imprint on the MHC I immunopeptidome.* Int Immunol, 2005. **17**(12): p. 1583-96.
121. Tai, H.C. and E.M. Schuman, *Ubiquitin, the proteasome and protein degradation in neuronal function and dysfunction.* Nat Rev Neurosci, 2008. **9**(11): p. 826-38.
122. Kisselev, A.F., et al., *The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation.* J Biol Chem, 1999. **274**(6): p. 3363-71.
123. Cascio, P., et al., *26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide.* EMBO J, 2001. **20**(10): p. 2357-66.
124. Stoltze, L., et al., *Two new proteases in the MHC class I processing pathway.* Nat Immunol, 2000. **1**(5): p. 413-8.

125. Beninga, J., K.L. Rock, and A.L. Goldberg, *Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase.* J Biol Chem, 1998. **273**(30): p. 18734-42.
126. York, I.A., et al., *The cytosolic endopeptidase, thimet oligopeptidase, destroys antigenic peptides and limits the extent of MHC class I antigen presentation.* Immunity, 2003. **18**(3): p. 429-40.
127. Reits, E., et al., *Peptide diffusion, protection, and degradation in nuclear and cytoplasmic compartments before antigen presentation by MHC class I.* Immunity, 2003. **18**(1): p. 97-108.
128. Geiss-Friedlander, R., et al., *The cytoplasmic peptidase DPP9 is rate-limiting for degradation of proline-containing peptides.* J Biol Chem, 2009. **284**(40): p. 27211-9.
129. Aki, M., et al., *Interferon-gamma induces different subunit organizations and functional diversity of proteasomes.* J Biochem, 1994. **115**(2): p. 257-69.
130. Shin, E.C., et al., *Virus-induced type I IFN stimulates generation of immunoproteasomes at the site of infection.* J Clin Invest, 2006. **116**(11): p. 3006-14.
131. Driscoll, J., et al., *MHC-linked LMP gene products specifically alter peptidase activities of the proteasome.* Nature, 1993. **365**(6443): p. 262-4.
132. Gaczynska, M., K.L. Rock, and A.L. Goldberg, *Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes.* Nature, 1993. **365**(6443): p. 264-7.
133. de Verteuil, D., et al., *Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules.* Mol Cell Proteomics, 2010. **9**(9): p. 2034-47.
134. Kincaid, E.Z., et al., *Mice completely lacking immunoproteasomes show major changes in antigen presentation.* Nat Immunol, 2011. **13**(2): p. 129-35.
135. Basler, M., et al., *Immunoproteasomes down-regulate presentation of a subdominant T cell epitope from lymphocytic choriomeningitis virus.* J Immunol, 2004. **173**(6): p. 3925-34.

136. Strehl, B., et al., *Immunoproteasomes are essential for clearance of Listeria monocytogenes in nonlymphoid tissues but not for induction of bacteria-specific CD8+ T cells.* J Immunol, 2006. **177**(9): p. 6238-44.
137. Xing, Y., S.C. Jameson, and K.A. Hogquist, *Thymoproteasome subunit-beta5T generates peptide-MHC complexes specialized for positive selection.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 6979-84.
138. Sasaki, K., et al., *Thymoproteasomes produce unique peptide motifs for positive selection of CD8(+) T cells.* Nat Commun, 2015. **6**: p. 7484.
139. van Endert, P., *Post-proteasomal and proteasome-independent generation of MHC class I ligands.* Cell Mol Life Sci, 2011. **68**(9): p. 1553-67.
140. Seifert, U., et al., *An essential role for tripeptidyl peptidase in the generation of an MHC class I epitope.* Nat Immunol, 2003. **4**(4): p. 375-9.
141. Parmentier, N., et al., *Production of an antigenic peptide by insulin-degrading enzyme.* Nat Immunol, 2010. **11**(5): p. 449-54.
142. Lopez, D., et al., *Caspases in virus-infected cells contribute to recognition by CD8+ T lymphocytes.* J Immunol, 2010. **184**(9): p. 5193-9.
143. Kessler, J.H., et al., *Antigen processing by nardilysin and thimet oligopeptidase generates cytotoxic T cell epitopes.* Nat Immunol, 2011. **12**(1): p. 45-53.
144. Vigneron, N. and B.J. Van den Eynde, *Proteasome subtypes and the processing of tumor antigens: increasing antigenic diversity.* Curr Opin Immunol, 2012. **24**(1): p. 84-91.
145. van Endert, P.M., et al., *A sequential model for peptide binding and transport by the transporters associated with antigen processing.* Immunity, 1994. **1**(6): p. 491-500.
146. Chang, S.C., et al., *The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism.* Proc Natl Acad Sci U S A, 2005. **102**(47): p. 17107-12.
147. Saric, T., et al., *An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides.* Nat Immunol, 2002. **3**(12): p. 1169-76.

148. Saveanu, L., et al., *Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum.* Nat Immunol, 2005. **6**(7): p. 689-97.
149. Evnouchidou, I., et al., *ERAP1-ERAP2 dimerization increases peptide-trimming efficiency.* J Immunol, 2014. **193**(2): p. 901-8.
150. Andres, A.M., et al., *Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation.* PLoS Genet, 2010. **6**(10): p. e1001157.
151. Serwold, T., et al., *ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum.* Nature, 2002. **419**(6906): p. 480-3.
152. York, I.A., et al., *The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues.* Nat Immunol, 2002. **3**(12): p. 1177-84.
153. Nagarajan, N.A., et al., *ERAAP Shapes the Peptidome Associated with Classical and Nonclassical MHC Class I Molecules.* J Immunol, 2016.
154. Park, B., et al., *A single polymorphic residue within the peptide-binding cleft of MHC class I molecules determines spectrum of tapasin dependence.* J Immunol, 2003. **170**(2): p. 961-8.
155. Zernich, D., et al., *Natural HLA class I polymorphism controls the pathway of antigen presentation and susceptibility to viral evasion.* J Exp Med, 2004. **200**(1): p. 13-24.
156. Sieker, F., S. Springer, and M. Zacharias, *Comparative molecular dynamics analysis of tapasin-dependent and -independent MHC class I alleles.* Protein Sci, 2007. **16**(2): p. 299-308.
157. Peaper, D.R., P.A. Wearsch, and P. Cresswell, *Tapasin and ERp57 form a stable disulfide-linked dimer within the MHC class I peptide-loading complex.* EMBO J, 2005. **24**(20): p. 3613-23.
158. Williams, A.P., et al., *Optimization of the MHC class I peptide cargo is dependent on tapasin.* Immunity, 2002. **16**(4): p. 509-20.

159. Wearsch, P.A. and P. Cresswell, *Selective loading of high-affinity peptides onto major histocompatibility complex class I molecules by the tapasin-ERp57 heterodimer.* Nat Immunol, 2007. **8**(8): p. 873-81.
160. Hermann, C., et al., *TAPBPR alters MHC class I peptide presentation by functioning as a peptide exchange catalyst.* Elife, 2015. **4**.
161. Boyle, L.H., et al., *Tapasin-related protein TAPBPR is an additional component of the MHC class I presentation pathway.* Proc Natl Acad Sci U S A, 2013. **110**(9): p. 3465-70.
162. Fleischmann, G., et al., *Mechanistic Basis for Epitope Proofreading in the Peptide-Loading Complex.* J Immunol, 2015. **195**(9): p. 4503-13.
163. Teng, M.S., et al., *A human TAPBP (TAPASIN)-related gene, TAPBP-R.* Eur J Immunol, 2002. **32**(4): p. 1059-68.
164. Thomas, C. and R. Tampe, *Proofreading of Peptide-MHC Complexes through Dynamic Multivalent Interactions.* Front Immunol, 2017. **8**: p. 65.
165. Spiliotis, E.T., et al., *Selective export of MHC class I molecules from the ER after their dissociation from TAP.* Immunity, 2000. **13**(6): p. 841-51.
166. Ladasky, J.J., et al., *Bap31 enhances the endoplasmic reticulum export and quality control of human class I MHC molecules.* J Immunol, 2006. **177**(9): p. 6172-81.
167. Steinman, R.M., et al., *Endocytosis and the recycling of plasma membrane.* J Cell Biol, 1983. **96**(1): p. 1-27.
168. Ljunggren, H.G., et al., *Empty MHC class I molecules come out in the cold.* Nature, 1990. **346**(6283): p. 476-80.
169. Neefjes, J.J., et al., *The fate of the three subunits of major histocompatibility complex class I molecules.* Eur J Immunol, 1992. **22**(6): p. 1609-14.
170. Mahmutefendic, H., et al., *Endosomal trafficking of open Major Histocompatibility Class I conformers--implications for presentation of endocytosed antigens.* Mol Immunol, 2013. **55**(2): p. 149-52.
171. Bartee, E., et al., *Downregulation of major histocompatibility complex class I by human ubiquitin ligases related to viral immune evasion proteins.* J Virol, 2004. **78**(3): p. 1109-20.

172. Duncan, L.M., et al., *Lysine-63-linked ubiquitination is required for endolysosomal degradation of class I molecules.* EMBO J, 2006. **25**(8): p. 1635-45.
173. Gromme, M., et al., *Recycling MHC class I molecules and endosomal peptide loading.* Proc Natl Acad Sci U S A, 1999. **96**(18): p. 10326-31.
174. Caplan, S., et al., *A tubular EHD1-containing compartment involved in the recycling of major histocompatibility complex class I molecules to the plasma membrane.* EMBO J, 2002. **21**(11): p. 2557-67.
175. Weigert, R., et al., *Rab22a regulates the recycling of membrane proteins internalized independently of clathrin.* Mol Biol Cell, 2004. **15**(8): p. 3758-70.
176. Xu, R.H., et al., *Direct presentation is sufficient for an efficient anti-viral CD8+ T cell response.* PLoS Pathog, 2010. **6**(2): p. e1000768.
177. Rock, K.L., S. Gamble, and L. Rothstein, *Presentation of exogenous antigen with class I major histocompatibility complex molecules.* Science, 1990. **249**(4971): p. 918-21.
178. Kovacsovics-Bankowski, M., et al., *Efficient major histocompatibility complex class I presentation of exogenous antigen upon phagocytosis by macrophages.* Proc Natl Acad Sci U S A, 1993. **90**(11): p. 4942-6.
179. Mora, J.R. and U.H. von Andrian, *T-cell homing specificity and plasticity: new concepts and future challenges.* Trends Immunol, 2006. **27**(5): p. 235-43.
180. Sei, J.J., et al., *Peptide-MHC-I from Endogenous Antigen Outnumber Those from Exogenous Antigen, Irrespective of APC Phenotype or Activation.* PLoS Pathog, 2015. **11**(6): p. e1004941.
181. Heipertz, E.L., et al., *Prolonged antigen presentation following an acute virus infection requires direct and then cross-presentation.* J Immunol, 2014. **193**(8): p. 4169-77.
182. Hickman, H.D., et al., *Direct priming of antiviral CD8+ T cells in the peripheral interfollicular region of lymph nodes.* Nat Immunol, 2008. **9**(2): p. 155-65.
183. Pfeifer, J.D., et al., *Phagocytic processing of bacterial antigens for class I MHC presentation to T cells.* Nature, 1993. **361**(6410): p. 359-62.

184. Harriff, M.J., et al., *TAP mediates import of Mycobacterium tuberculosis-derived peptides into phagosomes and facilitates loading onto HLA-I.* PLoS One, 2013. **8**(11): p. e79571.
185. Bertholet, S., et al., *Leishmania antigens are presented to CD8+ T cells by a transporter associated with antigen processing-independent pathway in vitro and in vivo.* J Immunol, 2006. **177**(6): p. 3525-33.
186. Subramanian, M., et al., *An AXL/LRP-1/RANBP9 complex mediates DC efferocytosis and antigen cross-presentation in vivo.* J Clin Invest, 2014. **124**(3): p. 1296-308.
187. Matheoud, D., et al., *Cross-presentation by dendritic cells from live cells induces protective immune responses in vivo.* Blood, 2010. **115**(22): p. 4412-20.
188. Matheoud, D., et al., *Dendritic cells crosspresent antigens from live B16 cells more efficiently than from apoptotic cells and protect from melanoma in a therapeutic model.* PLoS One, 2011. **6**(4): p. e19104.
189. Neijssen, J., et al., *Cross-presentation by intercellular peptide transfer through gap junctions.* Nature, 2005. **434**(7029): p. 83-8.
190. Wolfers, J., et al., *Tumor-derived exosomes are a source of shared tumor rejection antigens for CTL cross-priming.* Nat Med, 2001. **7**(3): p. 297-303.
191. Norbury, C.C., et al., *Class I MHC presentation of exogenous soluble antigen via macropinocytosis in bone marrow macrophages.* Immunity, 1995. **3**(6): p. 783-91.
192. Falo, L.D., Jr., et al., *Targeting antigen into the phagocytic pathway in vivo induces protective tumour immunity.* Nat Med, 1995. **1**(7): p. 649-53.
193. Cruz, F.M., et al., *The Biology and Underlying Mechanisms of Cross-Presentation of Exogenous Antigens on MHC-I Molecules.* Annu Rev Immunol, 2017. **35**: p. 149-176.
194. Wakim, L.M. and M.J. Bevan, *Cross-dressed dendritic cells drive memory CD8+ T-cell activation after viral infection.* Nature, 2011. **471**(7340): p. 629-32.
195. Harshyne, L.A., et al., *A role for class A scavenger receptor in dendritic cell nibbling from live cells.* J Immunol, 2003. **170**(5): p. 2302-9.

196. Laumont, C.M. and C. Perreault, *Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy.* Cell Mol Life Sci, 2018. **75**(4): p. 607-621.
197. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.
198. Wahl, A., et al., *HLA class I molecules reflect an altered host proteome after influenza virus infection.* Hum Immunol, 2010. **71**(1): p. 14-22.
199. Bassani-Sternberg, M., et al., *Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.* Nat Commun, 2016. **7**: p. 13404.
200. Esquivel, F., J. Yewdell, and J. Bennink, *RMA/S cells present endogenously synthesized cytosolic proteins to class I-restricted cytotoxic T lymphocytes.* J Exp Med, 1992. **175**(1): p. 163-8.
201. Croft, N.P., et al., *Kinetics of antigen expression and epitope presentation during virus infection.* PLoS Pathog, 2013. **9**(1): p. e1003129.
202. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
203. Schubert, U., et al., *Rapid degradation of a large fraction of newly synthesized proteins by proteasomes.* Nature, 2000. **404**(6779): p. 770-4.
204. Qian, S.B., et al., *Characterization of rapidly degraded polypeptides in mammalian cells reveals a novel layer of nascent protein quality control.* J Biol Chem, 2006. **281**(1): p. 392-400.
205. Pearson, H., et al., *MHC class I-associated peptides derive from selective regions of the human genome.* J Clin Invest, 2016. **126**(12): p. 4690-4701.
206. Anton, L.C. and J.W. Yewdell, *Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors.* J Leukoc Biol, 2014. **95**(4): p. 551-62.
207. Jonas, S. and E. Izaurralde, *Towards a molecular understanding of microRNA-mediated gene silencing.* Nat Rev Genet, 2015. **16**(7): p. 421-33.

208. Apcher, S., et al., *Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation.* Proc Natl Acad Sci U S A, 2011. **108**(28): p. 11572-7.
209. Eisenlohr, L.C., L. Huang, and T.N. Golovina, *Rethinking peptide supply to MHC class I molecules.* Nat Rev Immunol, 2007. **7**(5): p. 403-10.
210. Zarling, A.L., et al., *Phosphorylated peptides are naturally processed and presented by major histocompatibility complex class I molecules in vivo.* J Exp Med, 2000. **192**(12): p. 1755-62.
211. Cobbold, M., et al., *MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia.* Sci Transl Med, 2013. **5**(203): p. 203ra125.
212. Yague, J., J. Vazquez, and J.A. Lopez de Castro, *A post-translational modification of nuclear proteins, N(G),N(G)-dimethyl-Arg, found in a natural HLA class I peptide ligand.* Protein Sci, 2000. **9**(11): p. 2210-7.
213. Haurum, J.S., et al., *Presentation of cytosolic glycosylated peptides by human class I major histocompatibility complex molecules in vivo.* J Exp Med, 1999. **190**(1): p. 145-50.
214. Engelhard, V.H., et al., *Post-translational modifications of naturally processed MHC-binding epitopes.* Curr Opin Immunol, 2006. **18**(1): p. 92-7.
215. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
216. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.
217. Ingolia, N.T., et al., *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.
218. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.
219. Kim, M.S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575-81.
220. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.

221. Vanderperre, B., et al., *Direct detection of alternative open reading frames translation products in human significantly expands the proteome.* PLoS One, 2013. **8**(8): p. e70698.
222. Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**(3): p. 193-204.
223. Starck, S.R., et al., *Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I.* Science, 2012. **336**(6089): p. 1719-23.
224. Fields, A.P., et al., *A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation.* Mol Cell, 2015. **60**(5): p. 816-27.
225. Slavoff, S.A., et al., *A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining.* J Biol Chem, 2014. **289**(16): p. 10950-7.
226. D'Lima, N.G., et al., *A human microprotein that interacts with the mRNA decapping complex.* Nat Chem Biol, 2017. **13**(2): p. 174-180.
227. Lubec, G. and L. Afjehi-Sadat, *Limitations and pitfalls in protein identification by mass spectrometry.* Chem Rev, 2007. **107**(8): p. 3568-84.
228. Samandi, S., et al., *Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins.* Elife, 2017. **6**.
229. Boon, T., et al., *Cloning and characterization of genes coding for tum- transplantation antigens.* J Autoimmun, 1989. **2 Suppl**: p. 109-14.
230. Boon, T. and A. Van Pel, *T cell-recognized antigenic peptides derived from the cellular genome are not protein degradation products but can be generated directly by transcription and translation of short subgenic regions. A hypothesis.* Immunogenetics, 1989. **29**(2): p. 75-9.
231. Shastri, N. and F. Gonzalez, *Endogenous generation and presentation of the ovalbumin peptide/Kb complex to T cells.* J Immunol, 1993. **150**(7): p. 2724-36.
232. Shastri, N., V. Nguyen, and F. Gonzalez, *Major histocompatibility class I molecules can present cryptic translation products to T-cells.* J Biol Chem, 1995. **270**(3): p. 1088-91.

233. Schwab, S.R., et al., *Constitutive display of cryptic translation products by MHC class I molecules.* Science, 2003. **301**(5638): p. 1367-71.
234. Uenaka, A., et al., *Identification of a unique antigen peptide pRL1 on BALB/c RL male 1 leukemia recognized by cytotoxic T lymphocytes and its relation to the Akt oncogene.* J Exp Med, 1994. **180**(5): p. 1599-607.
235. Goodenough, E., et al., *Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR.* Proc Natl Acad Sci U S A, 2014. **111**(15): p. 5670-5.
236. Coulie, P.G., et al., *A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma.* Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7976-80.
237. Guilloux, Y., et al., *A peptide recognized by human cytolytic T lymphocytes on HLA-A2 melanomas is encoded by an intron sequence of the N-acetylglucosaminyltransferase V gene.* J Exp Med, 1996. **183**(3): p. 1173-83.
238. Robbins, P.F., et al., *The intronic region of an incompletely spliced gp100 gene transcript encodes an epitope recognized by melanoma-reactive tumor-infiltrating lymphocytes.* J Immunol, 1997. **159**(1): p. 303-8.
239. Wang, R.F., et al., *Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen.* J Exp Med, 1996. **183**(3): p. 1131-40.
240. Rosenberg, S.A., et al., *Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy.* J Immunol, 2002. **168**(5): p. 2402-7.
241. Van Den Eynde, B.J., et al., *A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription.* J Exp Med, 1999. **190**(12): p. 1793-800.
242. van Bergen, C.A., et al., *Selective graft-versus-leukemia depends on magnitude and diversity of the alloreactive T cell response.* J Clin Invest, 2017. **127**(2): p. 517-529.

243. Kracht, M.J., et al., *Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes.* Nat Med, 2017. **23**(4): p. 501-507.
244. Coulie, P.G., et al., *Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy.* Nat Rev Cancer, 2014. **14**(2): p. 135-46.
245. Vigneron, N., et al., *Database of T cell-defined human tumor antigens: the 2013 update.* Cancer Immun, 2013. **13**: p. 15.
246. Coulie, P.G., et al., *A new gene coding for a differentiation antigen recognized by autologous cytolytic T lymphocytes on HLA-A2 melanomas.* J Exp Med, 1994. **180**(1): p. 35-42.
247. Kawakami, Y., et al., *Identification of the immunodominant peptides of the MART-1 human melanoma antigen recognized by the majority of HLA-A2-restricted tumor infiltrating lymphocytes.* J Exp Med, 1994. **180**(1): p. 347-52.
248. Bakker, A.B., et al., *Identification of a novel peptide derived from the melanocyte-specific gp100 antigen as the dominant epitope recognized by an HLA-A2.1-restricted anti-melanoma CTL line.* Int J Cancer, 1995. **62**(1): p. 97-102.
249. Weinzierl, A.O., et al., *Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface.* Mol Cell Proteomics, 2007. **6**(1): p. 102-13.
250. Fortier, M.H., et al., *The MHC class I peptide repertoire is molded by the transcriptome.* J Exp Med, 2008. **205**(3): p. 595-610.
251. Fisk, B., et al., *Identification of an immunodominant peptide of HER-2/neu protooncogene recognized by ovarian tumor-specific cytotoxic T lymphocyte lines.* J Exp Med, 1995. **181**(6): p. 2109-17.
252. Inoue, K., et al., *Aberrant overexpression of the Wilms tumor gene (WT1) in human leukemia.* Blood, 1997. **89**(4): p. 1405-12.
253. Andersen, R.S., et al., *Dissection of T-cell antigen specificity in human melanoma.* Cancer Res, 2012. **72**(7): p. 1642-50.
254. Brown, S.D., et al., *Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival.* Genome Res, 2014. **24**(5): p. 743-50.

255. Charoentong, P., et al., *Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade.* Cell Rep, 2017. **18**(1): p. 248-262.
256. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade.* Science, 2016.
257. Yadav, M., et al., *Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.* Nature, 2014. **515**(7528): p. 572-6.
258. Kreiter, S., et al., *Mutant MHC class II epitopes drive therapeutic immune responses to cancer.* Nature, 2015. **520**(7549): p. 692-6.
259. Stevanovic, S., et al., *Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer.* Science, 2017. **356**(6334): p. 200-205.
260. Turajlic, S., et al., *Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis.* Lancet Oncol, 2017. **18**(8): p. 1009-1021.
261. Mertens, F., et al., *The emerging complexity of gene fusions in cancer.* Nat Rev Cancer, 2015. **15**(6): p. 371-81.
262. Hayward, N.K., et al., *Whole-genome landscapes of major melanoma subtypes.* Nature, 2017. **545**(7653): p. 175-180.
263. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.
264. Martin, S.D., et al., *Low Mutation Burden in Ovarian Cancer May Limit the Utility of Neoantigen-Targeted Vaccines.* PLoS One, 2016. **11**(5): p. e0155189.
265. Tran, E., et al., *Immunogenicity of somatic mutations in human gastrointestinal cancers.* Science, 2015. **350**(6266): p. 1387-90.
266. Nielsen, J.S., et al., *Toward Personalized Lymphoma Immunotherapy: Identification of Common Driver Mutations Recognized by Patient CD8+ T Cells.* Clin Cancer Res, 2016. **22**(9): p. 2226-36.
267. Tran, E., et al., *T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer.* N Engl J Med, 2016. **375**(23): p. 2255-2262.

268. Rooney, M.S., et al., *Molecular and genetic properties of tumors associated with local immune cytolytic activity.* Cell, 2015. **160**(1-2): p. 48-61.
269. Hunt, D.F., et al., *Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.* Science, 1992. **255**(5049): p. 1261-3.
270. Purbhoo, M.A., et al., *T cell killing does not require the formation of a stable mature immunological synapse.* Nat Immunol, 2004. **5**(5): p. 524-30.
271. Vincent, K., et al., *Rejection of leukemic cells requires antigen-specific T cells with high functional avidity.* Biol Blood Marrow Transplant, 2014. **20**(1): p. 37-45.
272. Karttunen, J., S. Sanderson, and N. Shastri, *Detection of rare antigen-presenting cells by the lacZ T-cell activation assay suggests an expression cloning strategy for T-cell antigens.* Proc Natl Acad Sci U S A, 1992. **89**(13): p. 6020-4.
273. Porgador, A., et al., *Localization, quantitation, and in situ detection of specific peptide-MHC class I complexes using a monoclonal antibody.* Immunity, 1997. **6**(6): p. 715-26.
274. Moore, M.W., F.R. Carbone, and M.J. Bevan, *Introduction of soluble protein into the class I pathway of antigen processing and presentation.* Cell, 1988. **54**(6): p. 777-85.
275. Rotzschke, O., et al., *Exact prediction of a natural T cell epitope.* Eur J Immunol, 1991. **21**(11): p. 2891-4.
276. Apcher, S., et al., *Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17951-6.
277. Deplaen, E., et al., *Immunogenic (Tum-) Variants of Mouse Tumor-P815 - Cloning of the Gene of Tum- Antigen-P91a and Identification of the Tum- Mutation .9.* Proceedings of the National Academy of Sciences of the United States of America, 1988. **85**(7): p. 2274-2278.
278. van der Bruggen, P., et al., *A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma.* Science, 1991. **254**(5038): p. 1643-7.

279. Lurquin, C., et al., *Structure of the gene of tum- transplantation antigen P91A: the mutated exon encodes a peptide recognized with Ld by cytolytic T cells.* Cell, 1989. **58**(2): p. 293-303.
280. Traversari, C., et al., *A nonapeptide encoded by human gene MAGE-1 is recognized on HLA-A1 by cytolytic T lymphocytes directed against tumor antigen MZ2-E.* J Exp Med, 1992. **176**(5): p. 1453-7.
281. Popovic, J., et al., *The only proposed T-cell epitope derived from the TEL-AML1 translocation is not naturally processed.* Blood, 2011. **118**(4): p. 946-54.
282. Rotzschke, O., et al., *Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells.* Nature, 1990. **348**(6298): p. 252-4.
283. Jardetzky, T.S., et al., *Identification of self peptides bound to purified HLA-B27.* Nature, 1991. **353**(6342): p. 326-9.
284. Rotzschke, O., et al., *Characterization of naturally occurring minor histocompatibility peptides including H-4 and H-Y.* Science, 1990. **249**(4966): p. 283-7.
285. Caron, E., et al., *Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry.* Mol Cell Proteomics, 2015. **14**(12): p. 3105-17.
286. Lanoix, J., et al., *Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods.* Proteomics, 2018. **18**(12): p. e1700251.
287. Gebreselassie, D., H. Spiegel, and S. Vukmanovic, *Sampling of major histocompatibility complex class I-associated peptidome suggests relatively looser global association of HLA-B*5101 with peptides.* Hum Immunol, 2006. **67**(11): p. 894-906.
288. Abelin, J.G., et al., *Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry.* Nat Protoc, 2015. **10**(9): p. 1308-18.
289. Gokce, E., et al., *Increasing proteome coverage with offline RP HPLC coupled to online RP nanoLC-MS.* J Chromatogr B Analyt Technol Biomed Life Sci, 2011. **879**(9-10): p. 610-4.

290. Peng, J., et al., *Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome.* J Proteome Res, 2003. **2**(1): p. 43-50.
291. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.
292. Osinalde, N., et al., *Targeted mass spectrometry: An emerging powerful approach to unblock the bottleneck in phosphoproteomics.* J Chromatogr B Analyt Technol Biomed Life Sci, 2017. **1055-1056**: p. 29-38.
293. Bourmaud, A., S. Gallien, and B. Domon, *Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications.* Proteomics, 2016. **16**(15-16): p. 2146-59.
294. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-8.
295. Henderson, C.M., et al., *Skyline Performs as Well as Vendor Software in the Quantitative Analysis of Serum 25-Hydroxy Vitamin D and Vitamin D Binding Globulin.* Clin Chem, 2018. **64**(2): p. 408-410.
296. Gallien, S., S.Y. Kim, and B. Domon, *Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM).* Mol Cell Proteomics, 2015. **14**(6): p. 1630-44.
297. Ronsein, G.E., et al., *Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics.* J Proteomics, 2015. **113**: p. 388-99.
298. Kockmann, T., et al., *Targeted proteomics coming of age - SRM, PRM and DIA performance evaluated from a core facility perspective.* Proteomics, 2016. **16**(15-16): p. 2183-92.
299. Peterson, A.C., et al., *Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics.* Mol Cell Proteomics, 2012. **11**(11): p. 1475-88.
300. Wu, C., et al., *Expediting SRM assay development for large-scale targeted proteomics experiments.* J Proteome Res, 2014. **13**(10): p. 4479-87.

301. Mommen, G.P., et al., *Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD).* Proc Natl Acad Sci U S A, 2014. **111**(12): p. 4507-12.
302. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
303. Zhang, J., et al., *PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.* Mol Cell Proteomics, 2012. **11**(4): p. M111 010587.
304. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for mass spectrometry-based proteomics.* Methods Mol Biol, 2010. **604**: p. 55-71.
305. Gupta, N., et al., *Target-decoy approach and false discovery rate: when things may go wrong.* J Am Soc Mass Spectrom, 2011. **22**(7): p. 1111-20.
306. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.
307. Granados, D.P., et al., *Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides.* Nat Commun, 2014. **5**: p. 3600.
308. Granados, D.P., et al., *Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers.* Leukemia, 2016.
309. Liepe, J., et al., *A large fraction of HLA class I ligands are proteasome-generated spliced peptides.* Science, 2016. **354**(6310): p. 354-358.
310. Vita, R., et al., *The immune epitope database (IEDB) 3.0.* Nucleic Acids Res, 2015. **43**(Database issue): p. D405-12.
311. Shao, W., et al., *The SysteMHC Atlas project.* Nucleic Acids Res, 2018. **46**(D1): p. D1237-D1247.
312. Nielsen, M., et al., *The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage.* Immunogenetics, 2005. **57**(1-2): p. 33-41.
313. Peters, B., et al., *Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors.* J Immunol, 2003. **171**(4): p. 1741-9.

314. Andreatta, M. and M. Nielsen, *Gapped sequence alignment using artificial neural networks: application to the MHC class I system.* Bioinformatics, 2016. **32**(4): p. 511-7.
315. Karosiene, E., et al., *NetMHCcons: a consensus method for the major histocompatibility complex class I predictions.* Immunogenetics, 2012. **64**(3): p. 177-86.
316. Jurtz, V., et al., *NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data.* J Immunol, 2017. **199**(9): p. 3360-3368.
317. Zhang, H., O. Lund, and M. Nielsen, *The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding.* Bioinformatics, 2009. **25**(10): p. 1293-9.
318. Larsen, M.V., et al., *Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction.* BMC Bioinformatics, 2007. **8**: p. 424.
319. Stranzl, T., et al., *NetCTLpan: pan-specific MHC class I pathway epitope predictions.* Immunogenetics, 2010. **62**(6): p. 357-68.
320. Castle, J.C., et al., *Exploiting the mutanome for tumor vaccination.* Cancer Res, 2012. **72**(5): p. 1081-91.
321. Castle, J.C., et al., *Immunomic, genomic and transcriptomic characterization of CT26 colorectal carcinoma.* BMC Genomics, 2014. **15**: p. 190.
322. Robbins, P.F., et al., *Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells.* Nat Med, 2013. **19**(6): p. 747-52.
323. Szolek, A., *HLA Typing from Short-Read Sequencing Data with OptiType.* Methods Mol Biol, 2018. **1802**: p. 215-223.
324. Assarsson, E., et al., *A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection.* J Immunol, 2007. **178**(12): p. 7890-901.

CHAPTER 2

2 Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames

Céline M. Laumont[1,2], Tariq Daouda[1,2,3], Jean-Philippe Laverdure[1], Éric Bonneil[1], Olivier Caron-Lizotte[1], Marie-Pierre Hardy[1], Diana P. Granados[1,2], Chantal Durette[1], Sébastien Lemieux[1,2,*], Pierre Thibault[1,4,*] & Claude Perreault[1,2,5,*]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[2]Department of Medicine, Faculty of Medicine, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[3]Department of Computer Science and Operations Research, Faculty of Arts and Sciences, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[4]Department of Chemistry, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[5]Division of Hematology, Hôpital Maisonneuve-Rosemont, 5415 de l'Assomption Boulevard, Montreal, Quebec, Canada H1T 2M4.

*These authors contributed equally to this work.

Correspondence and requests for materials should be addressed to:
C.P. (email: claude.perreault@umontreal.ca)

Nature Communications, Volume 7, 10.1038/ncomms10238 (January 5, 2016)

2.1 Context

With the realization that (i) the proteome is far more complex than previously anticipated and that (ii) proteogenomics can be used to study the repertoire of polymorphic MAPs, including minor histocompatibility antigens and mTSAs, we wondered if a similar strategy could be developed to study the impact of non-canonical translation on the MAP repertoire.

This article presents the first, and maybe naive, proteogenomic attempt ever made to comprehensively characterize the MAP repertoire of B-lymphoblastoid cell lines (B-LCLs), i.e. including conventional, cryptic and polymorphic MAPs. Thanks to this proteogenomic approach, we were able to show that cryptic proteins are indeed processed by the MHC I antigen presentation pathway and that, on normal cells, the resulting cryptic MAPs represent at least 10% of the MAP repertoire. Our study confirmed the several, yet isolated, reports of MAPs derived from non-coding regions or non-coding transcripts and clearly shows that CD8+ T cells can survey more than just the coding part of our genome, i.e. ~2%.

Having in hand the largest pool of cryptic MAPs ever reported, we were able to gain some important insights with regard to their biogenesis and how it might differ from the one of conventional MAPs. Lastly, we observed that regions coding for cryptic MAPs overlap with significantly more non-synonymous germline polymorphisms than the ones coding for conventional MAPs. From this unexpected observation, we inferred that, in cancer cells, cryptic MAPs could be an unprecedented, yet unexplored, source of mTSAs, thereby leading us to launch the project presented in **Chapter 4**.

2.2 Authors' contributions

Céline M. Laumont: designed the study, analyzed all data, prepared all figures and wrote the first draft of the manuscript.

Tariq Daouda: developed pyGeno and general discussion.

Jean-Philippe Laverdure: performed bioinformatics analyses required for Figures 2.3 a–c and 2.5 a–c.

Éric Bonneil: acquired mass spectrometry data and validated all cryptic peptides (Supplementary Figure 2.2).

Olivier Caron-Lizotte: performed the analysis presented in Supplementary Figure 2.4.

Marie-Pierre Hardy: performed experiments to test the immunogenicity of cryptic MAPs (Figure 2.7) and wrote the associated method section.

Diana P. Granados: prepared samples for mass spectrometry and RNA-sequencing (subject 3 and 4).

Chantal Durette: acquired mass spectrometry data.

Sébastien Lemieux: analyzed data, helped preparing Figure 2.1 and general discussion.

Pierre Thibault: analyzed data and general discussion.

Claude Perreault: designed the study, analyzed data, general discussion and wrote the first draft of the manuscript.

All authors edited and approved the final version of the manuscript.

2.3 Abstract

In view of recent reports documenting pervasive translation outside of canonical protein-coding sequences, we wished to determine the proportion of major histocompatibility complex (MHC) class I-associated peptides (MAPs) derived from non-canonical reading frames. Here we perform proteogenomic analyses of MAPs eluted from human B cells using high-throughput mass spectrometry to probe the six-frame translation of the B-cell transcriptome. We report that ~ 10% of MAPs originate from allegedly noncoding genomic sequences or exonic out-of-frame translation. The biogenesis and properties of these 'cryptic MAPs' differ from those of conventional MAPs. Cryptic MAPs come from very short proteins with atypical C termini, and are coded by transcripts bearing long 3'UTRs enriched in destabilizing elements. Relative to conventional MAPs, cryptic MAPs display different MHC class I-binding preferences and harbour more genomic polymorphisms, some of which are immunogenic. Cryptic MAPs increase the complexity of the MAP repertoire and enhance the scope of CD8 T-cell immunosurveillance.

2.4 Introduction

Breathtaking advances in genomics and proteomics are drastically changing our perspective of cell biology and, in particular, our understanding of protein synthesis and degradation. For instance, next-generation sequencing analyses have shown that three-quarters of the human genome is capable of being transcribed[1].Meanwhile, high-throughput mass spectrometry (MS) studies in normal and infected human cells have resulted in the identification of proteins representing more than 80% of canonical human and viral protein-coding genes[2,3]. Recently, a quantum leap in systems biology was made possible by the emergence of a new field, proteogenomics, that leverages on next-generation sequencing to perform 'genomically informed proteomics'[4]. In conventional shotgun proteomics, peptide sequencing is achieved by matching tandem MS spectra from an experimental sample against a reference protein sequence database (for example, UniProt). As a result, conventional MS sequencing suffers from a major limitation: it can only identify peptides encoded by the canonical reading frame of classic exons. The crux of proteogenomic studies is to perform MS-based peptide sequencing by searching customized databases containing the six-frame translation of genomic or transcriptomic sequences. In this way, proteogenomics studies can identify peptides encoded by all reading frames of any genomic region[5].

Proteogenomics has rapidly revolutionized our vision of the proteome of cells from numerous living organisms, including normal and neoplastic human cells[4,5]. A fundamental issue tackled by proteogenomics is the landscape of genomic regions that are expressed at the protein level. Ribosome-profiling experiments have provided strong evidence for pervasive translation outside of annotated protein-coding genes[6]. However, the definite proof of a genomic locus being protein-coding is the detection of its corresponding protein[7]. Accordingly, one salient concept emerging from proteogenomic analyses is that the proteome is more complex than previously thought. The proteome contains peptides arising from a variety of RNAs that were not supposed to encode proteins (noncoding RNAs) and are therefore not included in annotated protein databases. Many long noncoding RNAs, short open reading frames (ORFs) and pseudogenes, mislabelled as 'noncoding', were ultimately found to code for peptides[2-

[9]. Moreover, numerous peptides originate from non-canonical reading frames with non-AUG start codons[10].

We therefore hypothesised that proteogenomics might allow us to elucidate a fundamental question: the contribution of proteins derived from non-canonical transcripts to the repertoire of major histocompatibility complex (MHC) class I-associated peptides (MAPs). Endogenous MAPs are collectively referred to as the immunopeptidome and represent the essence of self for CD8 T lymphocytes[11,12]. Despite the fundamental importance of the immunopeptidome, its genesis remains ill-defined[13,14]. MAPs derive from proteolytic degradation of proteins found in all cell compartments; however, the immunopeptidome is not a random sample of the proteome: many abundant proteins do not generate MAPs, while some low-abundance proteins generate large amounts of MAPs[13-16]. In a series of seminal studies, Shastri and colleagues made startling observations showing that, similarly to the proteome, the immunopeptidome might be more complex than anticipated. Using an alloreactive T-cell clone as a probe, they screened a splenic cDNA library in transfected antigen-presenting cells (APCs) and isolated a cDNA clone that encoded the MAP recognized by the T-cell clone. The salient finding was that this MAP derived from a non-canonical reading frame initiated with a non-AUG start codon[17]. They discovered that synthesis of this peptide was initiated with a CUG codon decoded as a leucine rather than a methionine[18]. Studies by other groups provided evidence that MAPs could arise not only from alternate translational reading frames but also from untranslated regions (UTRs or introns)[19,20]. However, the structure of only a handful of these 'cryptic MAPs' has been confirmed by MS[20,21]. Therefore, in the absence of proteomic evidence, the existence of most reported cryptic MAPs must be considered with some scepticism because their identification relied on indirect methods fraught with high false discovery rates (FDRs). We therefore developed a novel proteogenomic approach to define the landscape of the cryptic immunopeptidome and answer the following questions: what proportion of MAPs derives from non-canonical reading frames and how are they generated? To this end, we performed an all-frames translation of the transcriptome of human B lymphoblastoid cell lines to generate databases of predicted

peptides/proteins. These databases were used to identify MAPs using high-throughput MS sequencing. Integration of transcriptomic and proteomic data revealed that cryptic MAPs constitute ~ 10% of the immunopeptidome and that their biogenesis and properties differ in many ways from those of conventional MAPs.

2.5 Results

2.5.1 Novel proteogenomic strategy to identify cryptic MAPs

MAPs were eluted from an Epstein-Barr virus-transformed B-cell line (B-LCL) obtained from a blood donor bearing the HLA-A*03:01, -A*29:02; -B*08:01, -B*44:03 MHC class I molecules (referred to as subject 1). Peptides were fractionated with strong cation exchange chromatography and analysed with liquid chromatography-MS/MS using high-resolution precursor and product ion spectra, as previously described[22]. To identify both conventional and cryptic MAPs present at the surface of this B-LCL, peptides were matched to two personalized databases referred to as the 'control' and the 'all-frames' databases (**Figure 2.1a**). Both databases were built by *in silico* translation of RNA-sequencing (RNA-seq) data from subject 1's B-LCL using the pyGeno python package (https://github.com/tariqdaouda/pyGeno)[23]. Two reasons led us to focus on the transcriptome rather than the genome of our B-LCL for database construction: (i) MAPs can only derive from transcripts expressed in the cell of interest and (ii) in proteogenomics, the risk of false discovery increases with the size of the database used for MS sequencing[4,5].

The control database corresponds to the canonical proteome of the B-LCL and was generated as follows (**Figure 2.1a, left**): RNA-seq reads were mapped on the reference genome (version GRCh37.75) to identify subject 1-specific high-quality non-synonymous single-nucleotide polymorphisms (ns-SNPs), which were then integrated in the reference genome to build the personalized genome of subject 1. All putative protein-coding genes were then *in silico* translated in their conventional reading frame to obtain the canonical proteome of the B-LCL. The all-frames database was built using the six-frame translation of RNA-seq data from the B-LCL (**Figure 2.1a, right**): reads passing the Illumina quality filters were in silico translated into six possible reading frames using a sliding window of 33 base pairs (bp), since the vast majority of MAPs are known to be 8–11 amino acids long and only rare MAPs contain more than 11 residues. Translation products having a length inferior to eight amino acids, due to the presence of a stop codon within the sliding window, were excluded. By not aligning the

reads before translating them, we are able to leverage the whole output of the sequencer, including reads resulting from rare elongation events that might otherwise be discarded. However, this approach also prevents us from using established filtering approaches such as coverage measures or base quality filters. To address the necessity of sequence filtering, we computed for each translated peptide an S-value (or Seen-value) that represents the number of times a peptide was seen following the *in silico* translation (**Supplementary Figure 2.1a**). The higher the S-value, the more confidence we have that the peptide sequence is indeed not due to a sequencing error. We therefore elected to use a stringent approach and kept only peptides having an S-value ≥ 10 to (i) obtain a database whose size was manageable using the Mascot search engine (**Supplementary Figure 2.1b**) and (ii) to minimize the risk of false discovery[4,5,24].

In our search for cryptic MAPs, the key question was whether the all-frames database would lead to the identification of MAPs missed with the control database, which only contains the *in silico* translation of sequences assumed to be translated (for example, protein-coding transcripts). Out of 3,037 MAPs identified by the all-frames database, 2,686 MAPs were also identified by the control database among which 2,435 were unambiguously assigned to a single gene (**Figure 2.1b**). However, the salient finding is that 351 MAPs were solely identified by the all-frames database. After these 351 putative cryptic MAPs were subjected to four stringent filtering and validation steps (see Methods), we found that 168 of them were unambiguously assigned to a single genomic region (**Figure 2.1b**). We validated 18 cryptic MAPs using the synthetic version of them (**Supplementary Figure 2.2**). Furthermore, we found that the Mascot score distribution (the confidence level of a peptide assignation using MS) and the transcriptomic coverage of the peptide-coding regions (PCRs) were similar for these 168 cryptic and the 2,435 conventional MAPs (**Supplementary Figure 2.3**). It should be noted that the multiple filtering steps were designed to be particularly stringent. We therefore expect that some of the 183 discarded peptides may, nevertheless, be genuine cryptic MAPs (**Figure 2.1b**), thereby increasing their total number up to 351 (13% of the immunopeptidome). However, at this discovery stage, we chose to conduct

further analyses using only the 168 cryptic MAPs identified using our most stringent criteria (6.5% of the immunopeptidome).

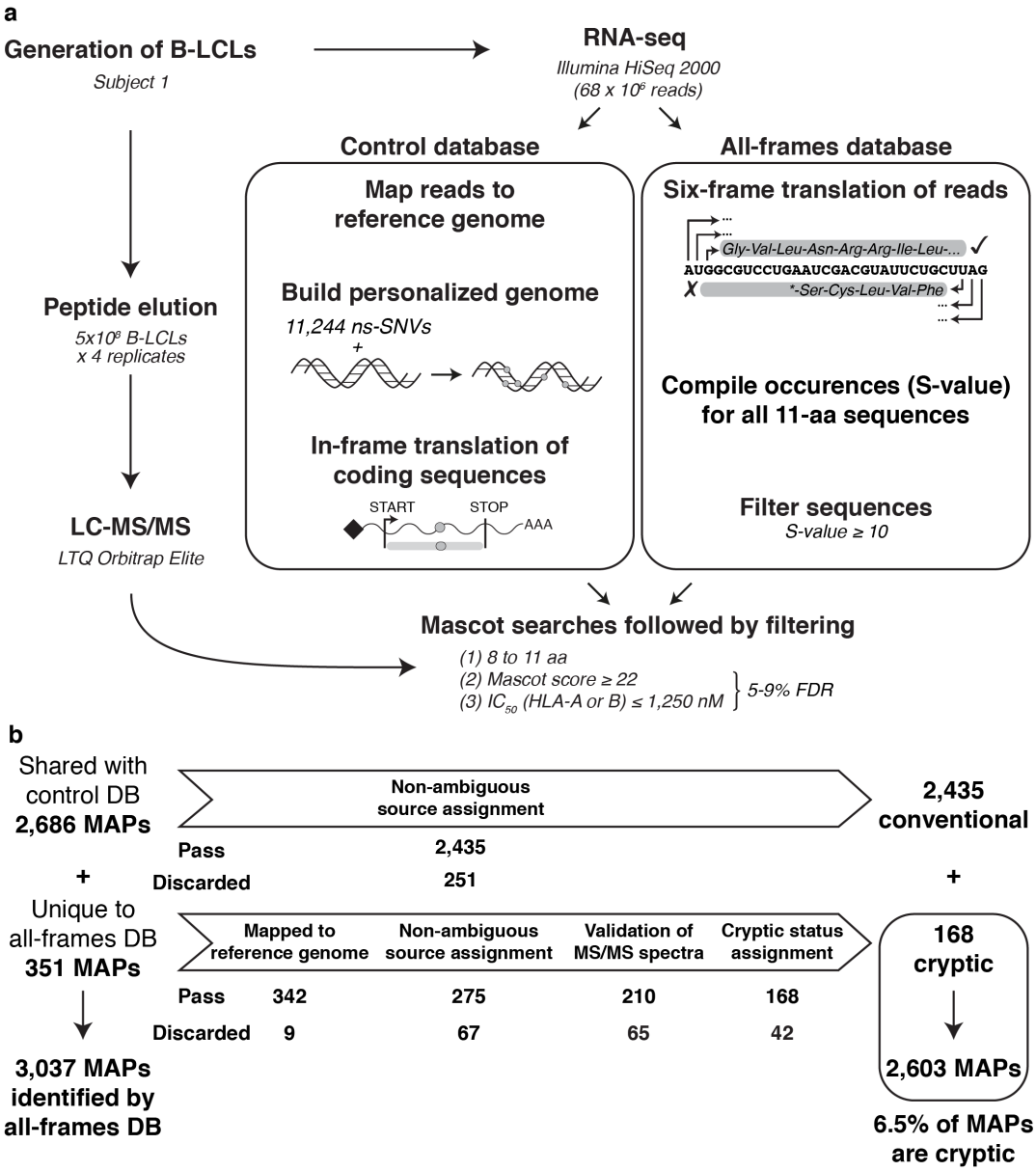


Figure 2.1 | Proteogenomic workflow used for high-throughput identification of cryptic MAPs. (**a**) General overview of the proteogenomic workflow used to identify conventional (Conv.) and cryptic (Crypt.) MAPs. Peptides were eluted from the cell surface of subject 1's B-LCL and were sequenced with liquid chromatography-MS/MS (LC-MS/MS). To determine the amino-acid (aa) sequence of those peptides, we built two databases (DBs), both derived from the analysis of RNA-seq data obtained from subject 1's B-LCL: the control DB and the all-frames DB (see **Methods** and **Supplementary Figure 2.1**). (**b**) Peptides solely identified by

the all-frames DB were considered as Crypt. MAP candidates and further filtered to remove ambiguous and false-positive identifications. See also **Supplementary Figures 2.2** and **2.3**.

2.5.2 The cryptic MAPs' repertoire is linked to the HLA genotype

Various human leukocyte antigen (HLA) allotypes have different peptide-binding motifs and therefore present different MAP repertoires. Accordingly, if a peptide eluted from cells of subject 1 is a genuine MAP, its presence on cells from other individuals should depend on the presence of the HLA allotype, presenting this peptide on cells from subject 1. In other words, the presence of authentic MAPs should be 'HLA-restricted'. The restriction should be strong but it does not need to be perfect because there are some overlap in the MAP repertoires presented by various allotypes[25]. On the contrary, no HLA restriction should be seen between the HLA genotype and the presence of MHC-unrelated peptides. Therefore, to test whether our cryptic MAPs were HLA-restricted, we analysed the immunopeptidome of three other subjects who shared four, two or no HLA allotypes with subject 1 (**Supplementary Figure 2.4**). For both conventional and cryptic MAPs, we found a very strong positive dependence between peptide detection in subjects 2–4 and the presence of the corresponding HLA-A or -B allotype (two-sided Fisher's exact test, $P < 2.2 \times 10^{-16}$; **Figure 2.2a,b**). The degree of HLA allotype restriction was similar for conventional and cryptic MAPs. Moreover, most of the MAPs detected in the absence of the relevant HLA allele were predicted to be promiscuous binders (**Figure 2.2c**). These data further validate that cryptic peptides detected with our proteogenomic approach are genuine MAPs.

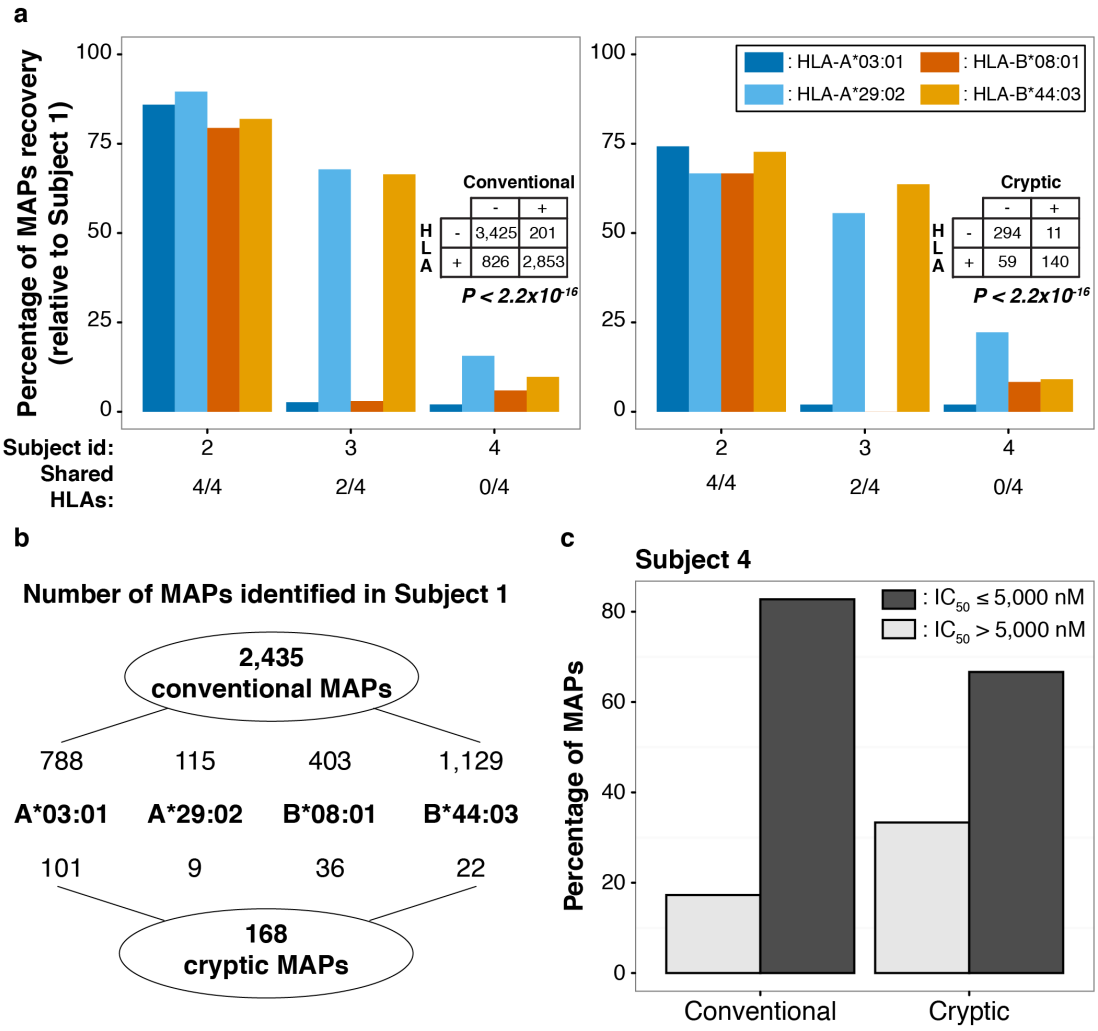


Figure 2.2 | Detection of Crypt. and Conv. MAPs is HLA-dependent. (**a**) Relationship between MAP detection and HLA genotype**.** We sequenced MAPs on B-LCLs from three subjects who shared four, two or no HLA alleles with subject 1. We then determined the number of Conv. (left) and Crypt. (right) MAPs found in subject 1 that were shared by subjects 2–4. Each bar represents one HLA allotype. A detailed schematic of the analysis can be found in **Supplementary Figure 2.4**. MAP detection in subjects 2–4 correlated with presence of the HLA allotype presenting the MAPs in subject 1: $P < 2.2 \times 10^{-16}$ for Conv. and Crypt. MAPs (two-sided Fisher's exact test). (**b**) Schematic detailing of the numbers of Conv. and Crypt. MAPs identified in subject 1 for the considered HLA alleles. (**c**) Most MAPs detected in subject 4 are promiscuous binders. Overall, 168 Conv. and 9 Crypt. MAPs detected in subject 1 were also detected in subject 4, even though the two subjects did not share any HLA alleles. Using NetMHCcons, we computed the predicted binding affinity (IC_{50}) of those MAPs for the four HLA-A and -B allotypes of subject 4, and we kept the lowest of the four IC_{50} values (corresponding to the highest MHC-binding affinity). The bar chart depicts the percentage of Conv. and Crypt. MAPs having an IC_{50} ≤ or > 5,000 nM. Peptides with an IC_{50} > 5,000 nM for the HLA-A/B

allotypes of subject 4 were assumed to be promiscuous binders, that is, to bind subject 4 allotypes in addition to subject 1 allotypes.

2.5.3 Cryptic MAPs derive from both coding and noncoding RNAs

Next, we analysed the origin of cryptic MAPs. A notable finding was that 20.2% of cryptic MAPs unambiguously allocated to one gene could be assigned exclusively to non-annotated antisense transcripts (transcribed from non-template DNA strand; **Figure 2.3a**). This suggests that, although antisense transcripts are generally assumed to be noncoding[26], their translation can generate substrates for the MHC class I antigen presentation pathway. Next, we focused our efforts on sense cryptic MAPs, as annotations were available for their respective gene source, and made two observations. First, by using the gene biotype nomenclature that classifies genes according to their biological relevance[27], we observed that 86.6% of sense cryptic MAPs derived from protein-coding genes, 9% from genes assumed to be noncoding such as pseudogenes, annotated antisenses, long intergenic noncoding RNAs or processed transcripts and finally 4.5% from unannotated intergenic regions (**Figure 2.3b**). Second, by analysing the location of sense cryptic MAPs within their respective gene source, we observed that 48.5% of them were produced by out-of-frame translation of exonic sequences. The remaining 51.5% originated from translation of allegedly noncoding sequences (**Figure 2.3c**). Among those, cryptic MAPs predominantly derived from the translation of 5'UTRs as opposed to 3'UTRs (24.6% versus 7.5%). This observation is coherent with the reinitiation model for translation initiation, which implies that the probability of translation initiation decreases along the transcript[28]. A small proportion of peptides (5.2%) derived from intronic sequences, a finding consistent with a report showing that a construct coding for the model SIINFEKL peptide, could generate MAPs after insertion into an intronic sequence[29]. Finally, we observed that 9.7% of cryptic MAPs derived from UTR–exon or intron–exon junctions and thus corresponded to translation products of overlapping short ORFs or retained intron transcripts, respectively. Overall, these results highlight the complexity of the

immunopeptidome by showing that the landscape of cryptic MAPs includes both sense as well as antisense coding and noncoding RNAs.

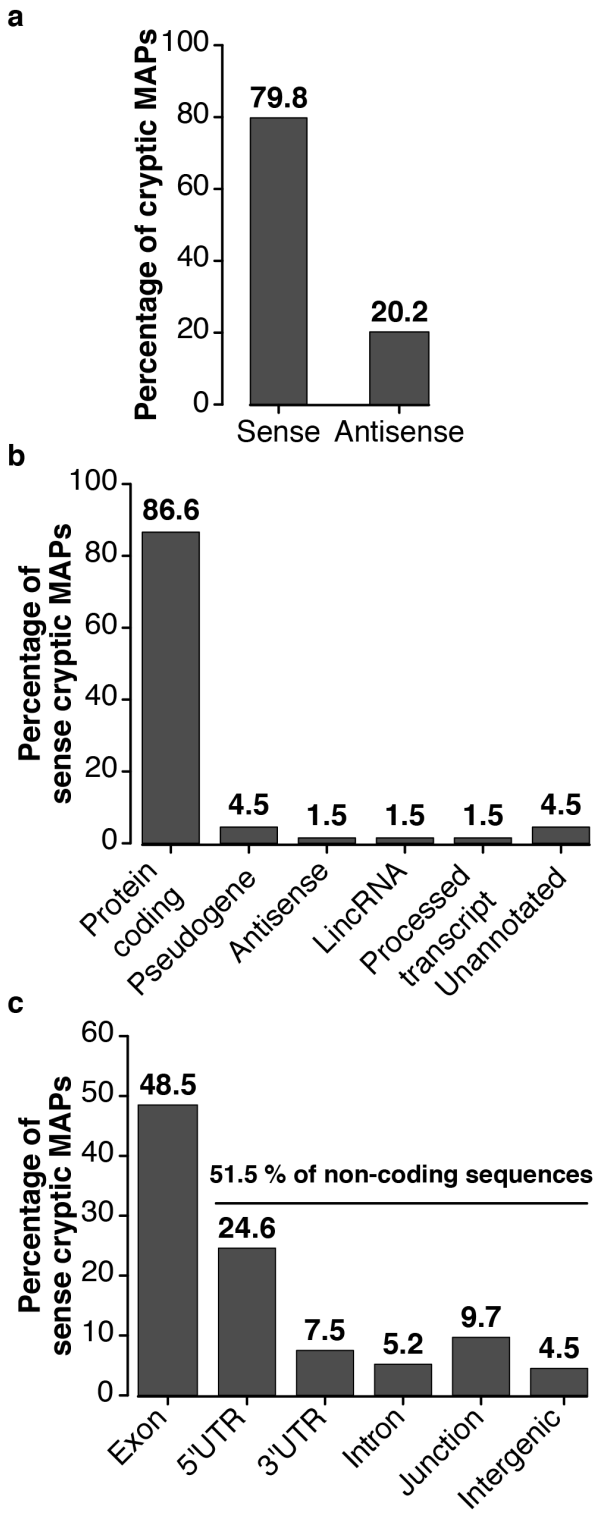


Figure 2.3 | Crypt. MAPs derive from both coding and noncoding transcripts. (**a**) Some Crypt. MAPs derive from novel antisense transcripts. Bar plot showing the percentages of Crypt. MAPs derived from sense and antisense transcriptions. (**b,c**) For Crypt. MAPs derived from sense transcription, we determined the percentage of each gene biotype in MAP source genes (**b**) and the proportion of Crypt. MAPs generated by six types of genomic regions (**c**). The 'exon' class refers to out-of-frame Crypt. MAPs, while the 'junction' category corresponds to peptides encoded by intron–exon or UTR–exon junction. LincRNA, long intergenic noncoding RNAs.

2.5.4 Cryptic MAPs derive from ORFs with a 5' end positional bias

We next sought to determine whether specific types of genes would preferentially generate cryptic as opposed to conventional MAPs. We first noted that very few genes generated both conventional and cryptic MAPs: (i) among the 121 cryptic MAP source genes, only 17 (that is, 14%) also gave rise to conventional MAPs and (ii) only 1% of the 1,731 conventional MAP source genes generated cryptic MAPs (**Figure 2.4a**). The small overlap between genes coding cryptic versus conventional MAP suggests that these two gene sets possess some intrinsic differential feature(s). Further analyses highlighted two conspicuous differences between genes coding conventional versus cryptic MAPs. First, cryptic PCRs were located much closer to the 5' end of their source transcript than conventional PCRs (**Figure 2.4b**). This shift in PCR location was observed not only for cryptic MAPs coded by 5'UTRs and 5'UTR/exons but also for the entire set of exonic cryptic MAPs (**Supplementary Figure 2.5a**). Second, the expression level of genes coding cryptic and conventional MAPs was different. Conventional MAPs have been shown to derive preferentially from abundant transcripts[30,31], and we observed that this was also the case for cryptic MAPs. However, the expression of cryptic MAP-coding genes was slightly but significantly inferior to that of conventional MAP source genes (**Figure 2.4c**).

MAPs derive primarily from rapidly degraded proteins, and evidence suggests that the nonsense-mediated decay (NMD) pathway plays a significant role in this process via translation-dependent degradation[32,33]. NMD targets messenger RNAs (mRNAs) containing a premature termination codon or normal mRNAs containing upstream ORFs[33,34]. Premature termination is predicted to result in more MAPs originating from the 5' end of the transcript[35], as we observed for cryptic but not conventional MAPs (**Figure 2.4b**). In addition, we found that the proportion of MAP-coding transcripts that harboured at least one upstream ORF was significantly higher for cryptic than for conventional MAPs (30% versus 13%), while transcripts generating both types of MAPs showed an intermediate percentage (20%, **Figure 2.4d**). Since transcripts with an upstream ORF generated cryptic MAPs from 5'UTRs but also from exons and 3'UTRs (**Supplementary Figure 2.5b**), NMD appears to be involved in the

generation of all types of cryptic MAPs. Moreover, NMD was also reported to target transcripts bearing long 3'UTRs or 3'UTRs containing intronic sequences. While the transcript source of conventional and cryptic MAPs displayed the same frequency of 3'UTR introns (**Supplementary Figure 2.5c**), cryptic MAP source transcripts had longer 3'UTRs than conventional MAP source transcripts (1,100 versus 687 nt, **Figure 2.4e**). Taken together, these observations suggest that NMD contributes to the generation of cryptic MAPs while lowering the abundance of cryptic MAP source transcripts relative to conventional ones (**Figure 2.4c**) because NMD reduces the steady-state levels of its target RNAs. Besides NMD, mRNA stability is also regulated by cis-regulatory elements that are located in 3'UTRs and interact with RNA-binding proteins[36]. In line with this, relative to conventional MAP source transcripts, the 3'UTRs of cryptic MAP source transcripts contained similar numbers of stabilizing elements but an increased number of destabilizing elements (**Figure 2.4f**). In other words, cryptic MAP source transcripts display longer 3'UTRs with a selective enrichment in destabilizing elements. Taken together, our data suggest that cryptic MAPs derive from unstable transcripts targeted by NMD or 3'UTR-destabilizing elements.

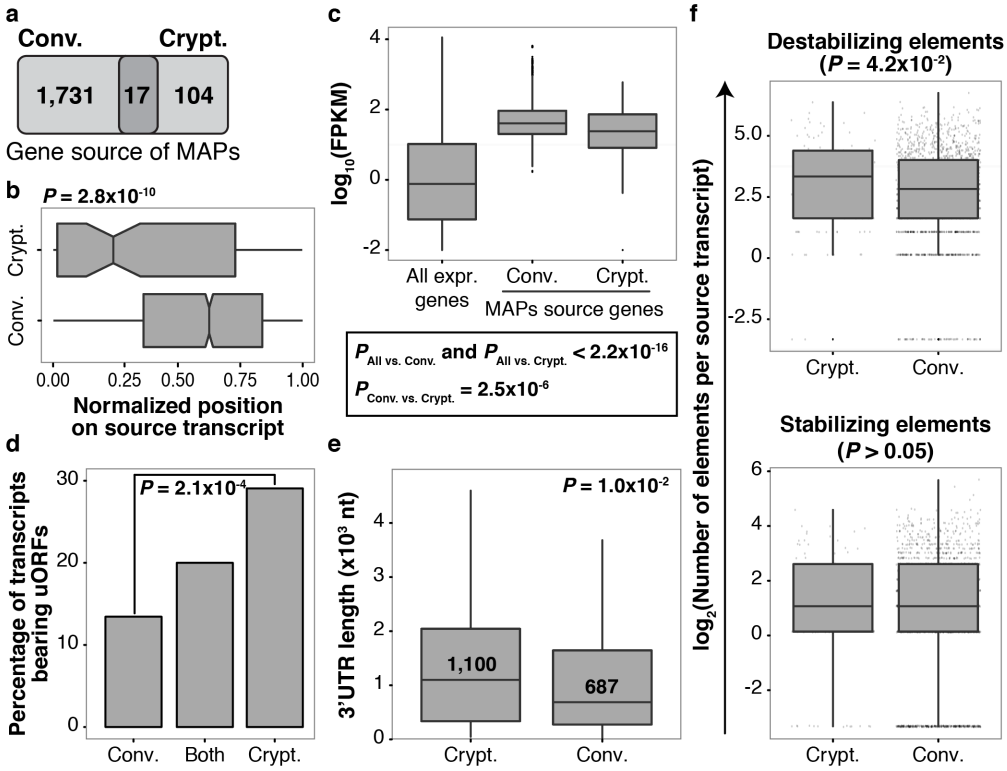


Figure 2.4 | Crypt. MAPs preferentially derive from unstable mRNAs. (**a**) Venn diagram showing minimal overlap between the gene source of Conv. and Crypt. MAPs. (**b**) Crypt. MAPs preferentially derive from the 5' end of their source transcript. The length of each source transcript was normalized to 1, and the start of each MAP was then positioned on a 0–1 scale (x axis), where 0 represents the 5' end of the source transcript. Crypt. MAPs deriving from intergenic and intronic regions were excluded from this analysis. See also **Supplementary Figure 2.5a**. (**c**) Log_{10} expression values, in FPKM, of all genes expressed in B-LCL versus the subset of the gene source of Conv. and Crypt. MAPs. (**d**) Crypt. source transcripts preferentially bear upstream ORFs (uORFs). For each MAP source transcript, we predicted the 5'UTR and 5'UTR–exon ORF initiating at an AUG embedded in an optimal or strong Kozak context. The bar graph shows the proportion of source transcripts bearing at least one uORF and generating a Conv. MAP, a Crypt. MAP or both. See also **Supplementary Figure 2.5b**. (**e**) Crypt. source transcripts display long 3'UTRs. Using pyGeno, we retrieved the 3'UTR of MAP source transcripts (when available) and computed their length in nucleotide (nt). The boxplot displays the resulting 3'UTR length distribution for Crypt. and Conv. MAP source transcripts excluding the upper outliers that represented 6 and 107 values out of 97 and 1,770 transcripts, respectively. (**f**) 3'UTRs of Crypt. but not Conv. MAP source transcripts are enriched in destabilizing elements. We looked for destabilizing and stabilizing elements identified in ref. 36 in the 3'UTR of Crypt. and Conv. MAP source transcripts. For each source transcript, we computed the number of destabilizing and stabilizing elements contained in its sequence. The resulting distributions are plotted for Crypt. and Conv. MAP source transcripts as the log_2 number of destabilizing (top panel) or stabilizing elements (bottom panel) per transcript. See

also **Supplementary Figure 2.5c**. Statistical significance was assessed with a two-sided (**b,c,e**) or one-sided (**f**) Wilcoxon rank sum test, or a two-sided Fisher's exact test (**d**). On box plots, boxes represent second and third quartiles, whiskers $\pm$1.5 the interquartile range, and dots the outliers.

2.5.5 Cryptic MAPs derive from precursors with atypical C termini

To gain further insights into the mechanisms responsible for the generation of cryptic MAPs, we analysed the nucleotide sequence of MAP source transcripts to predict their translation start and stop sites. Notably, we observed that translation initiation occurred at a known initiation codon for 69% of cryptic MAPs: AUG was used more often than near-cognate start codons, which differ from AUG by a single nucleotide (62% versus 7%). This suggests that, even for those atypical proteins, AUG is the preferential translation initiation codon (**Figure 2.5a**). Among near-cognate start codons, CUG was the most commonly observed (**Figure 2.5b**). This observation is in agreement with several reports demonstrating that CUG is the most efficient near-cognate start codon to initiate translation[18,37,38]. Other near-cognate start codons that were used more than one time included ACG and GUG, which were both shown to be enriched at translation initiation sites by ribosome profiling[38]. Finally, 31% of cryptic MAPs did not display any of the known translation initiation codons upstream of their respective PCR (**Figure 2.5a**). In accordance with similar observations based on analyses of ribosome-profiling data[38], these data suggest that translation can be initiated at other codons than the classical AUG or near-cognate start codons.

The median length of conventional proteins is ~ 400 amino acids and, simply by virtue of their size, longer proteins generate more MAPs than shorter proteins[14]. Accordingly, the median length of conventional MAP source proteins in our data set was 523 amino acids. In stark contrast, the median length of cryptic MAP source proteins was 39 amino acids, and 75% of them had less than 62 amino acids (**Figure 2.5c**). The shortest predicted cryptic proteins (3 out of 168) had a length of 10 amino acids and generated cryptic MAPs of 9 amino acids; MHC processing of these cryptic MAPs only required trimming of the N-terminal methionine. The generation of conventional MAPs is initiated by proteasomal cleavage followed in general by

exopeptidase trimming of the N terminus but not the C terminus[39-41]. Therefore, with few exceptions, the C terminus created by the proteasome remains intact in conventional MAPs[42,43]. Given the remarkably short size of cryptic MAP source proteins, we hypothesized that many cryptic MAPs may not need proteasomal degradation before entering the MHC class I antigen presentation pathway. We reasoned that, if cryptic MAPs were proteasome- independent, their C terminus might be different from that of (proteasome-dependent) conventional MAPs. To test this hypothesis, we analysed amino-acid usage at the four C-terminal amino acids of individual MAPs and the four amino acids downstream of the C terminus (in the source protein) for conventional versus cryptic MAPs. The 20 amino-acid residues were grouped into four categories based on their bulkiness and hydrophobicity[44], and we analysed these data to determine which categories were enriched or depleted at each position for the two types of MAPs. We found that, out of the eight considered positions, five displayed significant differential amino-acid class usage between cryptic and conventional MAPs (**Figure 2.5d**). Together, the facts that cryptic MAPs originate from very short proteins and that amino-acid usage around their C termini is different from that of conventional MAPs suggest that processing of cryptic MAPs may be proteasome-independent.

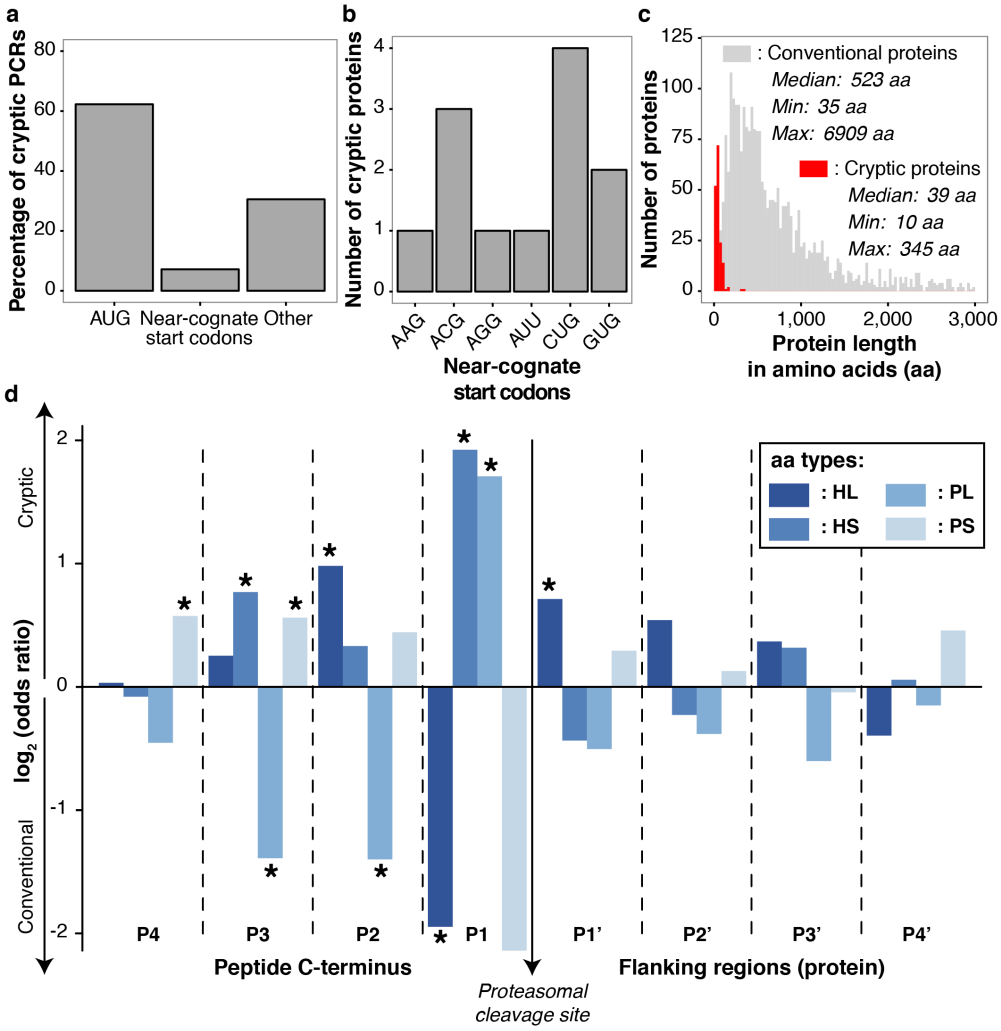


Figure 2.5 | Features of ORFs coding Crypt. MAPs. (**a**) Most Crypt. PCRs are in-frame with an upstream start codon. To predict the probable start codon of each Crypt. PCR, we sequentially applied the following rules: (i) presence of an upstream AUG within an optimal (GCC[R]CCstartG[V]), strong ([R]NNstartG[V]) or weak (anything else) Kozak context, (ii) presence of an upstream near-cognate start codon within an optimal or strong Kozak context, (iii) any other codon downstream of the first upstream stop codon. Bars represent the percentage of Crypt. PCRs displaying an upstream in-frame AUG, near-cognate start codon or any other codon as a probable initiation codon. (**b**) Bar plot showing near-cognate start codon usage at putative translational start sites of 12 Crypt. source proteins. (**c**) Length distribution of Conv. and predicted Crypt. proteins. Median, minimum (Min) and maximum (Max) observed lengths are indicated on the graph for both types of proteins. Conv. proteins having a length 43,000 amino acids are not displayed on the graph. (**d**) Crypt. and Conv. MAPs do not have the same amino-acid composition at their C termini. Amino acids (aa) were classified in four categories: Hydrophobic/Large (HL), Hydrophobic/Small-Medium (HS), Polar/Large (PL) and Polar/Small-Medium (PS)[44]. For the MAP C terminus (positions P4 to P1) and its C-terminal flanking region (positions P1' to P4'), we compared the usage of those four aa categories at

each position between Crypt. and Conv. MAPs. The graph displays the $\log_2$(odds ratio) and significant differences are marked with an asterisk (*P < 0.05; two-sided Fisher's exact test).

2.5.6 Cryptic MAPs display distinct features and are immunogenic

We next evaluated relevant structural and functional features of cryptic MAPs *per se*. Relative to conventional MAPs, we found that cryptic MAPs exhibited three distinctive characteristics: they were shorter, had different allotype-binding preferences and harboured more genomic polymorphisms (**Figure 2.6**). The length distribution of cryptic MAPs revealed a significant enrichment in 8-mers and depletion in 10–11-mers when compared with conventional MAPs (**Figure 2.6a**). This further supports the idea that cryptic and conventional MAPs are processed differently by peptidases. Unexpectedly, we found that cryptic MAPs were preferentially presented by HLA-A*03:01, while conventional MAPs were preferentially presented by HLA-B*44:03 in subject 1 (**Figure 2.6b**). Proteogenomic studies of MAPs presented by other HLA allotypes will be required to assess whether differential allotype preferences of cryptic and conventional MAPs can be generalized. If it were the case, one implication would be that the HLA genotype dictates the breadth of the cryptic immunopeptidome presented at the cell surface. No bias in favour or against ns-SNPs was found in conventional MAP PCRs[22]. However, we found that cryptic MAP PCRs contained a significantly higher frequency of ns-SNPs than conventional MAP PCRs (**Figure 2.6c**; $P < 5.625 \times 10^{-3}$). In other words, cryptic MAPs derive from genomic sequences that are more polymorphic at the population level than conventional protein-coding sequences.

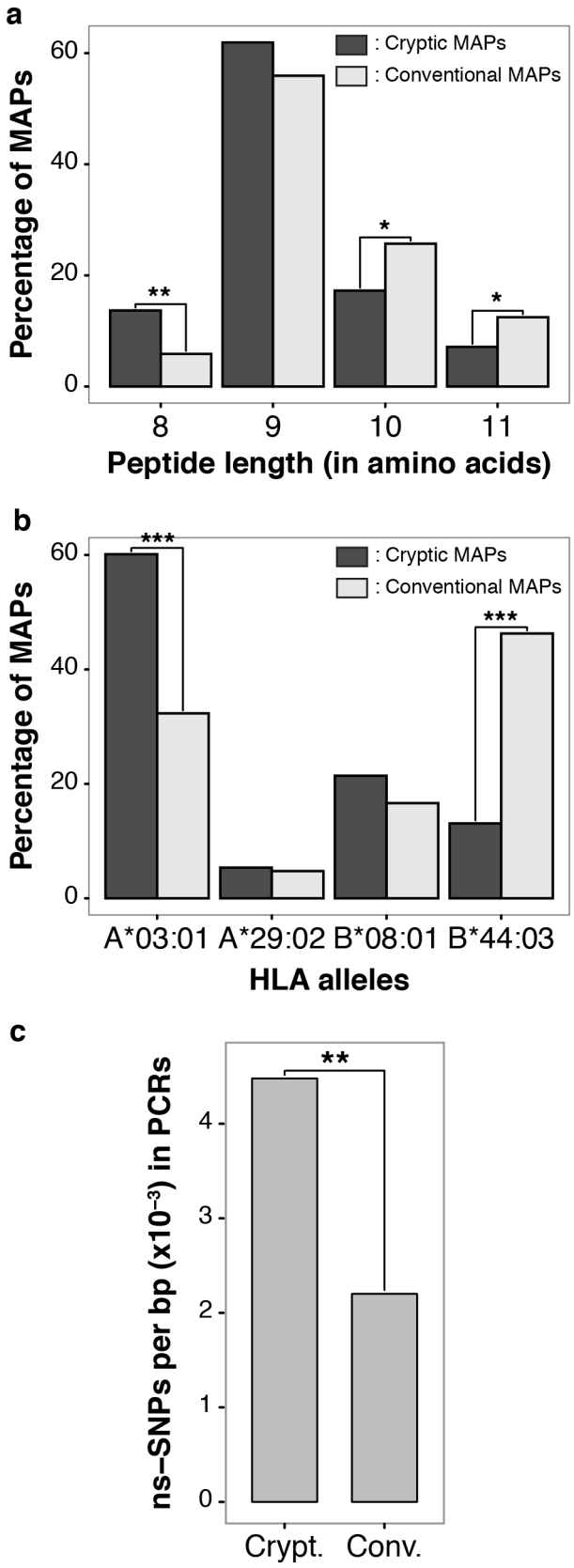


Figure 2.6 | Crypt. and Conv. MAPs display different features. (**a–c**) Bar plots showing that Crypt. and Conv. MAPs from subject 1 have different (a) length distribution, (**b**) allotype distribution and that (**c**) their PCRs exhibit different ns-SNP frequencies (from dbSNP138). In all cases, statistical significance was assessed using a two-sided Fisher's exact test: $*P < 0.05$, $**P \leq 0.006$, $***P \leq 1.1 \times 10^{-11}$ in the bar plots.

Finally, we wished to determine whether cryptic MAPs could be immunogenic. To this end, we studied the T-cell response of subjects 2 and 3 against four randomly selected cryptic MAPs, whose sequence was validated using synthetic peptides (**Supplementary Figure 2.2a–d**), and that were not detected on their own B-LCLs but were present on B-LCLs from subject 1. Two of these MAPs were present on B-LCLs from subject 1 but not subject 2 (HLA-identical to subject 1) because of an unshared ns-SNP in the genomic sequence coding for these MAPs (**Table 2.1**). Two other MAPs were detected in subject 1 but not subject 3, presumably because of an unidentified transacting factor since the MAP-coding transcripts and the relevant HLA allotypes were expressed in both subjects (**Table 2.2**). Peripheral blood mononuclear cells (PBMCs) from subject 1, 2 or 3 were co-cultured with autologous monocyte-derived dendritic cells (DCs) pulsed with one of the four cryptic MAPs (synthetic peptides). After culture for 12 days in the presence of interleukin (IL)-7 and IL-15, cells were harvested and $CD8^+$ cells were separated from $CD8^-$ cells using FACS. Elispot was then used to quantify interferon (IFN)-γ-producing cells in wells containing either CD8 T cells alone or together with peptide-pulsed or -unpulsed CD8 APCs. Non-polymorphic MAPs did not elicit a MAP-specific response (**Figure 2.7a**). However, polymorphic MAPs elicited a MAP-specific response since the frequency of IFN-γ-producing cells was much higher in the presence of peptide-pulsed than -unpulsed APCs (**Figure 2.7b**). We conclude that, at least *in vitro*, polymorphic cryptic MAPs can be immunogenic.

Table 2.1 | Features of polymorphic cryptic MAPs presented in Figure 2.7.

Polymorphic MAPs	Cryptic status	HLA	IC_{50} (nM)	Subject 1	Subject 2
I/MKQIKGGSL	Novel antisense	B*08:01	(I) 5,071.92 (M) 335.50	I/<u>M</u>	I
QPNF/LRVSTV	Exon - out	B*08:01	(F) 739.13 (L) 784.45	<u>F</u>/<u>L</u>	<u>F</u>

HLA, human leucocyte antigen; IC_{50}, half-maximal inhibitory concentration; MAP, MHC I-associated peptide; MHC, major histocompatibility complex; MS, mass spectrometry. The columns Subject 1 and Subject 2 indicate the peptide variant coded by transcripts found in each subject as well as a positive MS detection when the amino acid is underlined.

Table 2.2 | Features of non-polymorphic cryptic MAPs presented in Figure 2.7.

Non-polymorphic MAPs	Cryptic status	HLA	IC_{50} (nM)	Subject 1	Subject 3
AEARPTTVGF	Exon - out	B*44:03	119.38	AEA	AEA
VMKEKLLF	Intron	A*29:02	883.60	VMK	VMK

HLA, human leucocyte antigen; IC_{50}, half-maximal inhibitory concentration; MAP, MHC I-associated peptide; MHC, major histocompatibility complex; MS, mass spectrometry. The columns Subject 1 and Subject 3 indicate the peptide variant coded by transcripts found in each subject as well as a positive MS detection when the amino acid is underlined.

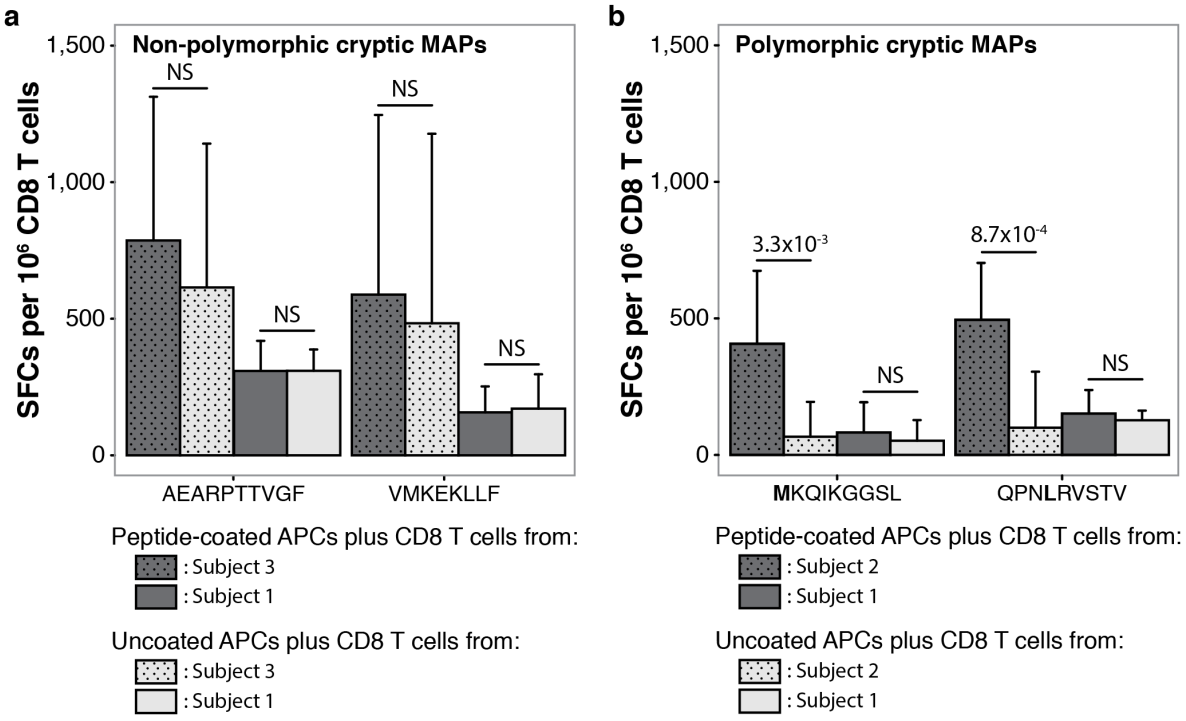


Figure 2.7 | Immunogenicity of Crypt. MAPs. (**a**,**b**) Only polymorphic Crypt. MAPs are immunogenic. IFN-γ Elispot counts showing the number of spot-forming cells (SFCs) per million CD8 T cells for two non-polymorphic (**a**) and two polymorphic (**b**) Crypt. MAPs. Final counts were obtained following the subtraction of background spots (peptide-coated APCs alone) from the spots obtained when CD8 T cells were exposed to peptide-coated or uncoated APCs. The experiment was performed in biological triplicates (each with three technical replicates), error bars represent s.d. and statistical significance was assessed using a two-tailed Student's t-test (NS: not significant, $P > 0.05$). Features of the four tested Crypt. MAPs are detailed in **Tables 2.1** and **2.2**.

2.6 Discussion

The present work demonstrates that proteogenomics can provide a systems-level perspective on the landscape of the cryptic immunopeptidome. The fact that a sizeable proportion of MAPs are cryptic (6.5–13% depending on stringency criteria) enhances the complexity of the immunopeptidome. If anything, we might have underestimated the proportion of cryptic MAPs in the immunopeptidome because our RNA-seq was performed on poly(A) tailed RNAs. The prevailing dogma holds that polyadenylation of RNA precursors is required for nuclear export and stability of mature transcripts and for efficient translation of mRNAs[45]. However, recent reports suggest that immature mRNA precursors can be translated in the nucleus and generate MAPs[29,46]. Further proteogenomic studies will therefore be needed to assess the potential contribution to the MAP repertoire of RNAs without poly(A) tail. In addition, RNA-seq-based proteogenomic studies may miss the rare MAPs derived from non-contiguous protein sequences via proteasome-mediated splicing[14].

About 50% of cryptic MAPs result from out-of-frame translation and the other half from translation of allegedly noncoding sequences. The ultimate biological role of cryptic translation remains elusive. However, it might be unwise to assume that this phenomenon merely represents ‘translational noise’. Protein synthesis is demanding: it is the most energy-consuming process in the cell as it monopolizes 45% of cellular ATP supplies[47]. Furthermore, any RNA sequence subject to translation will experience selection against encoding a protein with detrimental impact on cell function[6]. In any case, noncoding RNAs are vital, and our demonstration that several noncoding RNAs generate MAPs means that CD8 T cells have an opportunity to scrutinize these transcripts.

The gene source of conventional MAPs are enriched in microRNA-binding elements, suggesting that mRNA destabilization favours MAP generation[30]. By comparing transcripts coding conventional and cryptic MAPs, we obtained meaningful evidence suggesting that cryptic MAPs derive from particularly unstable transcripts targeted by NMD or 3’UTR destabilizing elements: (i) cryptic MAP transcripts were

enriched in upstream ORFs and their PCRs showed a strong 5' end positional bias (suggestive of premature termination) and (ii) cryptic MAP transcripts displayed longer 3'UTR enriched in destabilizing but not stabilizing elements when compared with conventional source transcripts. Together with previous work by us and others, these data allow for the development of an emerging model in which mRNA instability is instrumental in the genesis of all types of MAPs. This model is an extension of the idea that most MAPs derive from defective ribosomal products[32,48,49]: unstable RNAs targeted by NMD, microRNAs or other 3'UTR-destabilizing elements would generate more defective ribosomal products and therefore more MAPs. The validity of this model can be submitted to high-throughput experimental validation: if it is correct, mRNA half-life should be negatively correlated to MAP generation. We do not exclude that translation efficiency, which partly depends on codon usage[50], might also regulate MAP generation. Indeed, although we did not find evidence for a codon bias in conventional source transcripts versus cryptic MAP source ORFs ($P = 0.34$, odds ratio = 1.02), we observed that MAP source transcripts or ORFs in general use rare codons slightly more frequently than transcripts that do not generate MAPs ($P < 2.2 \times 10^{-16}$; odds ratio = 1.14; **Supplementary Tables 2.2** and **2.3**). Therefore, it might be interesting to further investigate the impact of codon bias on MAP generation.

Some 25 years ago, Boon and van Pel[51] proposed that MAPs might derive in a proteasome-independent manner from translation of short subgenic regions (peptons). This unorthodox hypothesis has progressively fallen into disfavour because no such MAPs were discovered with MS[39]. The present work argues that such MAPs do exist but can, in practice, be detected only by proteogenomics. Indeed, our cryptic MAPs were coded by extremely short ORFs, and the amino-acid composition of their C termini suggests that they are, at least in part, proteasome-independent.

One area where cryptic MAPs may be most relevant is cancer immunology. Although the vast majority of cancer mutations involve non-exomic regions, searches for tumour-specific antigens (TSAs) have focused on exomic mutations[31,52-54]. Nonetheless, since numerous noncoding transcripts are expressed only in cancer cells[55,56], a number of cryptic MAPs may be genuine TSAs. Furthermore, we

demonstrated that (i) cryptic MAP PCRs displayed a higher frequency of germline polymorphisms (ns-SNPs) than the conventional exome (**Figure 2.6c**) and that (ii) polymorphic cryptic MAPs discovered by proteogenomics were immunogenic (**Figure 2.7b**). Hence, it is reasonable to expect that cryptic MAPs bearing somatic mutations (that is, TSAs) should also be immunogenic. Accordingly, in melanoma and renal cell carcinoma, pioneering studies using more traditional approaches have uncovered unique immunogenic cryptic TSAs derived from noncoding regions[19,21]. Assuming that cryptic MAPs may be a rich source of heretofore overlooked TSAs, it is imperative to directly explore the presence of cryptic TSAs using systems-level approaches. Expanding the repertoire of TSAs would be highly beneficial because the low number of immunogenic exome-derived TSAs is a major hurdle for cancer immunotherapy[57-59].

2.7 Methods

2.7.1 Subject recruitment

Written informed consent was obtained from all study participants. The study protocol was approved by the Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont. Relative to subject 1 (HLA-A*03:01, -A*29:02; -B*08:01, -B*44:03), subjects 2–4 were HLA-identical, HLA-haploidentical (HLA-A*02:01, -A*29:02; -B*57:01, -B*44:03) or HLA-disparate (HLA-A*01:01, -A*02:01; -B*18:01, -B*39:24). See also **Supplementary Table 2.1**.

2.7.2 Analysis of RNA-seq data

RNA-seq was performed as described[22]. Paired-end RNA-seq data of subject 1 were mapped on the human reference genome (GRCh37.75) with the Casava 1.8.1 and Eland v2e mapping softwares (Illumina). This alignment was used to perform SNP calling with the Casava 1.8.1 software as previously described[22]. Only ns-SNPs having a Qmax_gt ≥ 20 were used to build the customized control database.

To obtain an expression value for each transcript of a given gene, paired-end RNA-seq data from subject 1 were mapped on the reference genome (GRCh37.75) using TopHat 2.0.10[60]. Cufflinks 2.2.1[61] was then run on the output- sorted BAM file in addition to the Ensembl gtf file to obtain FPKM (fragments per kilobase of transcript per million mapped reads) values for all known transcripts. Only transcripts having an FPKM value > 0 were considered as expressed.

2.7.3 Generation of the control and all-frames databases

We generated two customized databases based on the RNA-seq data of subject 1. To generate the control database, we applied a workflow similar to the one of Granados et al.[22]: ns-SNPs identified in subject 1 were integrated at their correct position in the reference genome (GRCh37.75) to build a personalized genome. Using the Ensembl gtf file, we extracted all known transcripts and further *in silico* translated them in their canonical reading frame to obtain the canonical proteome of subject 1. To

generate the all-frames database, we used all reads passing the Illumina quality filters and *in silico* translated them in the six possible reading frames using a sliding window of 33 bp to obtain all theoretical peptides having a length between 8 and 11 amino acids. For each peptide, we computed an S-value, that is, the number of times it was seen following the *in silico* translation process. Only peptides having an S-value ≥ 10 as well as a length between 8 and 11 amino acids were included in the predicted peptidome of subject 1. Both the canonical proteome and the predicted peptidome of subject 1 were compiled in fasta files to obtain the control and the all-frames database, respectively. Both databases were then concatenated with their respective decoy counterpart and submitted to the Mascot database search engine along with subject 1's immunopeptidomic data.

2.7.4 MS analyses

Immunopeptidomics raw data from subjects 1 and 2 B-LCL were obtained from a previous study[30]. For subjects 3 and 4, MAPs were eluted from B-LCLs and sequenced using MS as previously described (three to four biological replicates per subject)[22]. Each replicate was separated in six fractions using strong cation exchange chromatography. Vacuum-dried fractions were then suspended in 5% acetonitrile and 0.2% formic acid and injected into the LTQ-Orbitrap Elite operating at a resolving power of 60,000 (at *m/z* 400) for both full spectra and MS/MS spectra modes. Up to 10 precursor ions were accumulated to the target value of 50,000 with a maximum injection time of 100 ms. Mass spectra were analysed using the Xcalibur software and peak lists were generated with Mascot Distiller.

2.7.5 Control and all-frames database searches

The Mascot search engine (Matrix Science) was used in combination with the control or the all-frames database concatenated to their reverse database to identify peptides present in the immunopeptidome of subject 1. Mass tolerances on precursor and fragment ions were set to 5 p.p.m and 0.02 Da, respectively. Searches were performed without enzyme specificity, and cysteinylation, phosphorylation (on Ser, Thr

and Tyr), oxidation (Met) and deamidation (Asn, Gln) were used as variable modifications. Following each database search, we converted raw files to peptide maps containing m/z values, charge state, retention time and intensity above detection threshold (≥ 8,000) using ProteoProfile (http://proteomics.iric.ca/tools/ProteoProfile/)[62]. The peptide maps were used to extract the abundance of the identified peptides across the four replicates.

On the 8–11 amino-acid-long peptides identified with the control database, we computed the FDR[63] for all combinations of the Mascot score (which represents the confidence level of a peptide assignation) and predicted MHC-binding affinity (computed with NetMHCcons[64]). FDRs were computed as (number of decoy identifications/number of target identifications) × 100. We then selected the combination of the Mascot score and MHC-binding affinity yielding the higher number of MAPs at 5% FDR, as described[22]. The same Mascot score (≥ 22) and MHC-binding thresholds (≤ 1,250 nM) were then applied to the peptide list identified with the all-frames database. As expected, considering the unavoidable effect of database size on FDRs calculated according to decoy approaches[5,24,65], applying the thresholds defined with the control database to the all-frames database increased the decoy-based FDR to 9% for the all-frames database.

2.7.6 Identification of cryptic and conventional MAPs

Peptides identified with both the control and the all-frames databases were considered as conventional MAPs. Peptides solely identified by the all-frames database were considered as putative cryptic MAPs. To validate whether they were genuine cryptic MAPs, we mapped the subset of peptide-encoding reads using TopHat to discard peptides coming from multiple locations in the genome. The remaining cryptic MAP candidates were assigned to their respective source gene and their MS/MS spectra were manually validated. To determine the type of sequence (within the source gene) generating each cryptic MAP, we used the intersect function of the BEDTools suite on the bed file of our cryptic candidates as well as Ensembl gtf file. Peptides assigned to a gene source in the opposite orientation were classified as antisense

cryptic MAPs, those deriving from noncoding RNAs, 5'UTR, intronic, 3'UTR or intergenic sequences were classified as sense noncoding cryptic MAPs. Peptides deriving from exons of protein-coding genes were subjected to a reading frame validation: only peptides produced by non-canonical reading frames were classified as sense coding cryptic MAPs. For sense cryptic MAPs (except intergenic ones), we retrieved the gene biotype of their respective gene source from Ensembl annotations (when available) using pyGeno. Finally, since MAPs derive preferentially from highly abundant transcript[30,31,66], we assumed that the conventional and sense cryptic MAPs passing all of our filtering steps were generated by the most highly expressed isoform of their respective source gene. A complete list of identified conventional and cryptic MAPs can be found in Supplementary Data 1 and Supplementary Data 2, respectively.

2.7.7 Computation of PCR coverage

We computed the coverage of all identified PCRs by using the coverage function of the BEDTools suite. The sorted BAM file obtained following the TopHat alignment as well as the bed files of our cryptic and conventional PCRs were used as entry files. This coverage metrics, which represents the number of reads overlapping, by at least 1 bp, our PCRs were then correlated with the S-value metrics, which approximates the number of read fully overlapping the same PCRs (**Supplementary Figure 2.1a**).

2.7.8 Influence of the HLA genotype on the MAP repertoire

The Mascot search engine was used to perform database searches on the raw data of subjects 2–4 against a validation database that contained all identifications made in subject 1 as well as their decoy sequences. Mass tolerances on the precursor and fragment ions were set to 5 p.p.m and 0.02 Da, respectively. Peptide lists identified in each subject were extracted and compared with the 2,435 conventional and 168 cryptic MAPs identified in subject 1 (**Supplementary Figure 2.4**).

2.7.9 Prediction of upstream ORFs

For each transcript source of MAPs, we extracted the personalized mRNA sequences of subject 1 using pyGeno. We scanned the transcript from its 5'end to its 3'end to predict all possible ORFs initiating at an AUG embedded in an optimal (GCC[R]CCstartG[V]) or strong ([R]NNstartG[V]) Kozak context. ORFs located in the 5'UTR or at the 5'UTR–exon junction were considered as upstream ORFs. We computed the proportion of the transcript source of cryptic and/or conventional MAPs that presented at least one upstream ORF. Statistical significance between the cryptic and conventional source transcript categories was assessed using a two-sided Fisher's exact test. This analysis was performed on sense cryptic MAPs for which a source gene and transcript were available.

2.7.10 mRNA stability analysis

Using pyGeno, we retrieved the 3'UTR sequences of cryptic and conventional source transcripts to compute their length, their number of intronic sequences and to look for exact match of all destabilizing and stabilizing elements characterized by Zhao W. et al.[36]. The 3'UTR length distributions as well as the number of destabilizing and stabilizing elements per transcript were compared between the transcript source of conventional and cryptic MAPs. Statistical significance was assessed using a two- and a one-sided Wilcoxon rank sum test, respectively. Statistical significance for the proportion of conventional and cryptic MAP source transcripts containing no versus at least one intron was assessed using a two-sided Fisher's exact test. This analysis was performed on sense cryptic MAPs for which a source gene and transcript were available.

2.7.11 Prediction of cryptic source proteins

To predict the probable start codon of each cryptic PCR, we sequentially applied the following rules: (i) presence of an upstream AUG within an optimal (GCC[R]CCstartG[V]), strong ([R]NNstartG[V]) or weak (anything else) Kozak context, (ii) presence of an upstream near-cognate AUG within an optimal or strong Kozak

context, (iii) any other codon downstream of the first upstream stop codon. The probable stop codon was assumed to be the first in-frame stop codon downstream of the PCR. This analysis was performed on personalized mRNA sequences of cryptic source transcripts for most sense cryptic MAPs. Since no gene structures were known for antisense, intronic and intergenic cryptic MAPs, we simply extracted the personalized genomic sequences flanking the PCR (750 bp long) and performed the same analysis.

2.7.12 C-terminal amino-acid signature

At each position analysed, we compared the usage of each amino-acid class between cryptic and conventional MAPs using a two-sided Fisher's exact test. Hits were considered significant when they yielded a *P* value < 0.05.

2.7.13 ns-SNP frequency analysis

We used dbSNP138 (common_all set) to determine the frequency of ns-SNPs, at the population level, in the PCRs of conventional and cryptic MAPs. Since some cryptic MAPs derive from out-of-frame exonic translation, we could not rely on the synonymous versus non-synonymous dbSNP annotations. To circumvent this problem, we sequentially inserted all SNPs intersecting with our cryptic and conventional PCRs (stored in bed files). Those mutated PCRs were then *in silico* translated. If the resulting peptide was identical to the MAP initially identified in subject 1, the SNP was classified as synonymous. Otherwise, the SNP was classified as non-synonymous. Knowing the number of bp encoding our cryptic and conventional MAPs, we computed the frequency of ns-SNPs per bp observed in both types of PCRs. Statistical significance was assessed using a two-sided Fisher's exact test.

2.7.14 Rare codon usage analysis

Codons were classified as rare and common if their observed usage frequency (http://www.genscript.com/cgi-bin/tools/codon_freq_table)[67] was lower and greater than their expected usage frequency (1/number of codons encoding a given amino

acid), respectively. Out of 64 codons, 30 were classified as rare and 34 as common. Using an in-house python script, we computed the number of occurrence for each codon to further derive the number of rare and common codons used by each class of transcripts across (1) conventional source transcripts versus cryptic source ORFs and (2) MAP source transcripts or ORFs versus all the other transcripts for which a cDNA sequence was defined. Statistical significance was assessed using a two-sided Fisher's exact test.

2.7.15 T-cell priming and IFN-γ Elispot assays

Monocyte-derived DCs were generated from frozen PBMCs, as previously described[68]. Peptide-specific $CD8^+$ T cells were expanded as described, with some minor modifications[69]. Briefly, thawed PBMCs were first T-cell-enriched using the Easysep Human T Cell Enrichment Kit (StemCell Technologies) and co-incubated with autologous peptide-pulsed DCs at a DC:T cell ratio of 1:4 with the addition of IL-21 (30 $ng.mL^{-1}$). Cells were cultured in CellGro DC medium containing 5% human serum and L-glutamine. IL-15 (2.5 $ng.mL^{-1}$) and IL-7 (2.5 $ng.mL^{-1}$) were added on day 3 and every 3 days thereafter. On day 12, cells were harvested and stained with an anti-human CD8-PE as recommended by the manufacturer (clone RPA-T8, BD Biosciences). $CD8^+$ T and $CD8^-$ cells were sorted using a FACSAria apparatus and then used for the Elispot assays, which were performed as described[70]. IFN-γ production was expressed as the number of peptide-specific spot-forming cells per 10^6 $CD8^+$ T cells after subtracting the spot counts from negative control wells (CD8 T cells alone).

2.7.16 Data analysis and visualization

Unless stated otherwise, analyses were performed using the pyGeno python package (https://github.com/tariqdaouda/ pyGeno)[23]. The ggplot2 package from the R software was used for data visualization. All codes are available on request to the corresponding author.

2.8 Acknowledgements

We are grateful to our blood donors for their generosity. We also thank Caroline Côté for her help in the generation of immunopeptidomics data from subjects 3 and 4. This work was supported by the Canadian Cancer Society (Grant number 701564). C.P. and P.T. hold Canada Research Chairs in Immunobiology and in Proteomics and Bioanalytical Spectrometry, respectively.

2.9 Additional Information

2.9.1 Accession codes

RNA-seq data for the four subjects are available in the Gene Expression Omnibus database under accession code GSE67174 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=upkvweysnxabzkr&acc=GSE67174). MS data are available in PeptideAtlas for subjects 1 and 2 (http://www.peptideatlas.org/PASS/PASS00270); data from subjects 3 and 4 have been submitted to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifiers PXD001898 (project accession) and 10.6019/PXD001898 (project DOI). In addition, the entire list of MAPs identified in subject 1 has been deposited into the Immune Epitope Database (http://www.iedb.org) under accession code 1028836.

2.9.2 Competing financial interests

The authors declare no competing financial interests.

2.10 References

1. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
2. Kim, M.S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575-81.
3. Weekes, M.P., et al., *Quantitative temporal viromics: an approach to investigate host-pathogen interaction.* Cell, 2014. **157**(6): p. 1460-72.
4. Alfaro, J.A., et al., *Onco-proteogenomics: cancer proteomics joins forces with genomics.* Nat Methods, 2014. **11**(11): p. 1107-13.
5. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.
6. Ingolia, N.T., et al., *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.
7. Branca, R.M., et al., *HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics.* Nat Methods, 2014. **11**(1): p. 59-62.
8. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.
9. Zhang, B., et al., *Proteogenomic characterization of human colon and rectal cancer.* Nature, 2014. **513**(7518): p. 382-7.
10. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.
11. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.
12. Hassan, C., et al., *The human leukocyte antigen-presented ligandome of B lymphocytes.* Mol Cell Proteomics, 2013. **12**(7): p. 1829-43.
13. Mester, G., V. Hoffmann, and S. Stevanovic, *Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands.* Cell Mol Life Sci, 2011. **68**(9): p. 1521-32.
14. Granados, D.P., et al., *The nature of self for T cells-a systems-level perspective.* Curr Opin Immunol, 2015. **34**: p. 1-8.

15. Lev, A., et al., *Compartmentalized MHC class I antigen processing enhances immunosurveillance by circumventing the law of mass action.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6964-9.
16. de Verteuil, D., et al., *Origin and plasticity of MHC I-associated self peptides.* Autoimmun Rev, 2012. **11**(9): p. 627-35.
17. Malarkannan, S., M. Afkarian, and N. Shastri, *A Rare Cryptic Translation Product Is Presented by Kb Major Histocompatibility Complex Class I Molecule to Alloreactive T Cells.* J Exp Med, 1995. **182**: p. 1739-1750.
18. Starck, S.R., et al., *Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I.* Science, 2012. **336**(6089): p. 1719-23.
19. Starck, S.R. and N. Shastri, *Non-conventional sources of peptides presented by MHC class I.* Cell Mol Life Sci, 2011. **68**(9): p. 1471-9.
20. Goodenough, E., et al., *Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR.* Proc Natl Acad Sci U S A, 2014. **111**(15): p. 5670-5.
21. Weinzierl, A.O., et al., *A cryptic vascular endothelial growth factor T-cell epitope: identification and characterization by mass spectrometry and T-cell assays.* Cancer Res, 2008. **68**(7): p. 2447-54.
22. Granados, D.P., et al., *Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides.* Nat Commun, 2014. **5**: p. 3600.
23. Daouda, T., C. Perreault, and S. Lemieux, *pyGeno: A Python package for precision medicine and proteogenomics.* F1000Res, 2016. **5**: p. 381.
24. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* J Proteomics, 2010. **73**(11): p. 2092-123.
25. Sidney, J., et al., *Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype.* J Immunol, 2003. **171**(11): p. 5964-74.
26. Pelechano, V. and L.M. Steinmetz, *Gene regulation by antisense transcription.* Nat Rev Genet, 2013. **14**(12): p. 880-93.

27. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
28. Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**(3): p. 193-204.
29. Apcher, S., et al., *Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17951-6.
30. Granados, D.P., et al., *MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements.* Blood, 2012. **119**(26): p. e181-191.
31. Yadav, M., et al., *Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.* Nature, 2014. **515**(7528): p. 572-6.
32. Yewdell, J.W., *DRiPs solidify: progress in understanding endogenous MHC class I antigen processing.* Trends Immunol, 2011.
33. Apcher, S., et al., *Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation.* Proc Natl Acad Sci U S A, 2011. **108**(28): p. 11572-7.
34. Smith, J.E., et al., *Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae.* Cell Rep, 2014. **7**(6): p. 1858-66.
35. Kim, Y., et al., *Positional bias of MHC class I restricted T-cell epitopes in viral antigens is likely due to a bias in conservation.* PLoS Comput Biol, 2013. **9**(1): p. e1002884.
36. Zhao, W., et al., *Massively parallel functional annotation of 3' untranslated regions.* Nat Biotechnol, 2014. **32**(4): p. 387-91.
37. Ivanov, I.P., et al., *Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1).* Proc Natl Acad Sci U S A, 2010. **107**(42): p. 18056-60.
38. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.

39. Yewdell, J.W., E. Reits, and J. Neefjes, *Making sense of mass destruction: quantitating MHC class I antigen presentation.* Nat Rev Immunol, 2003. **3**(12): p. 952-61.
40. Blum, J.S., P.A. Wearsch, and P. Cresswell, *Pathways of antigen processing.* Annu Rev Immunol, 2013. **31**: p. 443-73.
41. Weimershaus, M., et al., *Peptidases trimming MHC class I ligands.* Curr Opin Immunol, 2013. **25**(1): p. 90-6.
42. de Verteuil, D., et al., *Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules.* Mol Cell Proteomics, 2010. **9**(9): p. 2034-47.
43. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nat Rev Immunol, 2011. **11**(12): p. 823-36.
44. Mishto, M., et al., *Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation.* Eur J Immunol, 2014. **44**(12): p. 3508-21.
45. Elkon, R., A.P. Ugalde, and R. Agami, *Alternative cleavage and polyadenylation: extent, regulation and function.* Nat Rev Genet, 2013. **14**(7): p. 496-506.
46. David, A., et al., *Nuclear translation visualized by ribosome-bound nascent chain puromycylation.* J Cell Biol, 2012. **197**(1): p. 45-57.
47. Princiotta, M.F., et al., *Quantitating protein synthesis, degradation, and endogenous antigen processing.* Immunity, 2003. **18**(3): p. 343-54.
48. Apcher, S., B. Manoury, and R. Fahraeus, *The role of mRNA translation in direct MHC class I antigen presentation.* Curr Opin Immunol, 2012. **24**(1): p. 71-6.
49. Bourdetsky, D., C.E. Schmelzer, and A. Admon, *The nature and extent of contributions by defective ribosome products to the HLA peptidome.* Proc Natl Acad Sci U S A, 2014. **111**(16): p. E1591-9.
50. Quax, T.E., et al., *Codon Bias as a Means to Fine-Tune Gene Expression.* Mol Cell, 2015. **59**(2): p. 149-61.
51. Boon, T. and A. Van Pel, *T cell-recognized antigenic peptides derived from the cellular genome are not protein degradation products but can be generated*

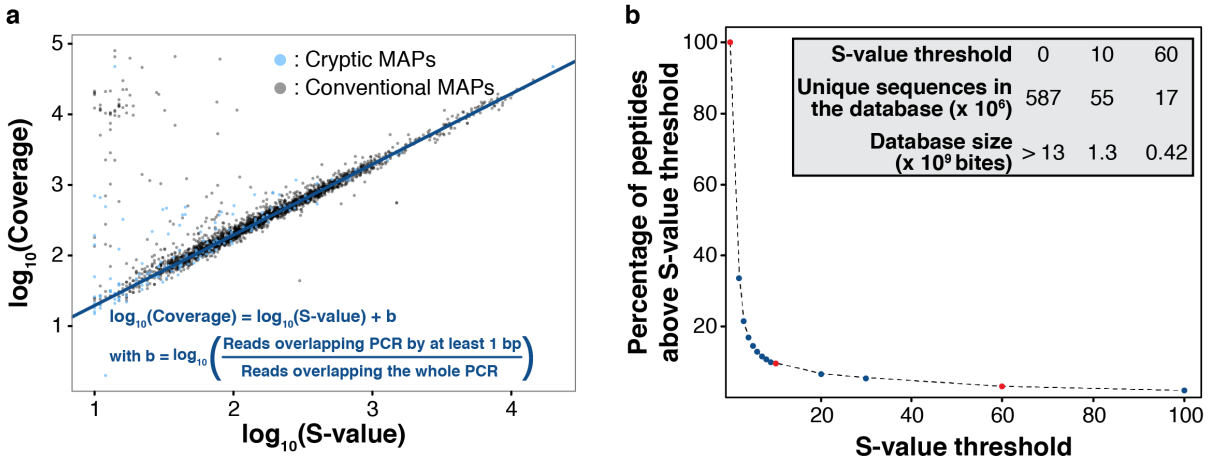
directly by transcription and translation of short subgenic regions. A hypothesis. Immunogenetics, 1989. **29**(2): p. 75-9.

52. van Rooij, N., et al., *Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma.* J Clin Oncol, 2013. **31**(32): p. e439-42.
53. Robbins, P.F., et al., *Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells.* Nat Med, 2013. **19**(6): p. 747-52.
54. Gubin, M.M., et al., *Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens.* Nature, 2014. **515**(7528): p. 577-81.
55. White, N.M., et al., *Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer.* Genome Biol, 2014. **15**(8): p. 429.
56. Trimarchi, T., et al., *Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia.* Cell, 2014. **158**(3): p. 593-606.
57. Coulie, P.G., et al., *Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy.* Nat Rev Cancer, 2014. **14**(2): p. 135-46.
58. Hinrichs, C.S. and N.P. Restifo, *Reassessing target antigens for adoptive T-cell therapy.* Nat Biotechnol, 2013. **31**(11): p. 999-1008.
59. Heemskerk, B., P. Kvistborg, and T.N. Schumacher, *The cancer antigenome.* EMBO J, 2013. **32**(2): p. 194-203.
60. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.
61. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
62. Thibault, P. *ProteoProfile*. 2015; Available from: https://proteomics.iric.ca/tools/ProteoProfile/.
63. Sennels, L., J.C. Bukowski-Wills, and J. Rappsilber, *Improved results in proteomics by use of local and peptide-class specific false discovery rates.* BMC Bioinformatics, 2009. **10**: p. 179.

64. Karosiene, E., et al., *NetMHCcons: a consensus method for the major histocompatibility complex class I predictions.* Immunogenetics, 2012. **64**(3): p. 177-86.
65. Blakeley, P., I.M. Overton, and S.J. Hubbard, *Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies.* J Proteome Res, 2012. **11**(11): p. 5221-34.
66. Fortier, M.H., et al., *The MHC class I peptide repertoire is molded by the transcriptome.* J Exp Med, 2008. **205**(3): p. 595-610.
67. GenScript bioinformatic tools. *GenScript codon usage frequence table tool*. 2015; Available from: http://www.genscript.com/cgi-bin/tools/codon_freq_table.
68. Bollard, C.M., et al., *Complete responses of relapsed lymphoma following genetic modification of tumor-antigen presenting cells and T-lymphocyte transfer.* Blood, 2007. **110**(8): p. 2838-45.
69. Wolfl, M. and P.D. Greenberg, *Antigen-specific activation and cytokine-facilitated expansion of naive, human CD8+ T cells.* Nat Protoc, 2014. **9**(4): p. 950-66.
70. Vincent, K., et al., *Rejection of leukemic cells requires antigen-specific T cells with high functional avidity.* Biol Blood Marrow Transplant, 2014. **20**(1): p. 37-45.

2.11 Supplementary Information

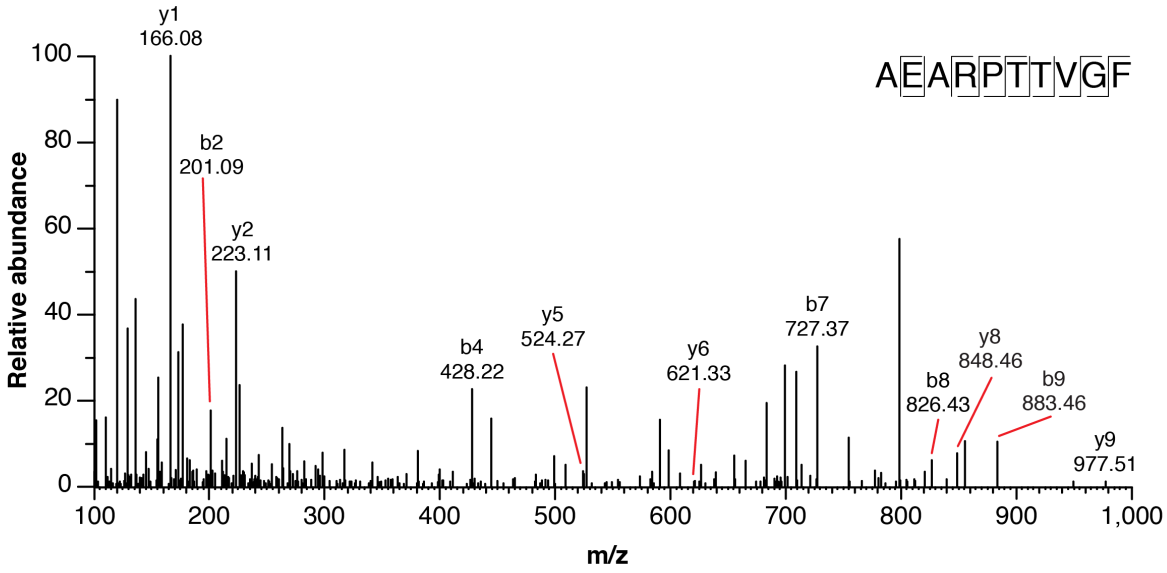
2.11.1 Supplementary Figures



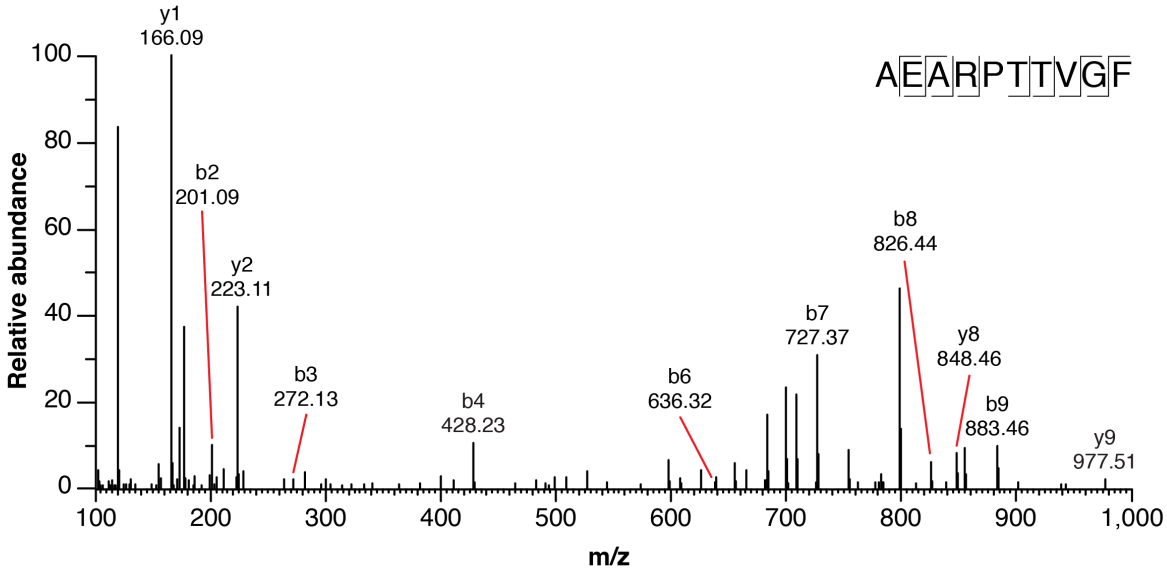
Supplementary Figure 2.1 | Selection of the S-value threshold. (**a**) The S-value metrics correlates with the coverage. Following RNA-seq reads mapping on the reference genome (version GRCh37.75) using TopHat, we computed the coverage of each peptide-coding region (PCR). For each identified MAP, we then plotted this coverage metrics as a function of the S-value, both in $\log_{10}$. Since the S-value metrics approximates the number of reads spanning the whole PCR while the coverage also takes into account reads spanning PCR by at least one base pair (bp), we reasoned that the coverage would slightly overestimate the S-value. For a few MAPs, the S-value appeared to underestimate the coverage: analysis of those MAPs revealed that most of them derived from genes in the RPS and RPL families. Highly similar in sequence but not identical, these genes will generate many 11-amino acids entries in the all-frames database, having different S-value but all containing the 9-amino acids MAPs of interest. Since we kept only one S-value among all possible ones, the real S-value of this 9-mers was therefore a strong underestimation of the coverage given by TopHat especially since this mapper maps multihit reads. (**b**) An S-value threshold ≥ 10 yields a database having a size manageable by Mascot search engine. This graph represents the percentage of peptides above the S-value threshold as a function of the S-value threshold. Number of unique sequences in the database as well as its size (in bites) were computed for three S-value thresholds (red dots) as detailed in the table. For comparison, the size of a typical reference protein database, such as UniProt, is about 0.5×10^9 bites.

a

Endogenous peptide (related to Fig. 7)

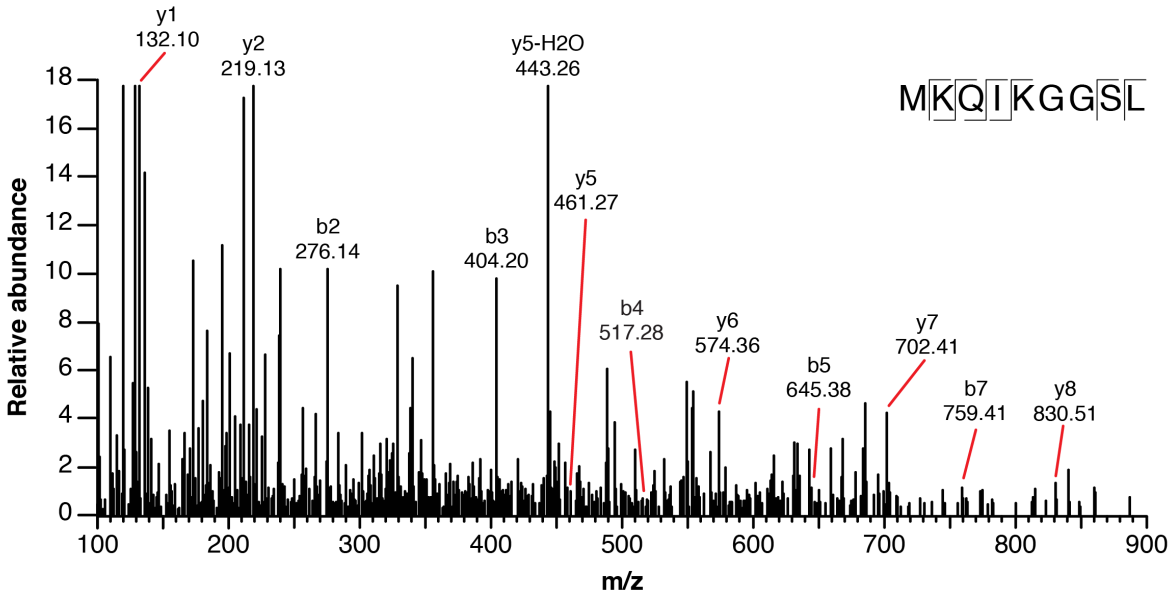


Synthetic peptide (related to Fig. 7)

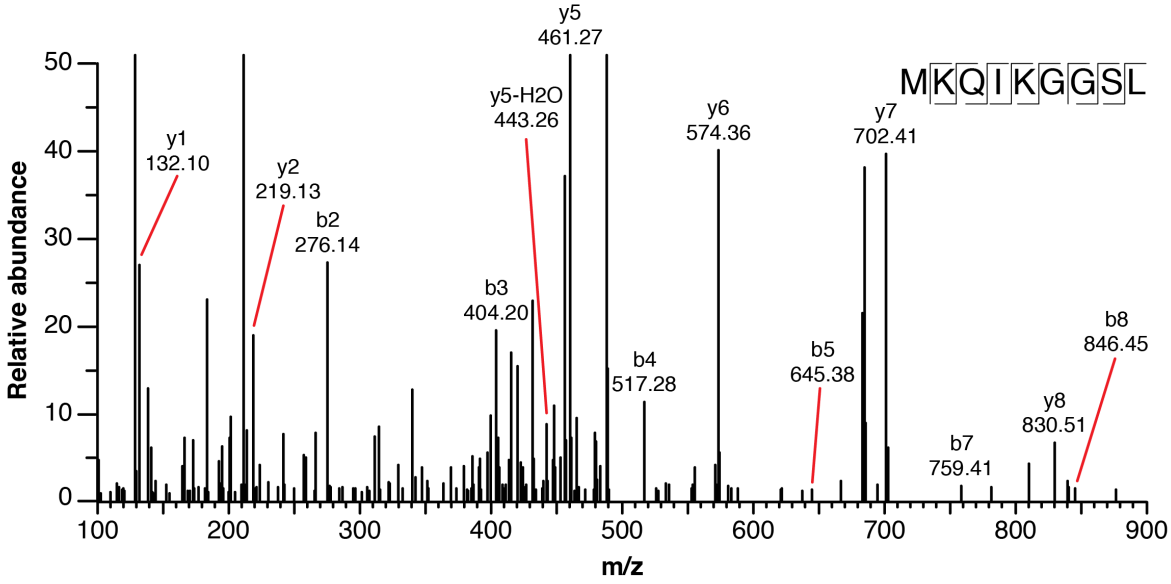


b

Endogenous peptide (related to Fig. 7)

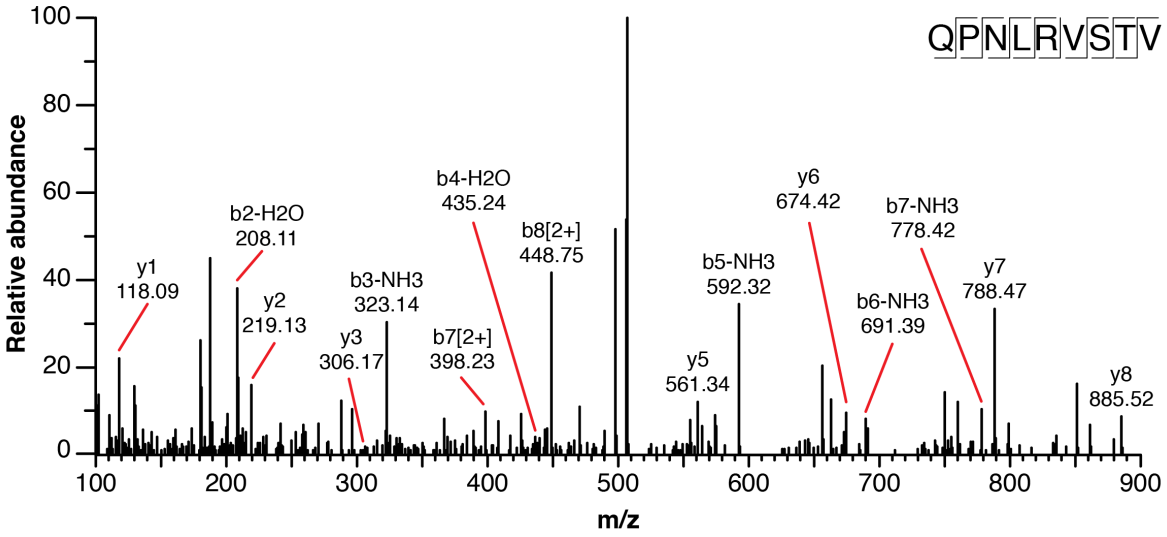


Synthetic peptide (related to Fig. 7)

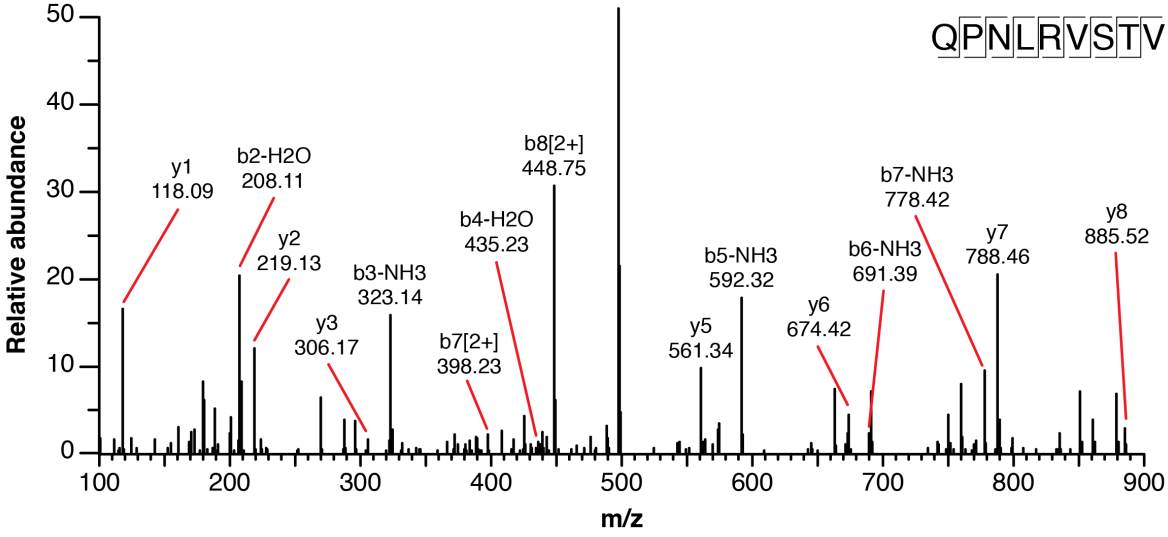


c

Endogenous peptide (related to Fig. 7)

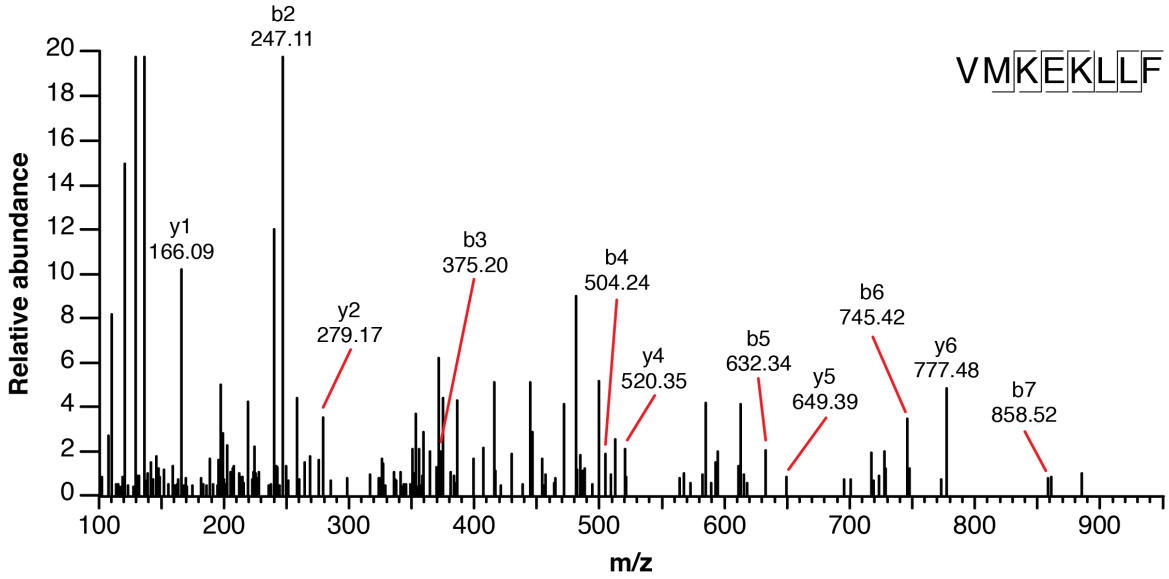


Synthetic peptide (related to Fig. 7)

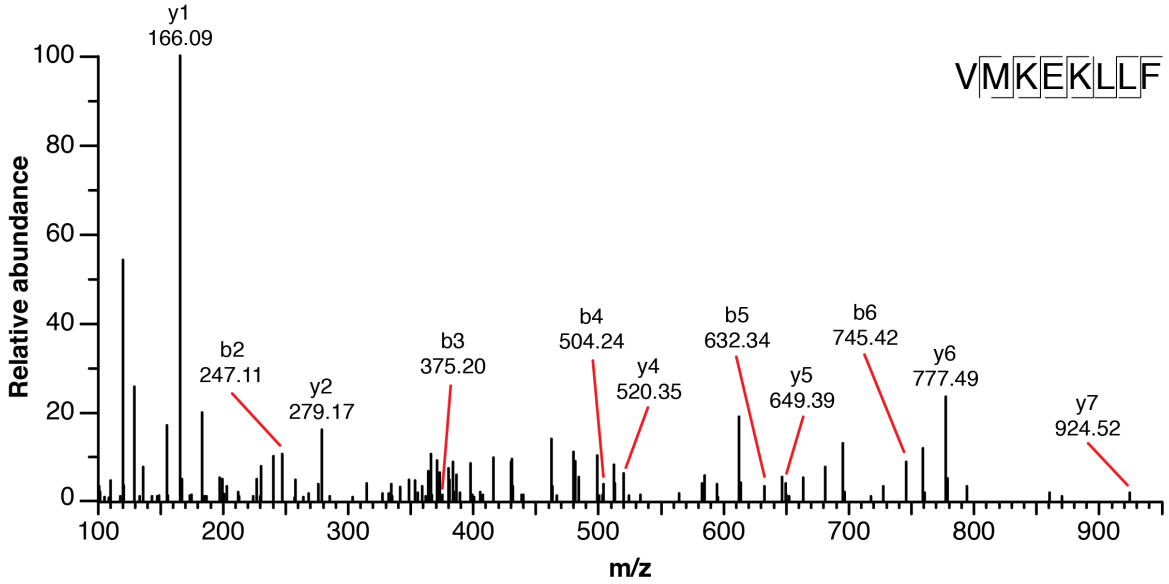


d

Endogenous peptide (related to Fig. 7)

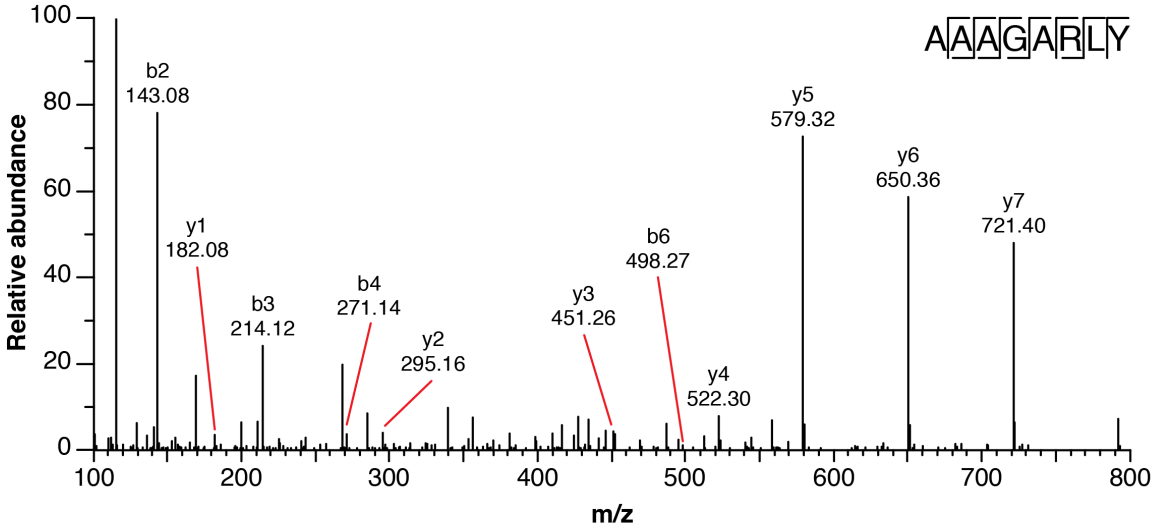


Synthetic peptide (related to Fig. 7)

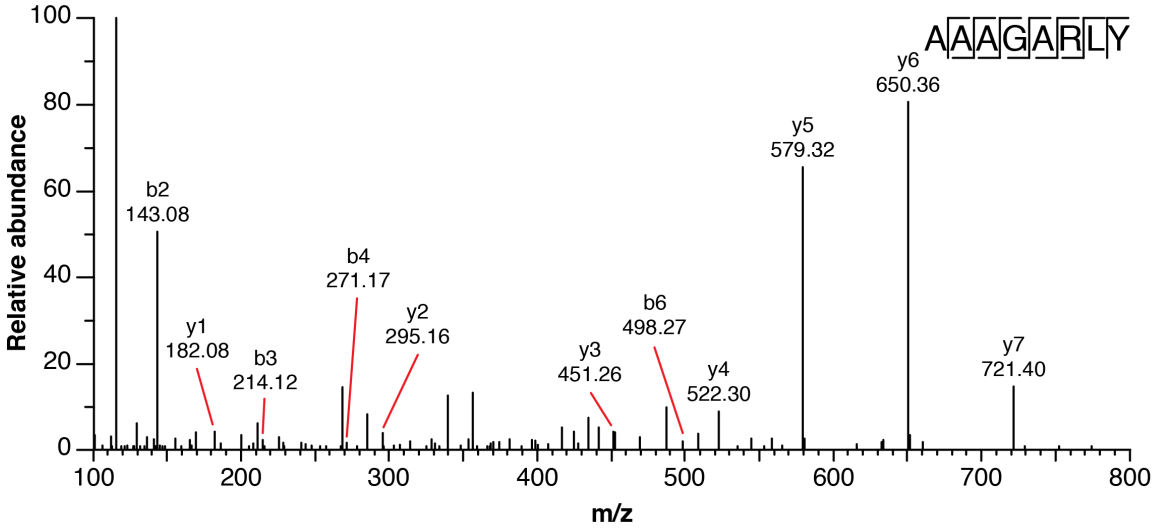


e

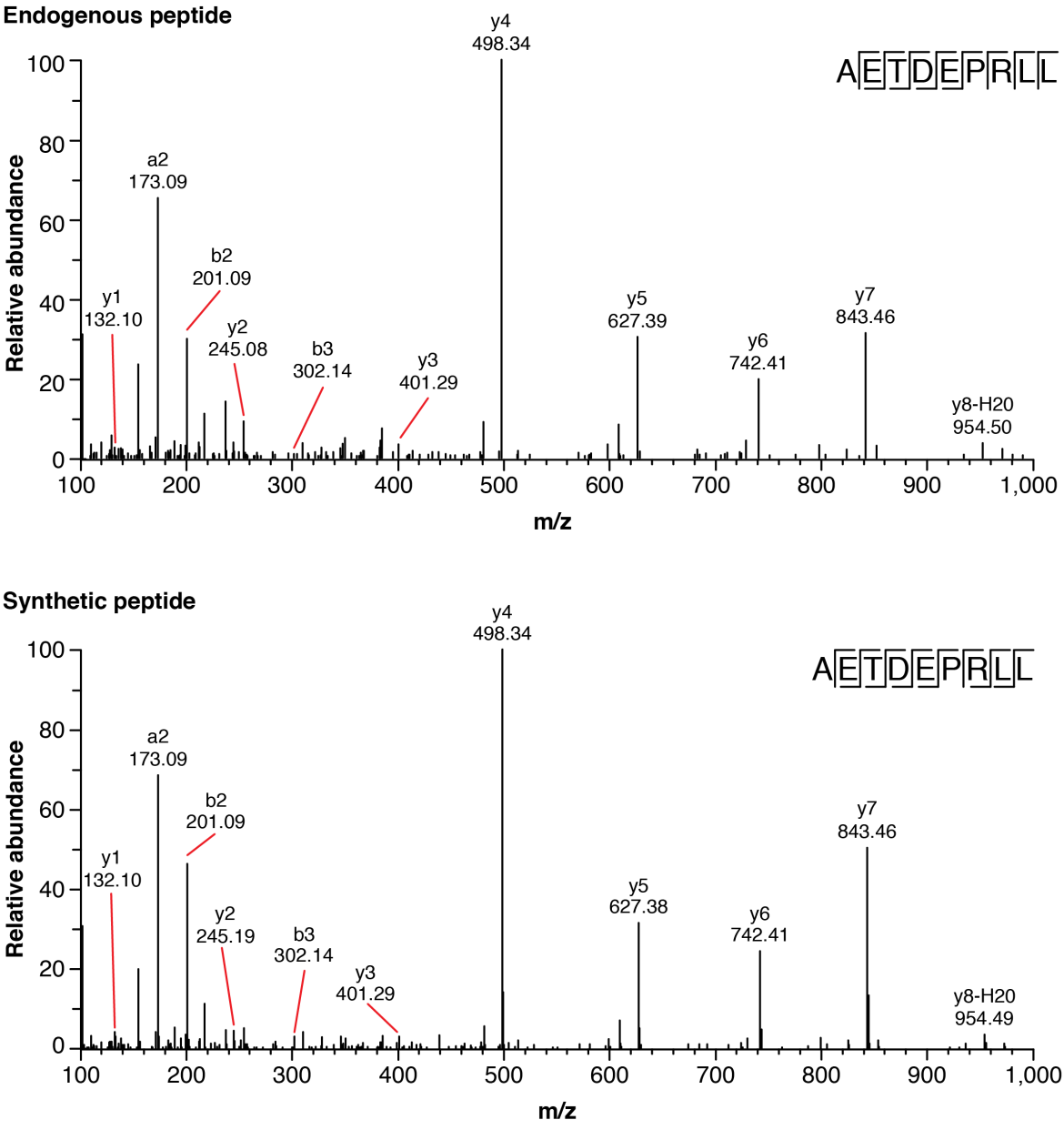
Endogenous peptide



Synthetic peptide

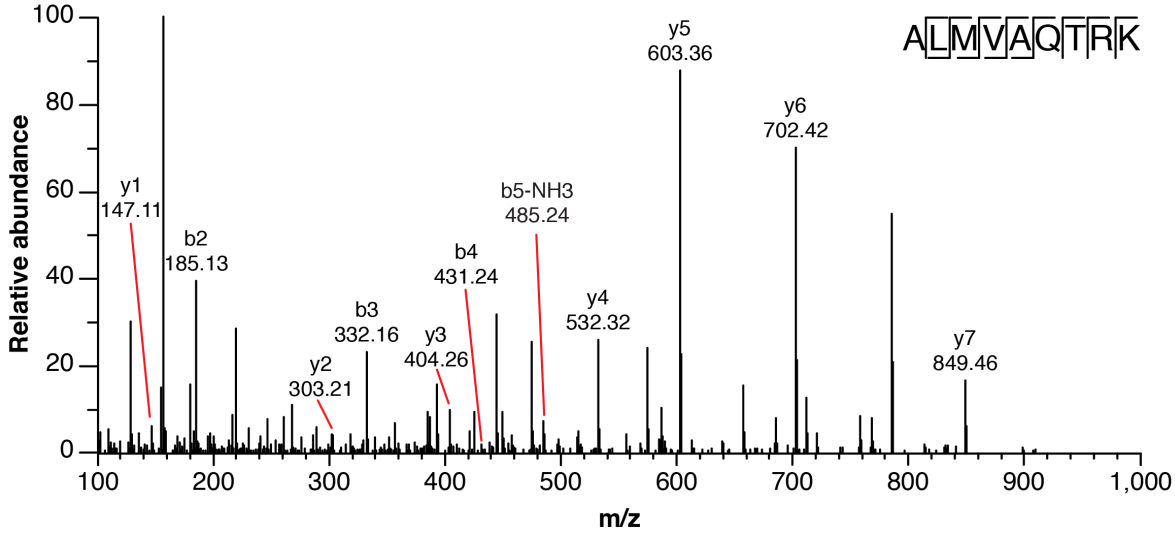


f

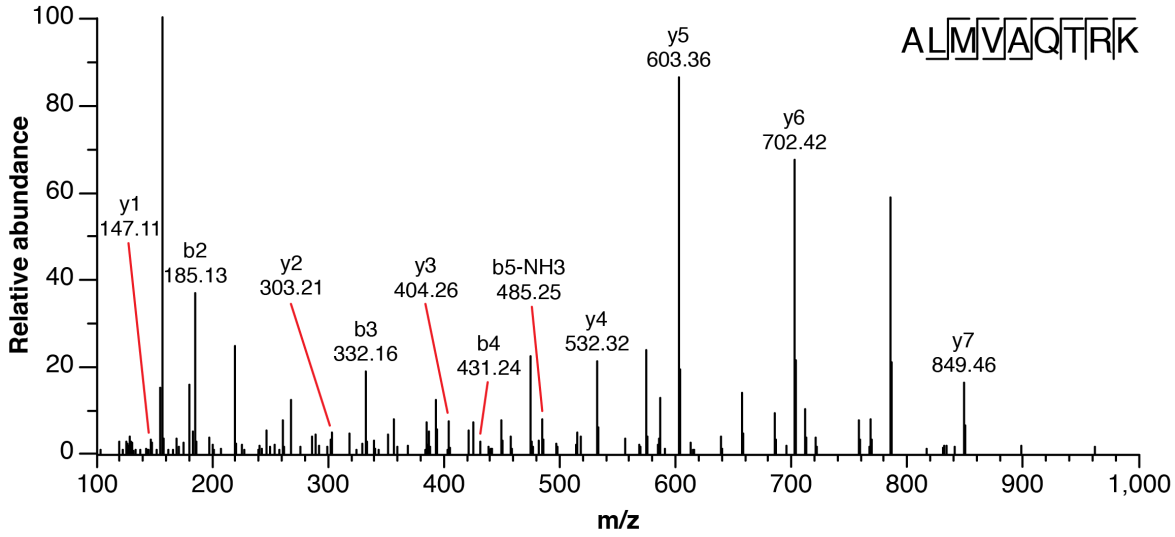

Endogenous peptide
AETDEPRLL
Relative abundance
m/z
y4 498.34
a2 173.09
b2 201.09
y1 132.10
y2 245.08
b3 302.14
y3 401.29
y5 627.39
y6 742.41
y7 843.46
y8-H20 954.50
Synthetic peptide
AETDEPRLL
Relative abundance
m/z
y4 498.34
a2 173.09
b2 201.09
y1 132.10
y2 245.19
b3 302.14
y3 401.29
y5 627.38
y6 742.41
y7 843.46
y8-H20 954.49

g

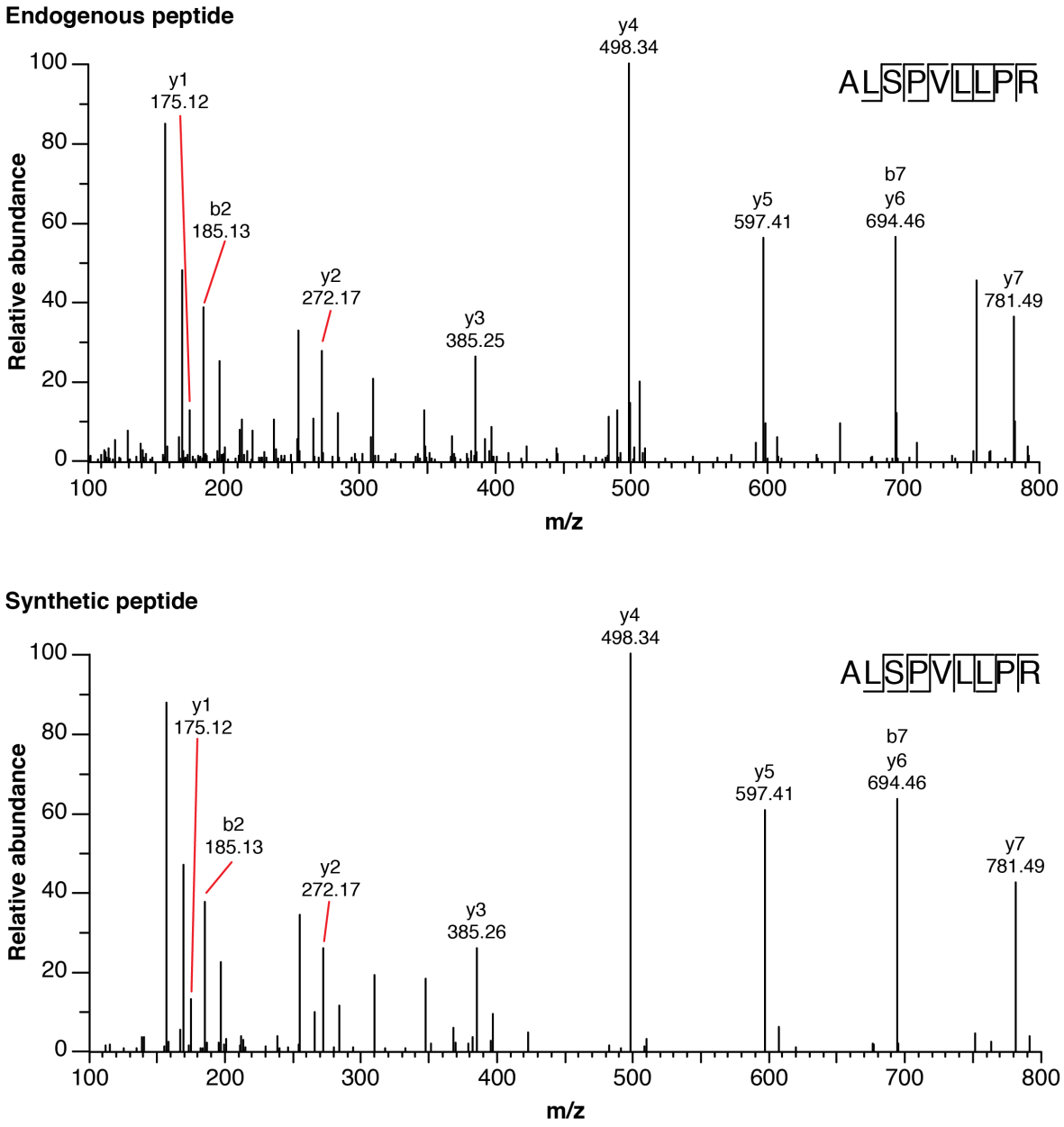
Endogenous peptide



Synthetic peptide

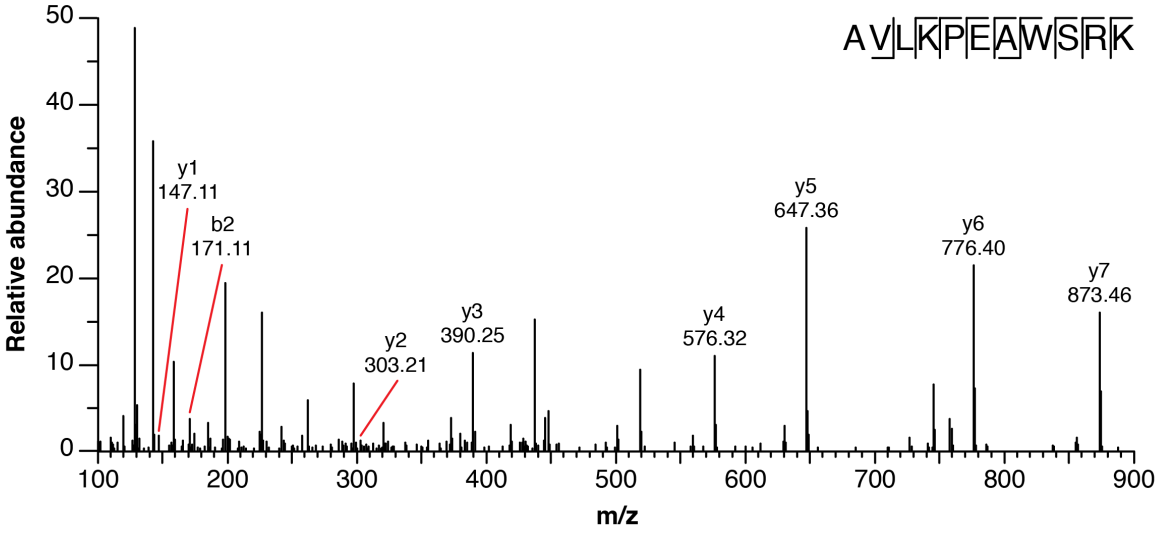


h

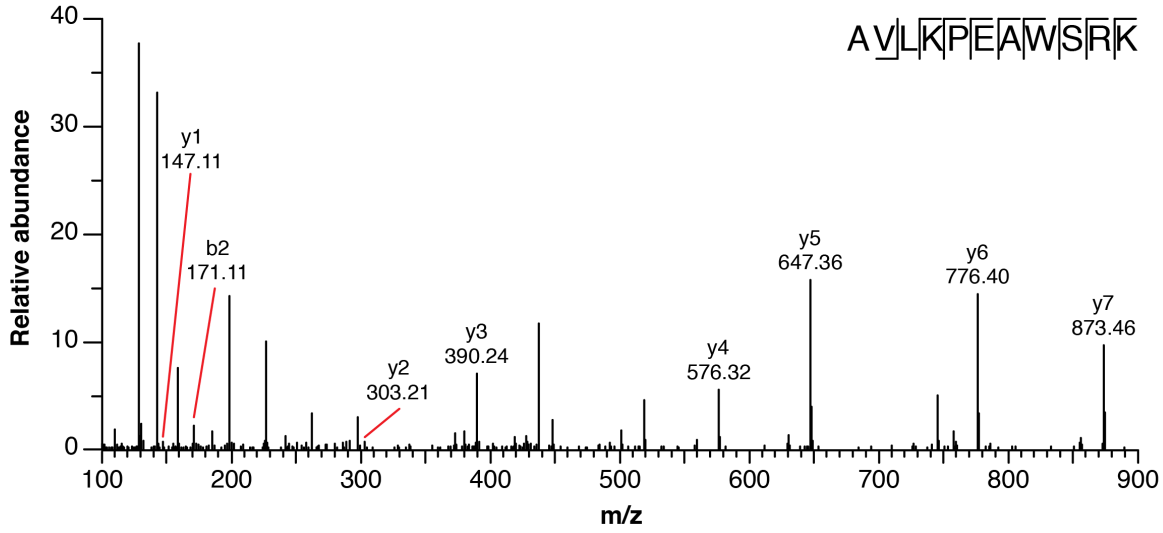

Endogenous peptide
Relative abundance
m/z
ALSPVLLPR
y1
175.12
b2
185.13
y2
272.17
y3
385.25
y4
498.34
y5
597.41
b7
y6
694.46
y7
781.49
Synthetic peptide
Relative abundance
m/z
ALSPVLLPR
y1
175.12
b2
185.13
y2
272.17
y3
385.26
y4
498.34
y5
597.41
b7
y6
694.46
y7
781.49

i

Endogenous peptide

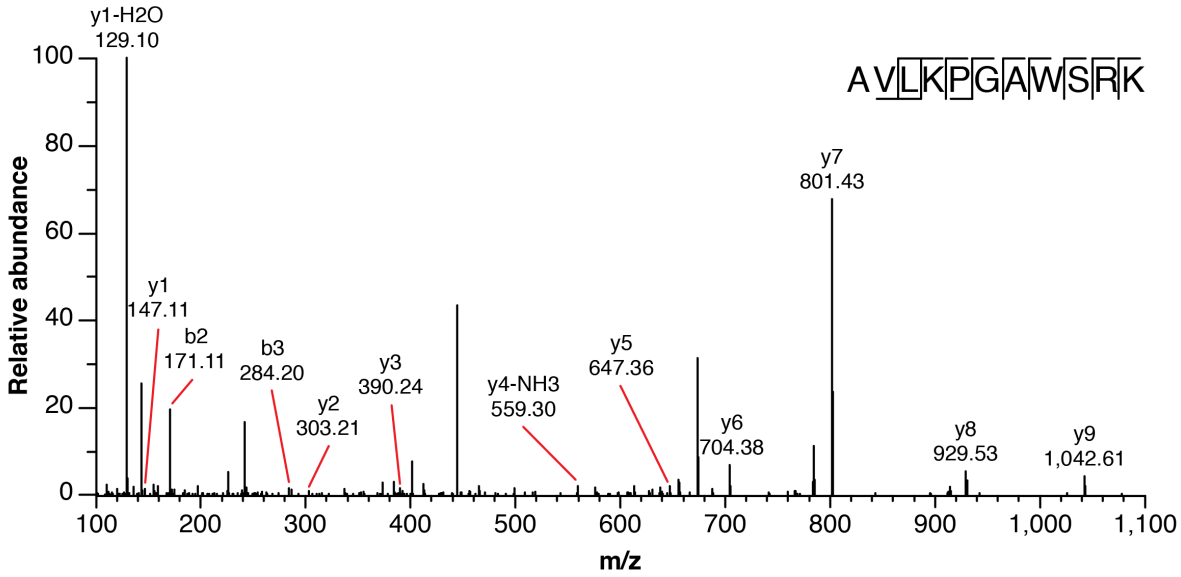


Synthetic peptide

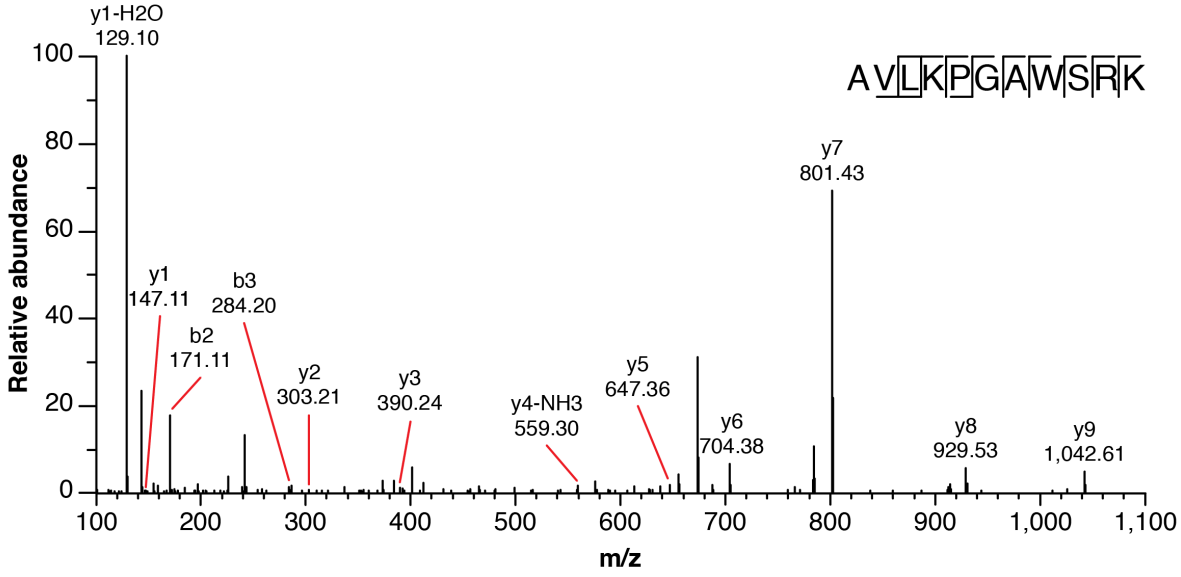


j

Endogenous peptide

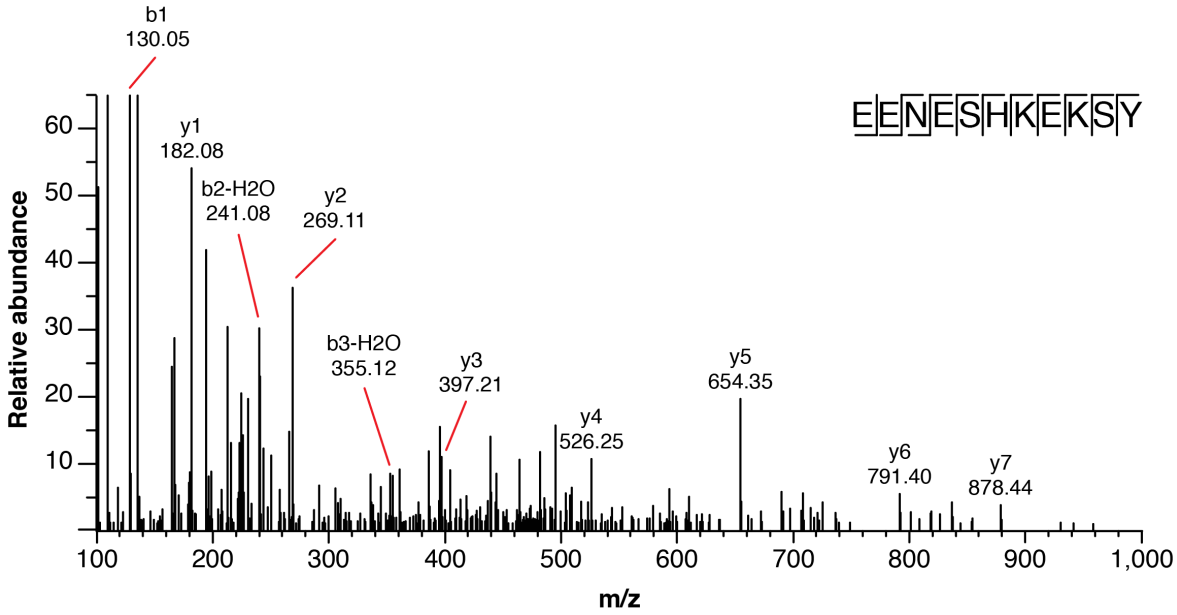


Synthetic peptide

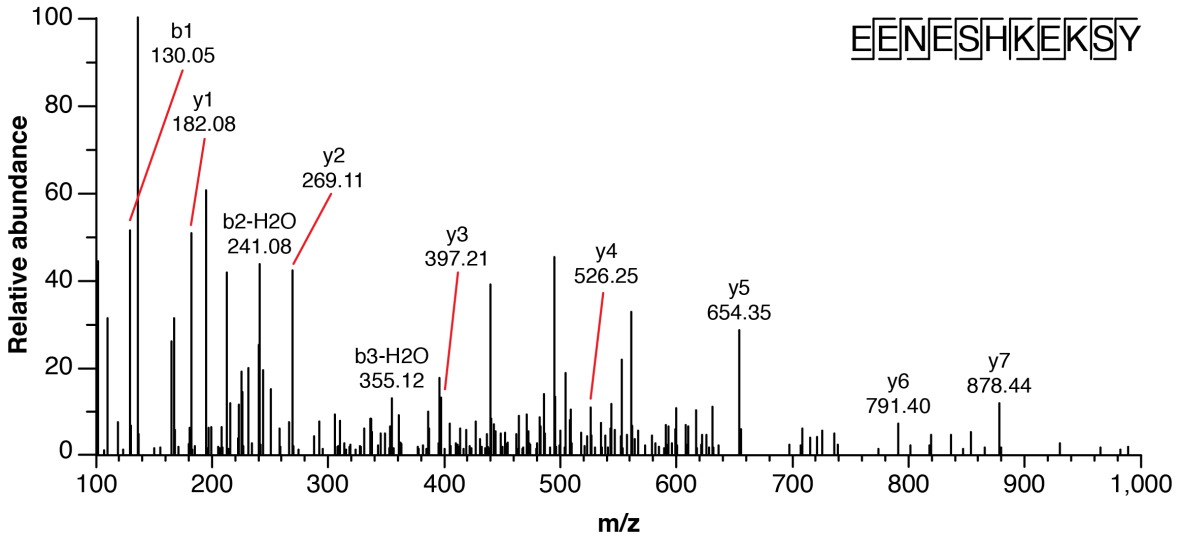


k

Endogenous peptide

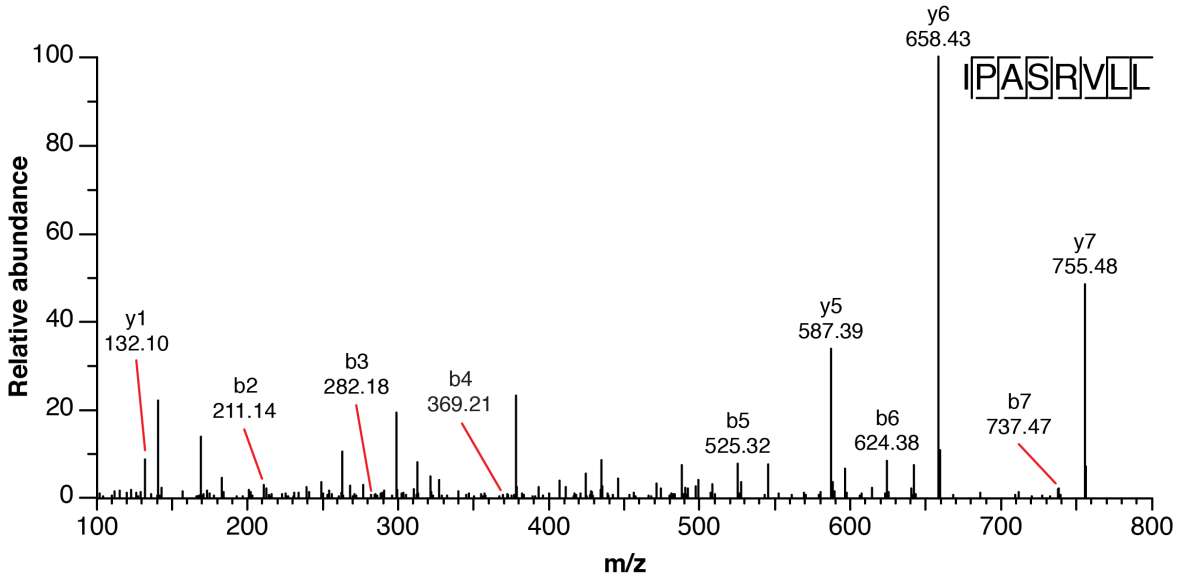


Synthetic peptide

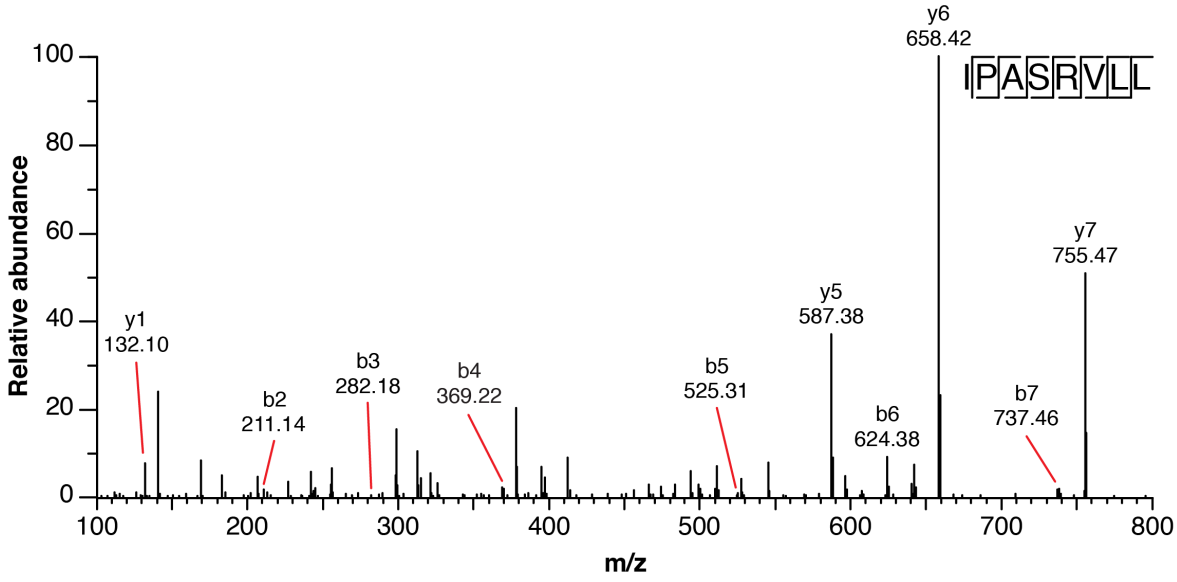


I

Endogenous peptide

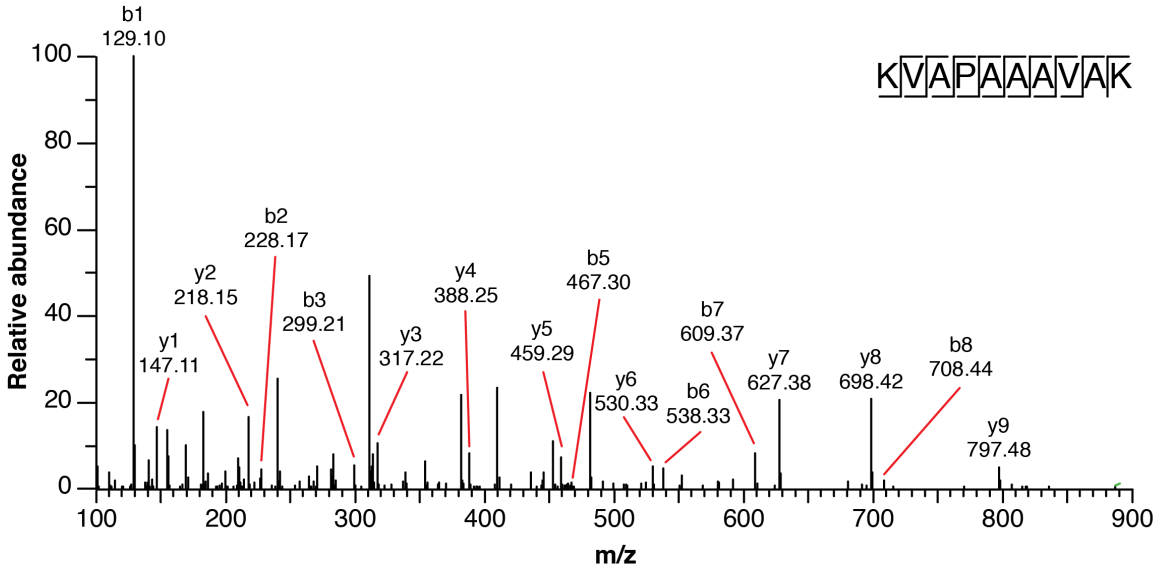


Synthetic peptide

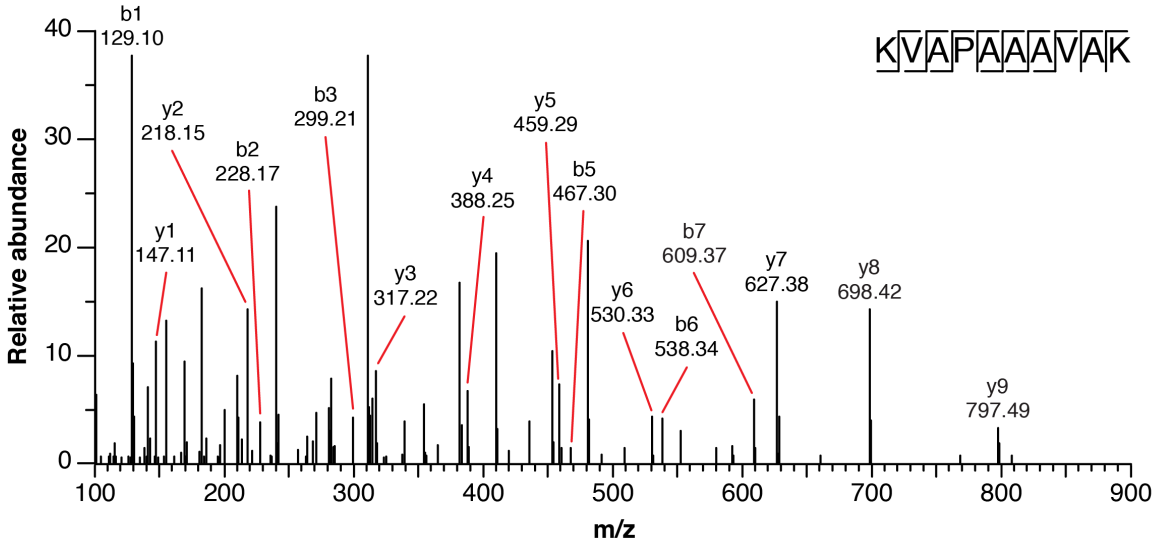


m

Endogenous peptide

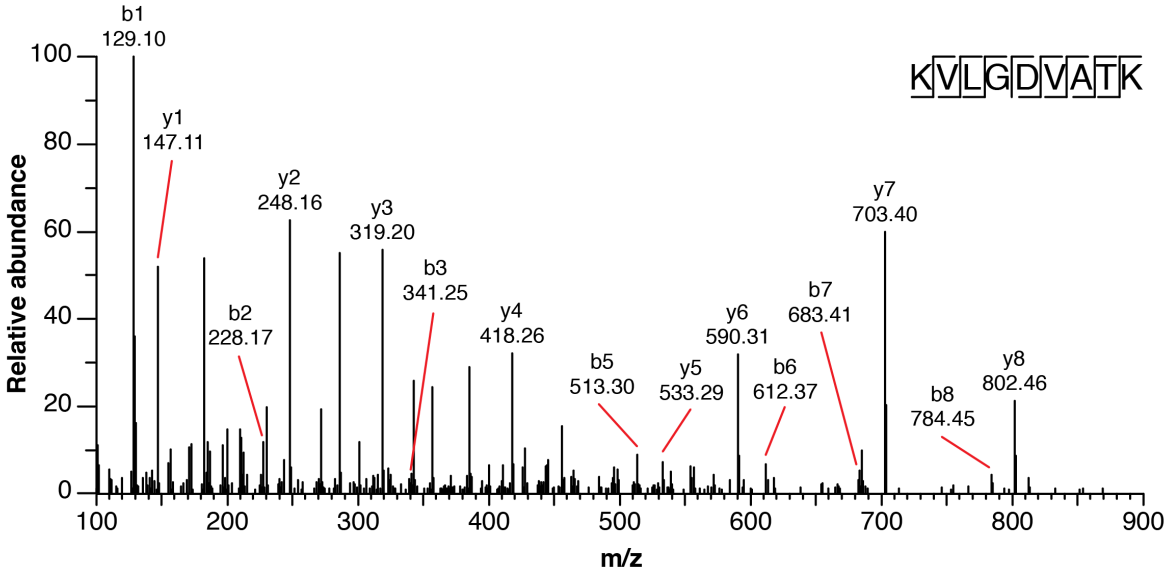


Synthetic peptide

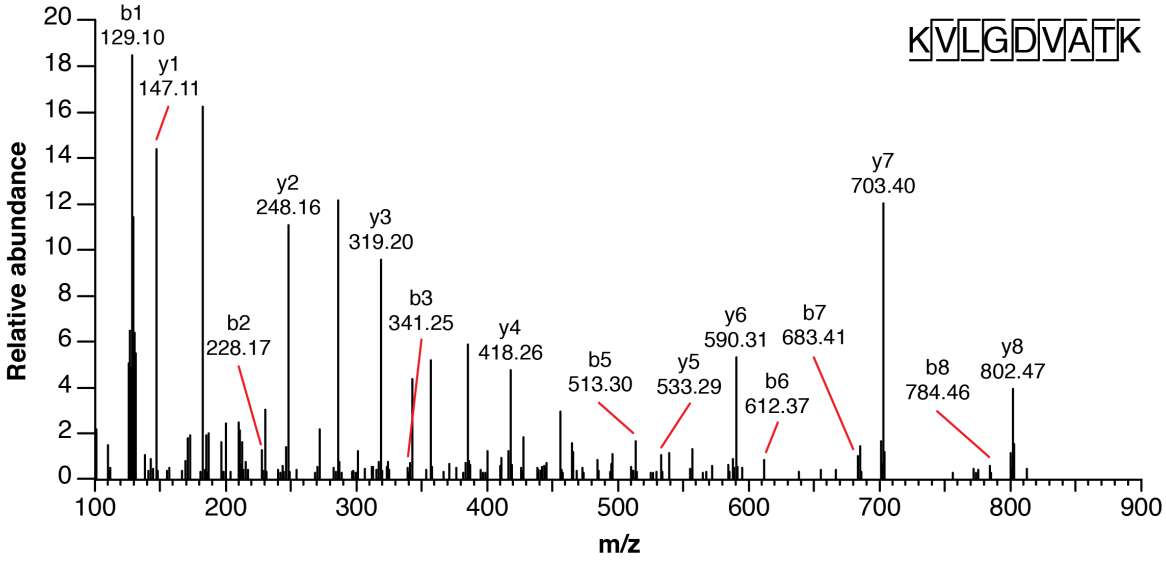


n

Endogenous peptide

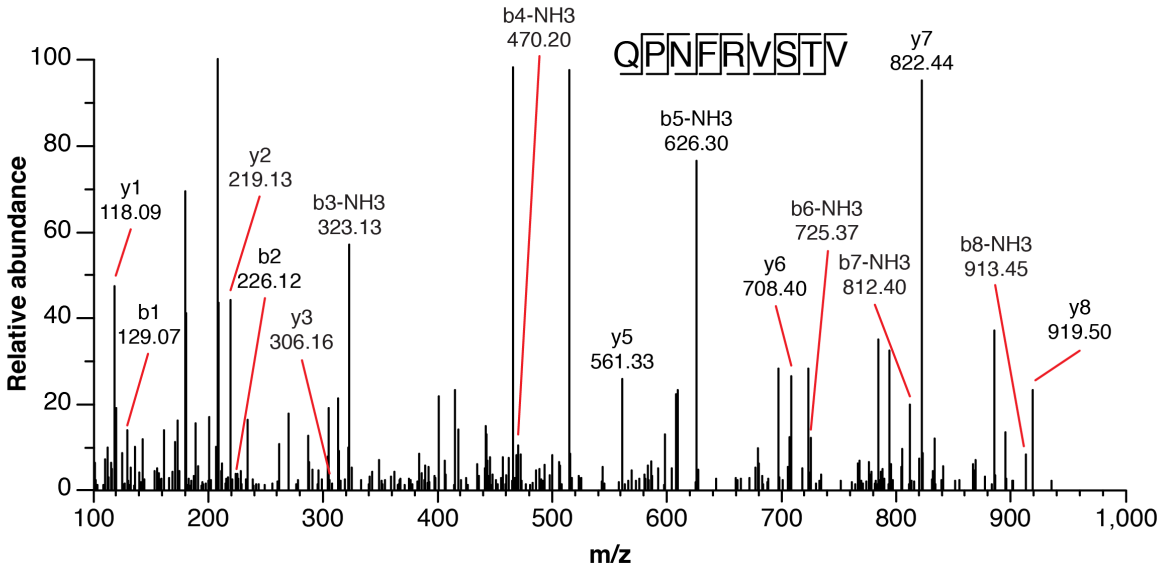


Synthetic peptide

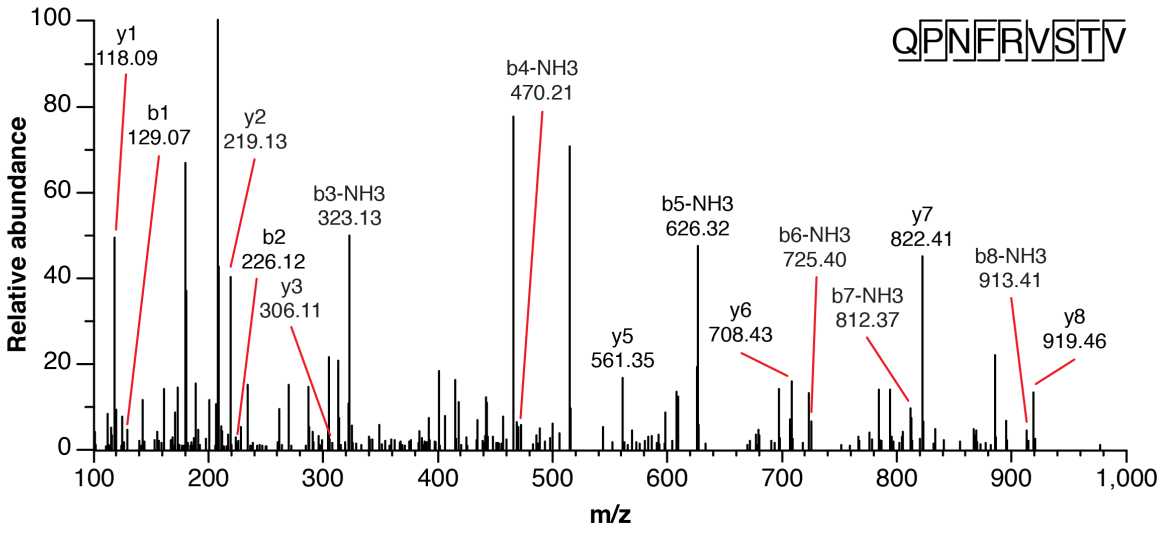


o

Endogenous peptide

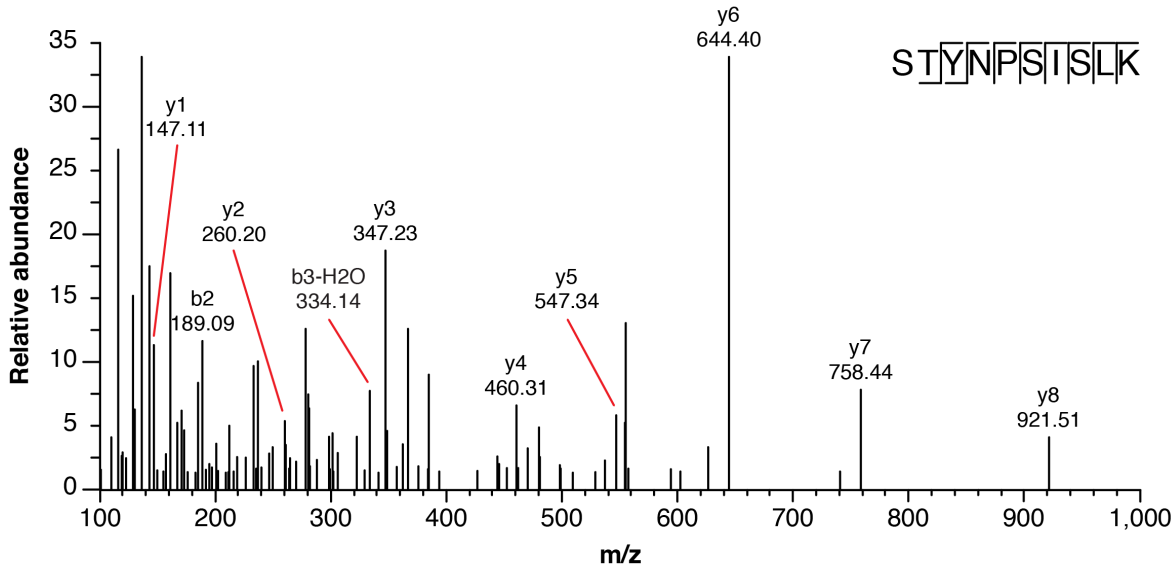


Synthetic peptide

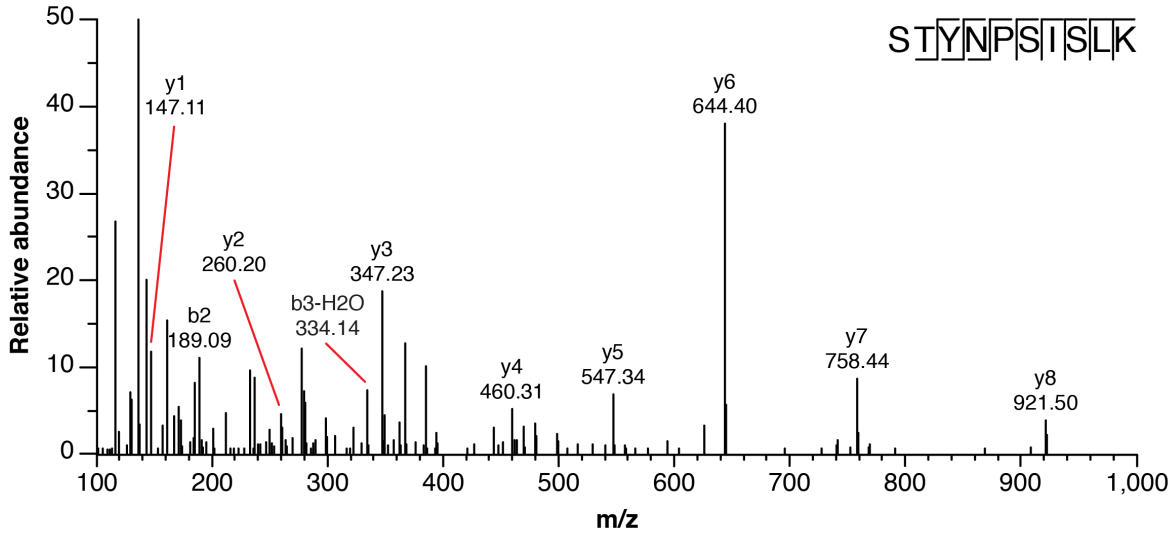


p

Endogenous peptide

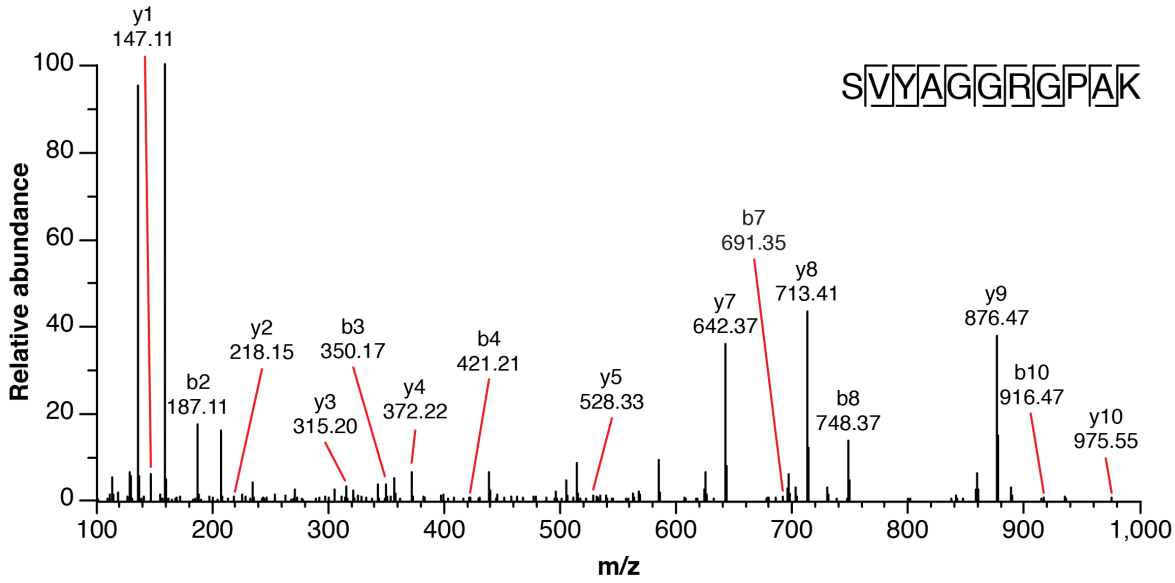


Synthetic peptide

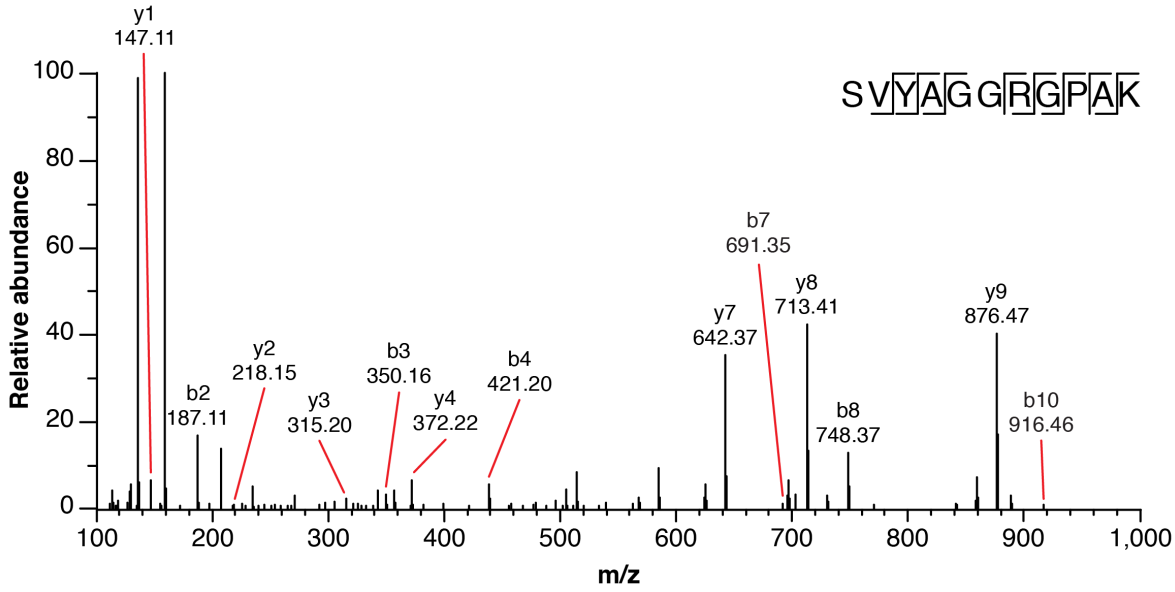


q

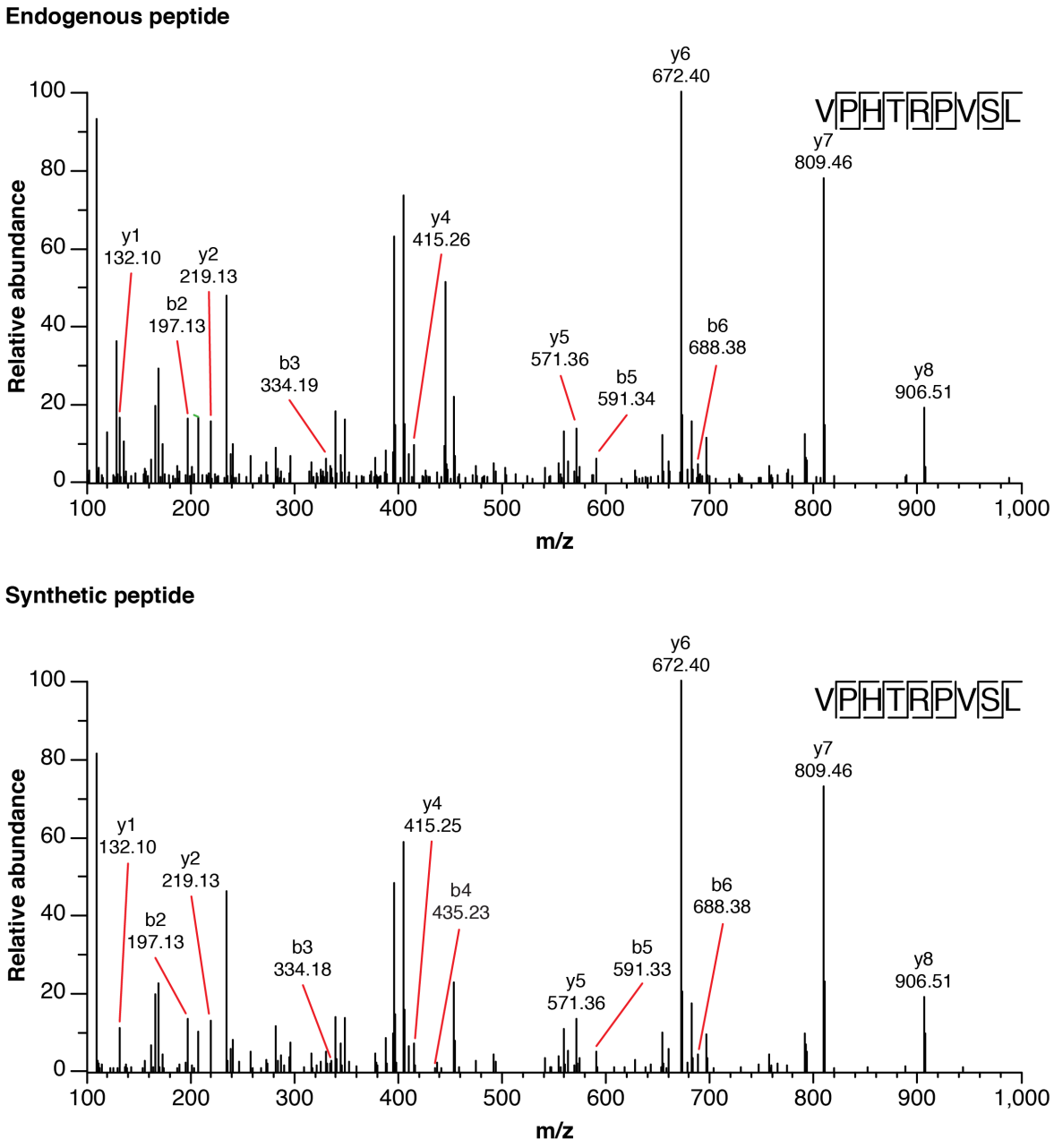
Endogenous peptide



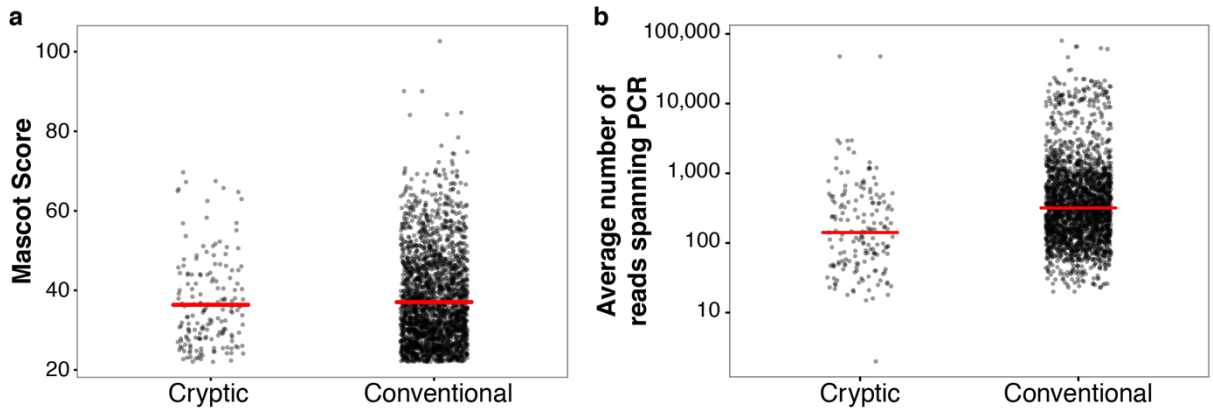
Synthetic peptide



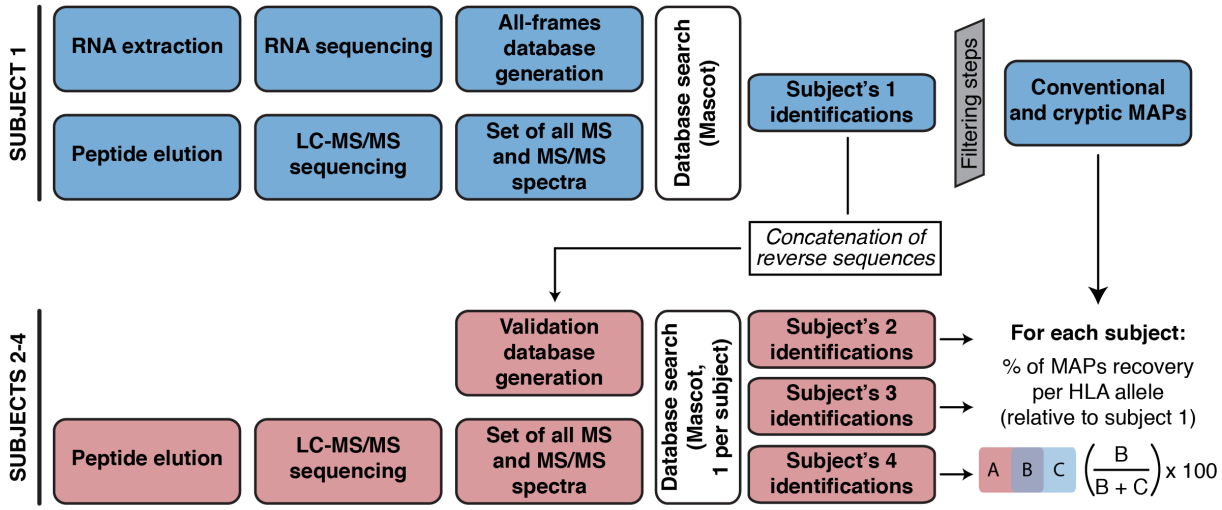
r



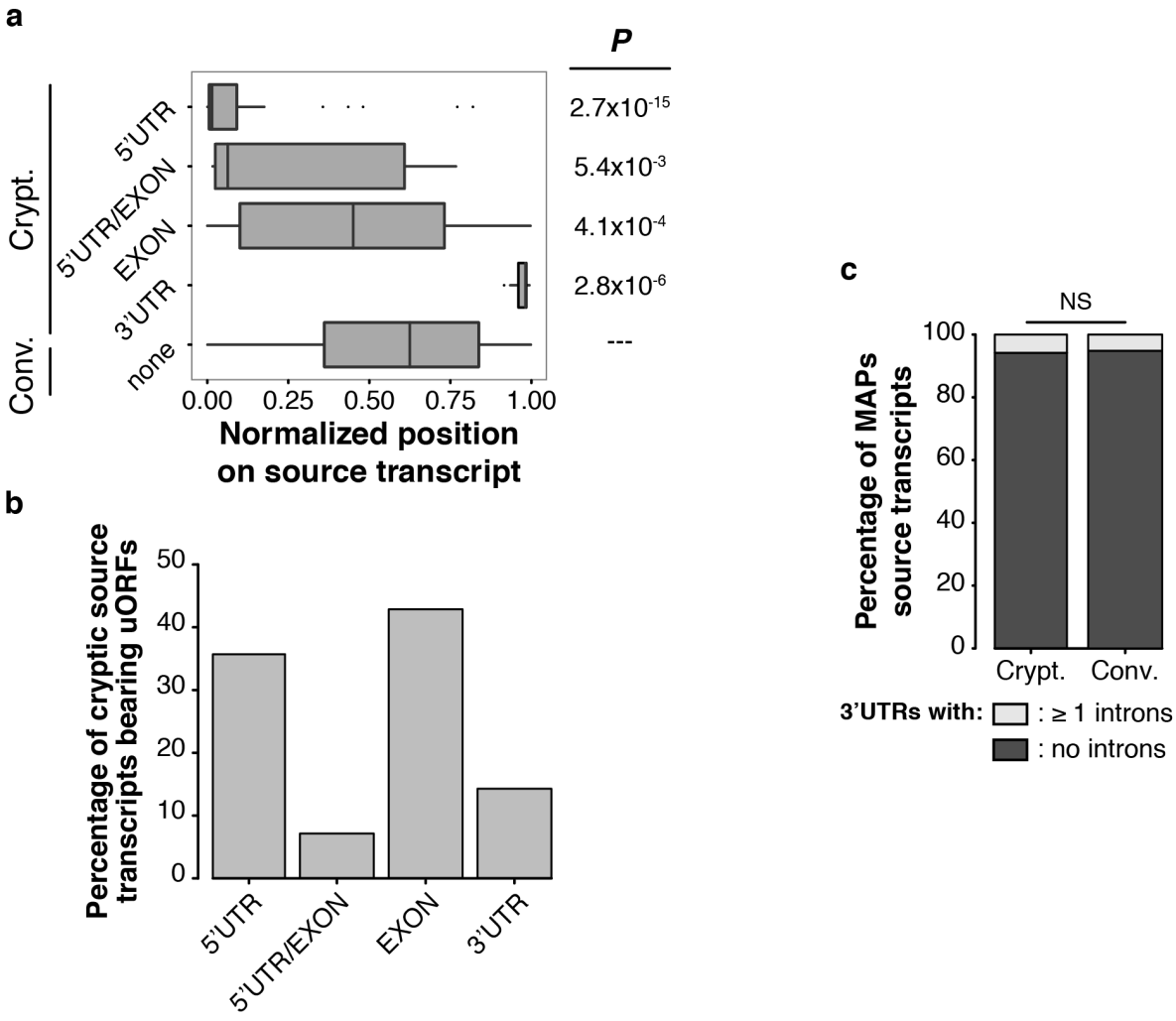
Supplementary Figure 2.2 | MS validation of 18 cryptic MAPs. Among the 168 cryptic MAPs identified in our study, 18 were randomly selected and subjected to MS validation using synthetic version of them. (**a-d**) Four cryptic MAPs related to **Figure 2.7**. (**e-r**) 14 other cryptic MAPs.



Supplementary Figure 2.3 | Cryptic and conventional MAPs are similarly detected by mass spectrometry and RNA-seq. (**a**) Cryptic and conventional MAPs display similar Mascot score distributions. Dot plot representing the Mascot score distribution for cryptic and conventional MAPs, with the Mascot score being an indicator for the goodness of a peptide-spectrum match. (**b**) Cryptic and conventional MAPs derive from regions covered by our RNA-seq experiment. Dot plot showing the average number of reads spanning the peptide-coding regions (PCR) of cryptic and conventional MAPs obtained following TopHat mapping. In both panels, red lines depict the median of each group for the considered metrics.



Supplementary Figure 2.4 | Cryptic MAPs validation workflow. Detailed workflow of the analysis presented in **Figure 2.2a**. We performed peptide elution on B-LCLs from three other subjects (subjects 2-4) that shared 4, 2 or no HLA allele with subject 1, respectively. Since we only wanted to validate MAPs from subject 1 rather than explore the whole immunopeptidome of subjects 2-4, we constructed a validation database that contained all identifications made in subject 1 concatenated with their reverse sequences. Identifications made in subjects 2-4 using Mascot were compared to the list of conventional and cryptic MAPs from subject 1 allowing us to compute a percentage of MAPs recovery per HLA allele relative to subject 1 for each subject.



Supplementary Figure 2.5 | Transcripts source of cryptic MAPs appear less stable than transcripts source of conventional MAPs. (**a**) Exonic cryptic PCRs are shifted towards the 5'end of source transcripts. Box plot depicting the normalized position in source transcript of PCRs for conventional and cryptic MAPs. For cryptic MAP PCRs, distinct box plots are depicted as a function of PCR genomic location. The two-sided Wilcoxon rank sum test was used to compare the location of conventional MAP PCRs to that of PCRs for various types of cryptic MAPs. Resulting *P*-values are indicated at the right of each box. (**b**) Transcripts containing uORFs do not solely generate 5'UTR and 5'UTR/EXON cryptic MAPs. Bar plot representing the percentage of each type of cryptic MAPs generated by transcripts bearing at least one uORF. (**c**) Transcripts source of conventional and cryptic MAPs contain the same proportion of 3'UTR intronic sequences. Bar plot depicting the percentage of transcripts containing no (dark gray) or at least one (light gray) intronic sequence in their 3'UTR for both conventional and cryptic MAPs. Statistical significance was assessed using a two-sided Fisher's exact test (NS: not significant, $P = 6.53 \times 10^{-1}$).

2.11.2 Supplementary Tables

Supplementary Table 2.1 | HLA allotypes presented by subject 1–4

	A*01:01	A*02:01	A*03:01	A*29:02	B*08:01	B*18:01	B*39:24	B*44:03	B*57:01	Number of shared HLA with Subject 1
Subject 1			x	x	x			x		
Subject 2			x	x	x			x		4
Subject 3		x		x				x	x	2
Subject 4	x	x				x	x			0

Supplementary Table 2.2 | Rare codon usage in conventional vs. cryptic MAP source transcripts or ORFs

	Conventional MAP source transcripts	Cryptic MAP source ORFs
Rare codons	464,739	3,030
Common codons	767,785	5,119

Supplementary Table 2.3 | Rare codon usage in MAP vs. non-MAP source transcripts or ORFs

	MAP source transcripts or ORFs	Non-MAP source transcripts
Rare codons	467,769	12,118,695
Common codons	772,904	22,909,893

CHAPTER 3

3 Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy

Céline M. Laumont[1,2] • Claude Perreault[1,2,3]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, Station Centre-Ville, PO Box 6128, Montreal, QC H3C 3J7, Canada

[2]Department of Medicine, Faculty of Medicine, Université de Montréal, Station Centre-Ville, PO Box 6128, Montreal, QC H3C 3J7, Canada

[3]Division of Hematology, Hôpital Maisonneuve-Rosemont, 5415 de l'Assomption Boulevard, Montreal, QC H1T 2M4, Canada

Correspondence to:
Claude Perreault
claude.perreault@umontreal.ca

Cellular and Molecular Life Sciences, Volume 75(4), pages 607-621 (February 1, 2018)

3.1 Context

Reflecting on a sentence written by Aristotle in his Metaphysics (Book VIII, part 6, translation of W.D. Ross) stating that ‘the whole is something beside the parts’, often mentioned as ‘the whole is more/greater than the sum of its parts’, we thought that an integrative approach could help us enlarge our knowledge on the biogenesis of cryptic MAPs (presented in **Chapter 2**).

To do so, we decided to gather previous and current findings made on cryptic proteins and MAPs (the parts) to make a coherent review (the whole). Because the observations we gathered were obtained from different fields (systems biology, basic immunology or mechanistic of translation) that do not apply the same reasoning techniques (exploratory vs. hypothesis-driven) nor the same technologies (-omic vs. reductionist approaches), this exercise really enriched our views on cryptic MAPs. As a result, we present a very comprehensive overview of the molecular mechanisms that have been (or might be) involved in the generation of cryptic MAPs and then try to reflect on the importance of such cryptic MAPs for cancer immunotherapy.

3.2 Authors' contributions

Céline M. Laumont: prepared all figures and wrote the first draft of the manuscript.

Claude Perreault: general discussion and contributed to the writing of the manuscript.

All authors edited and approved the final version of the manuscript.

3.3 Abstract

Cryptic MHC I-associated peptides (MAPs) are produced via two mechanisms: translation of protein-coding genes in non-canonical reading frames and translation of allegedly non-coding sequences. In general, cryptic MAPs are coded by relatively short open reading frames whose translation can be regulated at the level of initiation, elongation or termination. In contrast to conventional MAPs, the processing of cryptic MAPs is frequently proteasome independent. The existence of cryptic MAPs derived from allegedly non-coding regions enlarges the scope of CD8 T cell immunosurveillance from a mere ~2% to as much as ~75% of the human genome. Considering that 99% of cancer-specific mutations are located in those allegedly non-coding regions, cryptic MAPs could furthermore represent a particularly rich source of tumor-specific antigens. However, extensive proteogenomic analyses will be required to determine the breath as well as the temporal and spatial plasticity of the cryptic MAP repertoire in normal and neoplastic cells.

3.4 Introduction

In vertebrates, all nucleated cells present major histocompatibility complex class I (MHC I) molecules in complex with 8–11-amino acid-long peptides[1], which will be referred to as MHC I-associated peptides (MAPs). In the thymus, a crucial role of MAPs is to initiate the establishment of central tolerance[2], a process whereby classic CD8 thymocytes with a strong affinity for self MAPs are eliminated. Central tolerance is orchestrated by medullary thymic epithelial cells (mTECs) which express more genes than any other cell type and have, therefore, a particularly broad MAP repertoire[3]. The CD8 thymocytes which are not self-reactive migrate to extrathymic tissues and organs[4,5], where they are involved in immunosurveillance[6]. Immunosurveillance refers to the ability of self-tolerant CD8 T cells to eliminate cells presenting non-self MAPs or neo-self MAPs derived from microbial or cancer-specific proteins, respectively.

The fact that MAPs are such central players in the development and function of our adaptive immune system raises the question of their biogenesis, i.e., what is the molecular definition of the immune self? In an attempt to answer this question, we first present the basics of the MHC I antigen presentation pathway and the generation of conventional MAPs, which derive from translation of protein-coding genes in their canonical reading frame. We then present an overview of the molecular mechanisms responsible for the generation of cryptic MAPs, which derive from two processes: (1) translation of protein- coding genes in non-canonical reading frames and (2) translation of allegedly non-coding sequences. Finally, we discuss the relevance of such cryptic MAPs in cancer immunotherapy.

3.5 The art of sampling: the MHC I antigen presentation pathway

Most MAP precursor peptides derive from the degradation of cytosolic proteins by the proteasome and other cytosolic aminopeptidases[7]. Once translocated into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP)[8], the N-terminal end of these longer precursors is trimmed by ER aminopeptidase 1 and 2[9] so that they reach an optimal length, between 8 and 11 amino acids, for loading onto MHC I molecules[10]. Then, tapasin[11], a key component of the peptide- loading complex, along with its homolog the TAP-binding protein-related[12], samples this pool of MAPs to select those forming the most stable peptide–MHC I complexes[13]. Finally, MHC I molecules, loaded with a MAP, are exported to the cell surface through the secretory pathway[14]. On a per cell basis, the MHC I antigen presentation pathway could be described as a dimensionality reduction function that compresses a large input dataset, the proteome, into a smaller output, the MAP repertoire or immunopeptidome, while keeping as much information as possible. In fact, knowing that an average cell expresses ~10^4 proteins with a median length of 449 amino acids[15,16], the challenge is to comprehensively represent this proteome with a MAP repertoire composed of ~10^4 MAPs at an average length of 9 amino acids displayed on ~10^5 MHC I molecules[1,17], and that represents about 2% of proteome. Because the MAP repertoire contains only a small fraction of the proteome, it is important to understand the mechanisms of MAP biogenesis and to determine whether MAPs originate from peculiar genomic regions.

3.5.1 The crucial role of HLA polymorphisms

In humans, classical MHC I allotypes are encoded by three loci, the human leukocyte antigen (HLA) A, B and C, all located on the short arm of chromosome 6 (6p21). These genes are the most polymorphic genes of the human genome as they totalize about 12,351 alleles according to the IPD-IMGT/HLA database (last consulted on 2017/06/18)[18]. Since most of these polymorphisms affect residues in the peptide-binding groove of MHC I molecules, they can modify the number and the location of

peptide-binding pockets that will in turn limit the pool of potential MAPs[19]. Indeed, these pockets impose strict amino acid preferences along the MAP sequence and thereby define a unique binding motif for each HLA allotype[20]. When Granados et al. used mass spectrometry (MS) to sequence the MAP repertoire of HLA-genotyped subjects they found only 0.4% overlap between the immunopeptidome of HLA-unmatched subjects[21]. By contrast, HLA-identical siblings displayed almost identical MAP repertoires. Because they present different repertoires of both self and non-self MAPs, MHC I allotypes are often associated with susceptibility to various infectious and autoimmune diseases[22]. For instance, several MHC I molecules have been associated with HIV progression or control in seropositive patients[23]. In particular, HLA-B*27 and -B*57 have been shown to correlate with a decreased viral load, probably because of their ability to present Gag-specific MAPs[24,25]. To discover MHC I molecules predisposing for autoimmune disorders, multiple genome-wide association scans of non-synonymous single nucleotide polymorphisms have been performed on cohorts of thousands of patients and controls. For ankylosing spondylitis (AS), a heritable form of inflammatory arthritis, HLA-B*27 was identified as the major genetic risk factor along with several other HLA-B allotypes. The non-synonymous single nucleotide polymorphism most associated with AS susceptibility affects a residue located in the C-terminal binding pocket of HLA-B molecules and has the potential to modify their peptide-binding capacity[26]. This observation suggests that, as for HIV susceptibility, MAP repertoires presented by specific MHC I molecules can facilitate the development of various diseases including AS. Altogether these observations show that the HLA genotype has a dominant influence on the landscape of MAPs presented by any individual.

3.5.2 The DRiPs hypothesis and its implications

The MAP repertoire is very plastic as it can be influenced by cell type, cell metabolism, drugs and infections[27-30]. New MAPs can be produced exceedingly fast. Accordingly, CD8 T cells can destroy cells within 1 h following infection[31], meaning that infected cells already present viral MAPs[32], despite the fact that viral proteins are expected to be as stable as cellular proteins (median half-life of 46 h)[33]. To explain the

swift presentation of viral MAPs on infected cells, Yewdell et al. provided compelling evidence that most MAPs originate from a pool of rapidly degraded proteins that includes a large proportion of defective ribosomal products (DRiPs)[34]. DRiPs derive from the accumulation of “errors” during mRNA translation, thereby leading to the production of truncated or misfolded protein products that preferentially enter the MHC I antigen presentation pathway[35,36]. The reason why short-lived proteins in general and DRiPs in particular are preferential MAP substrates remains a matter of debate.

An interesting corollary to the DRiP hypothesis is that MAP generation should be tightly linked to both the amount and the accuracy of translation[37,38]. Indeed, a recent analysis of ~25,000 MAPs isolated from 18 B-lymphoblastoid cell lines (B-LCLs) confirmed that MAP source transcripts were expressed at higher levels than transcripts that did not generate MAPs[21,39]. In addition, MAP source transcripts had features known to improve translation efficiency such as an increased number of exons[40] and shorter 5’UTRs with fewer upstream open reading frames (ORFs)[41]. Furthermore, proteins produced by MAP source transcripts had a higher degradation potential than non-source transcripts as their sequence was enriched in ubiquitination sites, degradation motifs and disordered regions, which are all known to favor protein degradation[42-44]. In addition, mRNA targeting by microRNAs and short hairpin RNAs as well as the degradation of faulty mRNAs through the nonsense-mediated decay pathway has been positively linked to MAP generation[21,45,46], probably because these translation-coupled mRNA destabilization mechanisms increase the DRiP rate of nascent proteins.

3.6 Unconventional proteins and MAPs

3.6.1 The dark matter in the proteome

Irrespective of whether they originate from DRiPs or not[47], MAPs have been assumed to originate from the translation of protein-coding transcripts in their canonical reading frame (**Figure 3.1 and 3.2a**), i.e., the sequence starting and ending at the transcript's primary AUG and stop codon, respectively. However, it is important to realize that mass spectrometrists can only find what they are looking for. The traditional aphorism "absence of evidence is not evidence of absence" has, therefore, to be kept in mind.

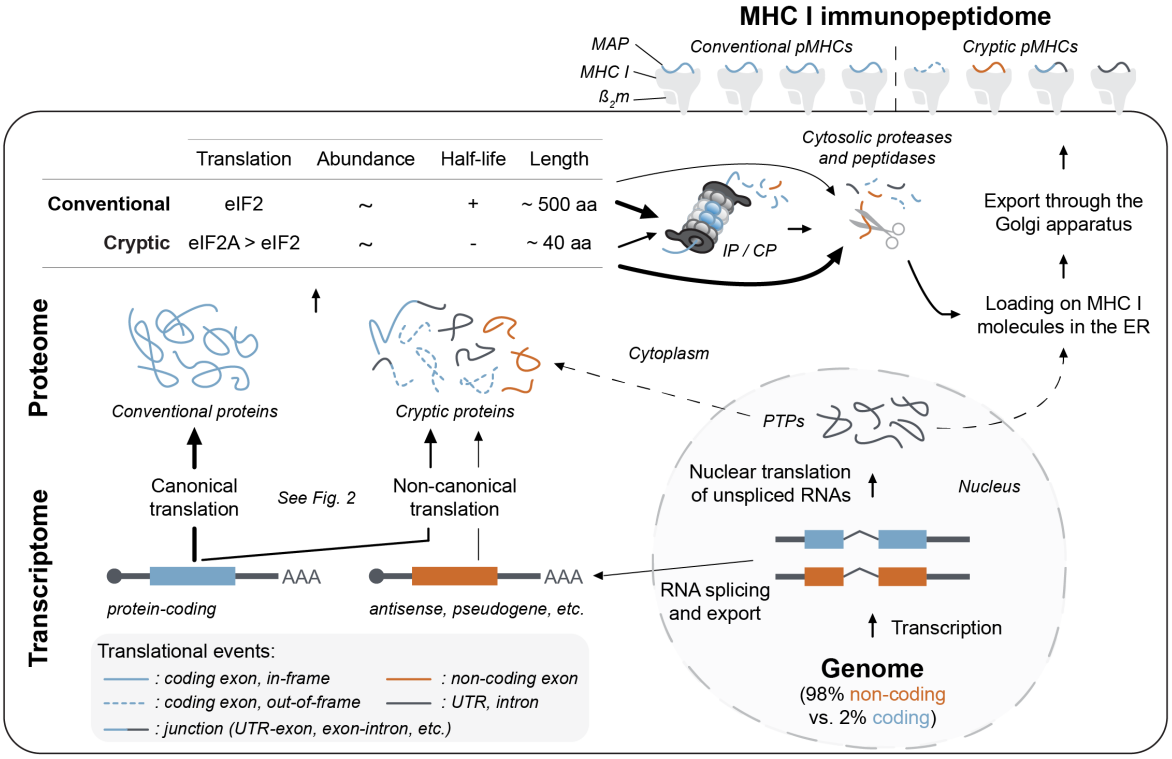


Figure 3.1 | Key features of conventional and cryptic MAPs biogenesis. *Thickness of the arrows* represents the relative contribution of considered sources while *dotted arrows* only represent suspected interactions. Peptide processing and loading in the ER as well as subsequent export steps of peptide–MHC complexes are not detailed here. *aa* amino acid, *CP* constitutive proteasome, *ER* endoplasmic reticulum, *IP* immunoproteasome, *MAP* MHC I-associated peptide, *MHC I* major histocompatibility complex class I, *pMHC* peptide–MHC I complex, *PTP* pioneer translation product, *UTR* untranslated region

Although protein-coding genes represent only 2% of the genome, in-depth transcriptomic analyses performed by the ENCODE consortium on 15 human cell lines of various origin revealed that, cumulatively, processed and primary transcripts cover ~62 and ~75% of the genome, respectively[48]. One could expect that this increased complexity would be restricted to the transcriptome, as only some transcripts are expected to be translated. However, analyses of ribosomal profiling data have revealed that ribosomes occupy both classical protein-coding and non-coding transcripts[49]. Moreover, ribosomes bound to protein-coding transcripts appear to translate them (1) in their canonical reading frame and (2) in their upstream or exonic non-canonical reading frames[50,51]. The existence of such atypical translation products was further confirmed by two large-scale MS analyses of the human proteome[52,53]. Focusing on non-canonical reading frames containing at least 40 codons, Vanderperre et al. identified a total of 1259 new proteins by MS in various human cell lines and tissues, where they appeared to represent ~2% of the proteome[54]. This is most likely an underestimate because, compared to conventional proteins, these atypical proteins are less likely to be caught by MS due to (1) their short size (median length of 57 vs. 449 amino acids), which makes them generate less tryptic peptides, and (2) their low molecular weight (<10 kDa) since most proteomic studies focus on larger proteins (≥10 kDa)[55]. Finally, another study used isotope dilution MS to quantify three atypical protein products and showed that their abundance ranged from 10 to 2000 copies per cell[56], well in the range observed for conventional proteins[15,33]. Altogether these studies provide clear evidence that the proteome is far more complex than we previously thought, as conventional proteins are not its sole component. Irrespective of whether these new protein products are functional[57,58] or not[59], if they generate MAPs, these MAPs should be as immunogenic as conventional MAPs.

3.6.2 Cryptic MAPs, from an odd observation to system levels analysis

In 1989, long before mass spectrometrists started to explore the dark matter in the proteome, Boon et al. made the seminal observation that cells transfected with tumor DNA fragments, lacking both transcriptional and translational regulatory elements, were nevertheless able to stimulate CD8 T cells, thereby suggesting that they were triggering the production of immunogenic MAPs[60]. To explain this odd observation, Boon and Van Pel proposed the "pepton hypothesis"[61]. The crux of this hypothesis was that, rather than deriving from the translation of protein-coding genes in their canonical reading frames, MAPs derived from the transcription and translation of short subgenic regions performed by dedicated RNA polymerases and ribosomes. As a follow-up, other groups modified the nucleotide context of known immunogenic MAPs and analyzed their production using MAP-specific T cell clones. In 1991, inserting a frameshift mutation that should have blocked the production of the two known immunogenic MAPs generated by the influenza nucleoprotein in mice, Fetten et al. demonstrated that these MAPs were still presented at levels sufficient to induce CD8 T cell activation, although the production of the full-length nucleoprotein and its associated polypeptides were undetectable[62]. In line with this, Shastri et al. placed the SIINFEKL-coding region out-of-frame with regard to the upstream AUG start codon and showed that cells transfected with such construct still produced SIINFEKL at level detectable by T cells[63]. In both mice and humans, several reports then showed that the generation of those unexpected MAPs was not restricted to model systems since translation of 5'UTRs[64], exon–intron junctions[65], intronic regions[66-68], non-canonical reading frames located in known exons[69-79] or not[80] and even antisense transcripts[81] could all generate tumor and viral antigens detected by T cells. Since they were always identified one at a time and in single studies, doubt was still permitted with regard to the physiological relevance of those atypical MAPs that we will hereafter refer to as cryptic MAPs. To demonstrate their importance, Shastri's team took advantage of a transgenic mouse model engineered to ubiquitously express a bicistronic transgene encoding a Uty- and a H60-derived MAP in a canonical and a non-canonical reading

frame, respectively. Although less efficiently presented than the conventional Uty-derived MAP, the cryptic H60- derived MAP was able to (1) induce central tolerance and (2) prime CD8 T cell responses in vivo, thereby proving the relevance of such cryptic antigens[82]. These observations were further indirectly confirmed in humans[83] and even extended, as cryptic MAPs were shown to elicit CD8 T cell responses strong enough to promote tumor regression in humans[69,72,76,84] or, in some cases, cause autoimmunity[85]. Finally, using the six-frame translation of RNA-sequencing reads to probe the conventional and cryptic MAP repertoire of B-LCLs, our group demonstrated that cryptic MAPs were more than just mere exceptions as they represent at least ~10% of the MAP repertoire. As prefigured by early findings, the cryptic MAPs that we identified originated from a large variety of non-canonical translation events: from translation of intergenic and non-coding regions, such as antisense loci, to translation of coding exons in non-canonical reading frames along with the translation of 5' and 3'UTRs, introns and various types of junctions (UTR–exon or exon–intron)[86]. Altogether these data show that the non-canonical (i.e., cryptic) proteome contributes to the MAP repertoire through the generation of cryptic MAPs (**Figure 3.1**). This discovery is of uttermost importance to the adaptive immune system, because it enlarges the scope of CD8 T cell immunosurveillance from a mere ~2% to as much as ~75% of the human genome[48].

3.7 The sinuous tale of cryptic MAPs' origin

The fact that, a priori, cryptic MAPs can derive from any transcribed sequence, regardless of its official coding status, suggests that the pepton hypothesis[61], which assumed that MAPs derived from the transcription and translation of subgenic regions, was correct. Since inserting a stop codon right upstream of a non-canonical reading frame encoding the SIINFEKL model MAP completely abrogated its presentation[63], we can infer that biogenesis of cryptic MAPs is regulated at the translational level. To review the translational mechanisms at play, we grouped them according to the translation phase to which they belong, e.g., initiation, elongation or termination. Moreover, we provide a brief overview of the post-translational processing of cryptic MAPs.

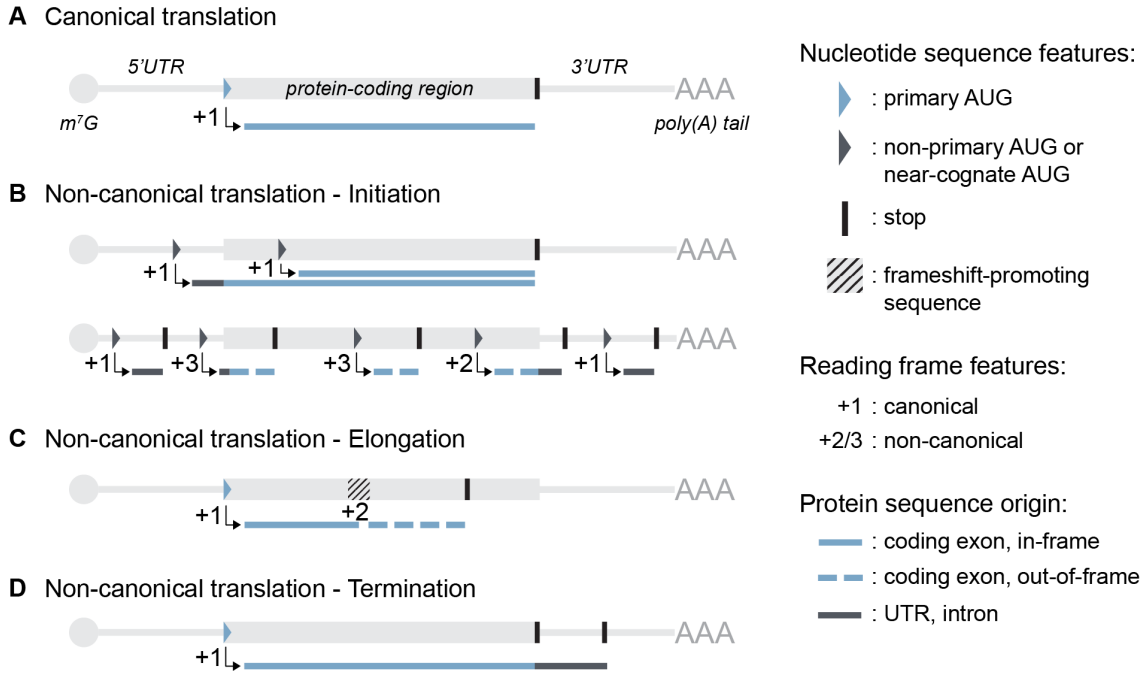


Figure 3.2 | Translational events involved in the generation of conventional and cryptic MAPs. (**a**) Schematic view of the canonical translation event that can lead to the generation of conventional MAPs. (**b-d**) Schematic representation of non-canonical initiation (**b**), elongation (**c**) or termination (**d**) translational events that can lead to the generation of cryptic MAPs.

3.7.1 Initiation: it is all about knowing where to start

In eukaryotes, translation initiation is a multistep process which requires several eukaryotic initiation factors (eIFs) and can be divided into three main phases[87]. First, formation of the 43S pre-initiation complex composed of the small 40S ribosome subunit and specific eIFs including the eIF2–ternary complex, i.e., the GTP-bound state of the eIF2 protein linked to a methionine-loaded initiator tRNA (Met-tRNA$_i^{Met}$). Then, attachment of the 43S pre-initiation complex on the 5' end of an mRNA with the help of eIF4F and poly(A)-binding proteins is followed by a 5'–3' mRNA scanning that settles the pre-initiation complex on the first AUG in an optimum Kozak context[88]. Finally, recognition of the start codon by the anticodon of the Met-tRNA$_i^{Met}$ promotes the release of eIF2 by hydrolysis of its GTP to a GDP, which ultimately results in the formation of the 80S initiation complex, containing both the 40S and 60S ribosome subunit, and that is now ready to translate the downstream protein-coding region. As regulated as it seems, we know for a fact that translation does not always start at the primary AUG, which produces the conventional protein product (**Figure 3.2a**). Start codons can be non-primary AUGs[85,89,90] and near-cognate start codons[91-93], which are codons differing from AUG by a single base such as CUG, UUG, and GUG. When located upstream or downstream of the primary AUG and leading to the production of N-terminal extended or truncated versions of the conventional protein, non-canonical translation start sites have a limited potential to generate cryptic MAPs (**Figure 3.2b**, upper panel). However, when initiation at such codons promotes the translation of non-canonical reading frames that are (1) fully upstream or downstream the protein-coding region or (2) fully or partially overlapping the protein-coding region but out-of-frame with regard to it, they are then likely to generate cryptic MAPs (**Figure 3.2b**, lower panel). This phenomenon is probably common because (1) 49% of human transcripts are predicted to bear at least one ORF in their 5'UTR[41], (2) each human transcript bears on average ~4 alternative ORFs anywhere along its sequence[54] and (3) nearly half of the genes expressed in HEK293 cells were shown to contain multiple translation initiation start sites both upstream and downstream of their primary AUG[94].

Several non-mutually exclusive mechanisms have been linked to those non-canonical translation initiation events producing cryptic MAPs such as start codon scan-through[95], in which ribosomes miss the primary AUG to initiate translation further downstream, and translation re-initiation[86], caused by the non-dissociation of the 43S pre-initiation complex following a first upstream translation event. On a more molecular note, translation initiation at the near-cognate codon CUG, instead of the classical AUG, has been shown to rely on both the expected wobble pairing between CUG and the anticodon of a Met-tRNA$_i^{Met}$ as well as the unexpected pairing between this same CUG and the CAG anticodon of a Leu-tRNA[96,97]. Upon certain stimuli such as viral infection or ER-stress, the integrated stress response shuts down translation by phosphorylation of eIF2 on its alpha subunit, which prevents further association with the Met-tRNA$_i^{Met}$, thereby blocking AUG initiation[98]. However, CUG and even UUG translation initiation was shown to be unaffected by the integrated stress response thanks to its eIF2A dependency, and could, therefore, help maintain the translation of some stress-privileged transcripts[99]. Now, although CUG appears to be the most commonly used near-cognate AUG[94], other near-cognate and even non-cognate start codons have been proven relevant to initiate translation in various systems[56,100,101] and even produce cryptic MAPs[86,102]. Whether this happens through the classical eIF2-dependent translation initiation or through its less classical eIF2A-dependent version is still under debate and could be explored using the eIF2A knock-out mouse[103]. Nevertheless, answering this question will clearly help us understand how various stimuli can affect the proteome and, therefore, its cell surface representation, the immunopeptidome. Finally, besides the cap-dependent translation mechanisms discussed above, cap-independent translation at internal ribosome entry site has also been involved in cryptic MAP generation, as demonstrated by the production of two immunogenic melanoma antigens from the long non-coding and polycistronic meloe transcript[104], while the cap-dependent translation of this very same transcript only produced a tolerized epitope[83].

3.7.2 Elongation: saying two things at once

During translation elongation, the ribosome moves along the mRNA three nucleotides at a time (codon by codon) to faithfully translate and produce the encoded protein product. At first, the ribosome's P site is filled by the initiation codon paired with its complementary initiator tRNA while the GTP-bound eukaryotic elongation factor 1A delivers the codon-complementary aminoacylated-tRNA to the ribosome's A site. Following formation of the peptide bond between the first and second amino acids, the elongation factor 2 catalyzes the 3-nucleotide translocation of the ribosome onto the next codon, thereby placing the deacetylated- and peptidyl-tRNA in the E and P sites of the ribosome, respectively. The newly emptied A site is now ready to receive the next aminoacylated-tRNA bearing an anticodon complementary to the codon it sits on for the elongation cycle to resume until a stop codon is met[105]. Although the initiation step defines the reading frame of translating ribosomes, it is not as fixed as one could imagine. In fact, 'slippery' nucleotide sequences followed by stem–loop secondary structures have been shown to promote programmed ribosomal frameshifting (PRF), i.e., translocation of the ribosome one nucleotide forward or backward instead of three[106]. This controlled ribosome slippage changes the initiation-defined reading frame during elongation, thereby leading to the production of a chimeric protein product (**Figure 3.2c**). For a long time PRF has been considered as an economical way for viruses to encode more proteins in their small genomes. In line with this, placing HIV or HSV frameshift-promoting sequences upstream of a region encoding a murine model peptide allows T cell recognition even when a +1 ribosomal frameshift is required for MAP production[107]. However, PRF is not a virus-specific characteristic since a preliminary estimate suggests that it could affect ~10% of genes in yeast[108]. In humans, the estimate, which is likely to be an underestimation, is lower with ~1% of genes shown to be translated in two frames[109], a dual decoding phenomenon that can be partly explained by PRF. In line with this, PRF has been involved in the production of at least one cryptic MAP[77] and a handful of proteins in mammalian cells[110-113]. Besides its capacity to produce chimeric proteins, PRF might also be a translation-coupled post-transcriptional gene regulation mechanism since it was shown to interact with

microRNAs as a means to target mRNAs to the nonsense-mediated decay pathway[114]. Notably, microRNAs and nonsense-mediated decay pathway are both already linked to MAP generation[21,38,46]. PRF-triggered mRNA degradation could help explain some of the cryptic MAPs that we identified and that did not present with a start codon upstream of their peptide-coding region[86]. Therefore, prediction of ribosomal frameshiftpromoting sequences[115,116] upstream of peptide-coding regions should provide further insights as to whether PRF in non-viral mRNAs is an anecdotal or an important player in the biogenesis of cryptic MAPs.

3.7.3 Termination: let's put an end to it… or not!

The final step of translation, called termination, involves the recognition of a stop codon in the ribosome's A site by the eukaryotic release factor 1, which in turn stalls the ribosome and promotes the release of the translated protein product. Through the recruitment of several other release factors the stalled complex is dismantled and can then be recycled before re-initiating translation on another mRNA. Although stop codons are known to efficiently stop translation, stop codon read-through exists[117]. During this process, the stop codon is often decoded by an aminoacylated near-cognate-tRNA rather than the classical eukaryotic release factor 1[118], thereby allowing the ribosome to translate the transcript's 3'UTR down to the next in-frame stop codon, which leads to the production of C-terminal-extended proteins, part conventional and part cryptic (**Figure 3.2d**). Although we are only starting to uncover the precise mechanistic of stop codon read-through[119], we know that its incidence can be influenced by several factors including (1) the nature of the stop codon per se, UGA and UAG being leakier than UAA[120], (2) the direct upstream and downstream nucleotide context[121,122] as well as (3) the availability of release factors[123,124]. At least one functional human protein is known to derive from stop codon read-through: the C-terminal-extended version of the vascular endothelial factor A, called VEGF-Ax, which exhibits an antiangiogenic activity as opposed to its longer proangiogenic isoform[125]. Even more recently, a study on *C. elegans* and cultured human cells demonstrated that inappropriate 3'UTR translation decreased the final protein expression level, which suggested that erroneously C-terminal-extended proteins are rapidly degraded[126]. The

rapid degradation of such extended proteins should ensure the fair representation of 3'UTR-derived cryptic MAPs in-frame with their upstream protein-coding region in the MAP repertoire. However, none of the 3'UTR-derived cryptic MAPs identified in our study[86] could have derived from such events since they were far from the primary stop codon, rarely in-frame with it and separated from it by at least one other stop codon (**Figure 3.3a**). Of note, one of our 3'UTR-derived cryptic MAPs started 22 codons after the primary stop codon and was not separated by any additional stop codons from the protein-coding region, at least in its +2 frame (rather than the +1 frame used to translate the conventional protein). Since stop codons have also been shown to trigger ribosomal frameshifting in Euplotes[127], one could speculate that this 3'UTR-derived cryptic MAP was generated by such mechanism (**Figure 3.3b**). In any case, both stop codon read-through and ribosomal frameshifting at stop codons appear to be a poor source of cryptic MAPs[128], probably because the occurrence of both processes is rare. Moreover, we know that stop codons are underrepresented in protein-coding regions but not elsewhere, which means that they are as used as any other codon in non-coding regions[129]. Consequently, the few ribosomes skipping the primary stop codon of a protein-coding region are very likely to meet another stop codon not so far down, which is coherent with the observation that predicted C-terminal extensions caused by stop codon read-through are way shorter than the median length of proteins (27.8 vs. 449 amino acids)[130]. Since we know that longer proteins generate more MAPs just by virtue of their size[37,38], it is logical to assume that the same principle is at play here: being longer, the conventional portion of the protein is more likely to generate MAPs than the shorter 3'UTR-derived cryptic portion.

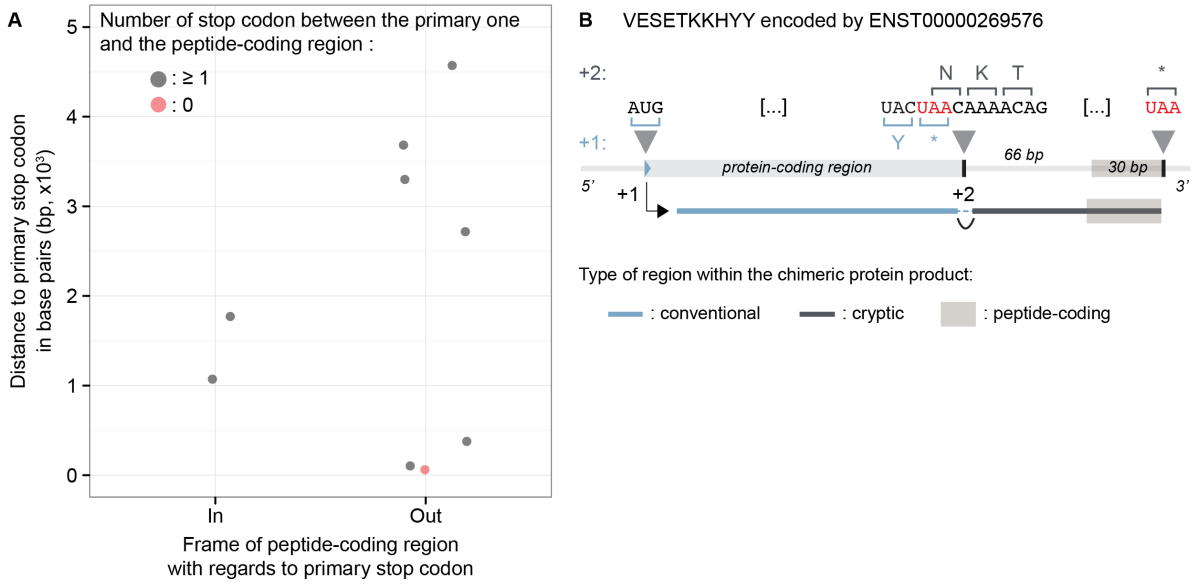


Figure 3.3 | Most 3′UTR-derived cryptic MAPs do not result from stop codon read-through. (**a**) Peptide-coding regions of 3'UTR-derived cryptic MAPs are far downstream of the primary stop codon. *Dot plot* representing the distance between the primary stop codon and the beginning of the peptide-coding region. (**b**) Schematic representation of the 3'UTR-derived cryptic MAP that potentially derives from a ribosomal frameshift event triggered at the primary stop codon. The nine 3'UTR-derived cryptic MAPs all come from Ref. 86.

3.7.4 Are cryptic MAPs proteasome independent?

As mentioned earlier, cryptic proteins identified by MS appear to have a short median length relative to conventional proteins[54,56]. Since this is also true for the predicted protein precursors of cryptic MAPs[86], we hypothesized that cryptic proteins required minimal if any proteasomal processing before entering the MHC I antigen presentation pathway. The proteasome is known to define the C-terminal end of MAPs[7]. Interestingly, we observed that 32 precursor proteins (out of 168) were an N-terminal extension of the final cryptic MAP, which is underlined in the following, non-exhaustive, list: MRATKPTVQK, MIASAGVKRVL, METVASAATVGAAREKVAV and MSNKLRKCAHLGKVGRKLEEMSEEGQMSEEGSDRRCRQEGSLSVI/VGGGPGECV TEEAKKVEIFKMRENGQWRPVSQRIQVREDQKHVHRIWQVSLATAISAK. They,

therefore, required no C terminus trimming. Moreover, looking at the C-terminal end of cryptic MAPs, we observed that the amino acid usage differed from that of conventional MAPs. Indeed, up to four amino acids before the proteasomal cleavage site (P4–P1), cryptic and conventional MAPs displayed major differences in the frequency of hydrophobic vs. polar and large vs. small/ medium amino acids, which suggested that the C terminus of cryptic MAPs was proteasome independent[86]. Although this is not absolute proof, these preliminary observations suggest that, as opposed to conventional proteins, cryptic proteins, or at least some of them, are unconventional DRiPs because they derive from the translation of allegedly non-coding sequences and are likely to be rapidly degraded as they require minimal processing for MAP generation and presentation (**Figure 3.1**).

3.8 On the use of cryptic MAPs in cancer immunotherapy

Despite their discovery some 30 years ago, cryptic MAPs have never been truly exploited in the context of cancer immunotherapy. The fact that some tumor-reactive T cells recognized cryptic MAPs has been a serendipitous retrospective finding. In these reports, investigators did not know which MAPs were targeted by re-infused tumor-infiltrating lymphocytes[131]. Only careful a posteriori studies on the antigen specificity of these tumor-infiltrating lymphocytes revealed that some of the recognized MAPs were indeed cryptic MAPs[69,72,76]. Besides the fact that some cryptic MAPs are sufficiently immunogenic to promote tumor regression, what are the pros and cons for their potential use as targets for cancer immunotherapy?

3.8.1 The pros…

First, cryptic MAPs are more than just translational noise. They derive from the degradation of atypical proteins produced by precise, and even novel, translational mechanisms implicating highly expressed transcripts[86]. Their coding regions might not be as conserved as the ones encoding conventional proteins; however, their translational activity is indeed conserved between mouse and human cells[132]. In humans, cryptic MAPs are shared by subjects expressing the relevant MHC I allotype[86].

Second, cryptic MAPs may considerably expand the landscape of tumor-specific antigens (TSAs). Efforts to discover TSAs are currently limited to the identification of cancer-specific non-synonymous somatic exonic mutations[133-136]. Extending the TSA search space to cryptic MAPs would significantly increase our chances to find mutated TSAs because 99% of cancer-specific mutations are in non-coding regions[137]. Furthermore, regions encoding cryptic MAPs are enriched in non-synonymous single nucleotide polymorphisms when compared to conventional ones[86], which suggests that cryptic MAPs may accumulate numerous somatic mutations in cancer. Moreover, as suggested by Townsend et al.[138], cancer-specific cryptic MAPs do not necessarily need to bear point mutations to be immunogenic. In fact, targeting cryptic MAPs would allow us to leverage a whole new world of cancer-specific mutations, namely structural variants such as large insertions/deletions or even fusions, which have all the potential

to create de novo ORFs or modifies existing ones by altering their frame and, therefore, produce MAPs recognized as non-self. Another advantage of these ORFs is that their TSA-generating potential should be higher than the one of point mutations. Indeed, while a point mutation can generate a single neo-MAP, an entire ORF can generate multiple neo-MAPs. Of course, not all cancer types would be able to generate such TSAs; however, they are worth exploring for microsatellite instable colorectal[139,140] and blood cancers[141,142] as well as cancer types presenting with a lot of structural variants such as melanoma[143], prostate cancer[144], bladder urothelial carcinoma, lung adenocarcinoma and squamous cell carcinoma, breast cancer, glioblastoma, head and neck squamous cell carcinoma, ovarian serous cystadenocarcinoma or low-grade glioma[145,146]. In a pre-clinical phase, such cancer-specific cryptic MAPs could be identified by proteogenomic approaches leveraging both ribosome profiling and RNA-sequencing data of tumor and normal-matched cells to identify any type of translated tumor-specific ORFs and mutations, including structural variants. However, efficient detection of such variants will require the development of novel methods to analyze RNA-sequencing data that would be more sensitive than current mappers and could be inspired by k-mer-based alignment-free methods[147,148].

Third, generation of cryptic MAPs can be modulated. This was suggested by at least three independent reports showing that specific stress conditions shut down canonical translation but increase non-canonical 5'UTR translation on stress-privileged transcripts[99,149,150]. Moreover, viral infection has been shown to increase the generation of cryptic MAPs via the effect of pro-inflammatory cytokines such as type I interferon and tumor necrosis factor α[151]. If this conclusion extends to the tumor microenvironment, which is known to contain both types of cytokines, one could hypothesize that inflammation could enhance tumor immunogenicity by increasing the generation of cryptic, yet cancer-specific, MAPs. Finally, we mentioned earlier that 3'UTR-derived cryptic MAPs caused by stop read-through are rare. However, several aminoglycoside-derived drugs are known to promote such stop codon read-through events and are currently used to treat diseases caused by nonsense mutations within protein-coding genes preventing the production of the full-length functional protein

product. If targeted to tumor cells, such drugs, as well as any other drug altering translation fidelity, could be used to artificially promote the generation of cryptic MAPs that are not produced by normal cells and should, therefore, have a strong immunogenic potential[152].

3.8.2 There are no cons, only unanswered questions!

First, we do not know the breadth of the cryptic MAP repertoire. For now, we have estimated that cryptic MAPs represent ~10% of MAPs presented at the cell surface of B-LCLs[86]. However, this proteogenomic analysis was performed on few samples of a single cell type, i.e., B cells. We speculate that this number is probably an underestimate because our database was built using the six-frame translation of reads derived from poly(A) RNA sequencing, while we know that (1) nuclear translation of unspliced and, therefore, non-polyadenylated, mRNAs has been involved in the generation of cryptic MAPs[153] and that (2) six- frame translation databases can limit MS identification due to an inaccurate overestimation of the false discovery rate[154]. Therefore, cogent assessment of the breadth of the cryptic MAP repertoire will require (1) analyses of more samples of different origins and (2) using databases with less noise. For the last point, performing stranded RNA sequencing and ribosomal profiling on the same sample should help predict which ORFs are actively translated and in which frame, thereby allowing the construction of smaller and truer databases, which should improve MS identification for cryptic MAPs as it did for proteins[155].

Second, we do not know to which extent the cryptic MAP repertoire is tolerated. We know that selected cryptic MAPs presented on mTECs are sufficient to trigger central tolerance[82,83], but we do not know how many cryptic MAPs can induce central tolerance. Because mTECs are rare and tricky to extract, both MS analysis and ribosomal profiling are unlikely to be performed on them any time soon. Therefore, answering this question will require the development of algorithm able to reliably predict both conventional and cryptic MAPs from RNA-sequencing reads.

Finally, we do not know to which extent the cryptic MAP repertoire is plastic. By this we mean, how different is the cryptic repertoire of normal vs. cancer cells, whether

they originate from the primary tumor or metastatic lesions. Because cancer cells are *de facto* stressed cells evolving in an abnormal microenvironment and because the generation of cryptic MAPs can be affected by these factors, one could hypothesize that the normal and cancer cryptic MAP repertoires may differ considerably. However, this issue needs to be analyzed as a function of time (e.g., cancer progression) and space (e.g., tumor heterogeneity). In the same vein, how do chemotherapeutic agents or any other stressors affect the cryptic MAP repertoire: are some compounds more prone than others to increase its breadth and, therefore, increase tumor immunogenicity? Analysis at the proteome level suggests that non-canonical translation is increased following chemotherapy[156], but we need complementary studies on (1) more compounds and (2) at the immunopeptidome level to see how those proteomic changes can translate into the MAP repertoire.

3.9 Concluding remarks

About 30 years ago, cryptic MAPs were just odd observations that no one could really explain. Today, they have unquestionably won their place in the MAP repertoire and several landmark mechanistic studies have helped us to understand their biogenesis, which significantly differs from that of conventional MAPs. Despite tremendous progresses made on the mechanistic front, more proteogenomic analyses are required to determine the breath as well as the temporal and spatial plasticity of the cryptic MAP repertoire. In particular, they will allow us to determine whether cryptic MAPs can become actionable targets for cancer immunotherapy.

3.10 Acknowledgements

Research performed in the authors' lab was supported by a Grant from the Quebec Breast Cancer Foundation. C.M.L. is supported by a Cole Foundation fellowship. C.P. holds the Canadian Research Chair in Immunobiology. We apologize to authors whose work has not been cited due to space limitations.

3.11 References

1. Bassani-Sternberg, M., et al., *Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.* Nat Commun, 2016. **7**: p. 13404.
2. Anderson, M.S. and M.A. Su, *AIRE expands: new roles in immune tolerance and beyond.* Nat Rev Immunol, 2016. **16**(4): p. 247-58.
3. Anderson, M.S., et al., *Projection of an immunological self shadow within the thymus by the aire protein.* Science, 2002. **298**(5597): p. 1395-401.
4. Liston, A., et al., *Aire regulates negative selection of organ-specific T cells.* Nat Immunol, 2003. **4**(4): p. 350-4.
5. Anderson, M.S., et al., *The cellular mechanism of Aire control of T cell tolerance.* Immunity, 2005. **23**(2): p. 227-39.
6. Burnet, F.M., *The concept of immunological surveillance.* Prog Exp Tumor Res, 1970. **13**: p. 1-27.
7. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nat Rev Immunol, 2011. **11**(12): p. 823-36.
8. Neefjes, J.J., F. Momburg, and G.J. Hammerling, *Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter.* Science, 1993. **261**(5122): p. 769-71.
9. Serwold, T., et al., *ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum.* Nature, 2002. **419**(6906): p. 480-3.
10. Peaper, D.R. and P. Cresswell, *Regulation of MHC class I assembly and peptide binding.* Annu Rev Cell Dev Biol, 2008. **24**: p. 343-68.
11. Lehner, P.J., M.J. Surman, and P. Cresswell, *Soluble tapasin restores MHC class I expression and function in the tapasin-negative cell line .220.* Immunity, 1998. **8**(2): p. 221-31.
12. Morozov, G.I., et al., *Interaction of TAPBPR, a tapasin homolog, with MHC-I molecules promotes peptide editing.* Proc Natl Acad Sci U S A, 2016. **113**(8): p. E1006-15.

13. Thomas, C. and R. Tampe, *Proofreading of Peptide-MHC Complexes through Dynamic Multivalent Interactions.* Front Immunol, 2017. **8**: p. 65.
14. Spiliotis, E.T., et al., *Selective export of MHC class I molecules from the ER after their dissociation from TAP.* Immunity, 2000. **13**(6): p. 841-51.
15. Beck, M., et al., *The quantitative proteome of a human cell line.* Mol Syst Biol, 2011. **7**: p. 549.
16. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Mol Syst Biol, 2011. **7**: p. 548.
17. Mester, G., V. Hoffmann, and S. Stevanovic, *Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands.* Cell Mol Life Sci, 2011. **68**(9): p. 1521-32.
18. Robinson, J., et al., *The IPD and IMGT/HLA database: allele variant databases.* Nucleic Acids Res, 2015. **43**(Database issue): p. D423-31.
19. van Deutekom, H.W. and C. Kesmir, *Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most?* Immunogenetics, 2015. **67**(8): p. 425-36.
20. Falk, K., et al., *Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.* Nature, 1991. **351**(6324): p. 290-6.
21. Granados, D.P., et al., *MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements.* Blood, 2012. **119**(26): p. e181-191.
22. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biol, 2017. **18**(1): p. 76.
23. Goulder, P.J. and B.D. Walker, *HIV and HLA class I: an evolving relationship.* Immunity, 2012. **37**(3): p. 426-40.
24. Goulder, P.J., et al., *Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS.* Nat Med, 1997. **3**(2): p. 212-7.
25. Kiepiela, P., et al., *CD8+ T-cell responses to different HIV proteins have discordant associations with viral load.* Nat Med, 2007. **13**(1): p. 46-53.

26. Cortes, A., et al., *Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1.* Nat Commun, 2015. **6**: p. 7146.
27. de Verteuil, D., et al., *Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules.* Mol Cell Proteomics, 2010. **9**(9): p. 2034-47.
28. Dudek, N.L., et al., *Constitutive and inflammatory immunopeptidome of pancreatic beta-cells.* Diabetes, 2012. **61**(11): p. 3018-25.
29. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.
30. Wahl, A., et al., *HLA class I molecules reflect an altered host proteome after influenza virus infection.* Hum Immunol, 2010. **71**(1): p. 14-22.
31. Esquivel, F., J. Yewdell, and J. Bennink, *RMA/S cells present endogenously synthesized cytosolic proteins to class I-restricted cytotoxic T lymphocytes.* J Exp Med, 1992. **175**(1): p. 163-8.
32. Croft, N.P., et al., *Kinetics of antigen expression and epitope presentation during virus infection.* PLoS Pathog, 2013. **9**(1): p. e1003129.
33. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
34. Anton, L.C. and J.W. Yewdell, *Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors.* J Leukoc Biol, 2014. **95**(4): p. 551-62.
35. Schubert, U., et al., *Rapid degradation of a large fraction of newly synthesized proteins by proteasomes.* Nature, 2000. **404**(6779): p. 770-4.
36. Cardinaud, S., et al., *The synthesis of truncated polypeptides for immune surveillance and viral evasion.* PLoS One, 2010. **5**(1): p. e8692.
37. Bassani-Sternberg, M., et al., *Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.* Mol Cell Proteomics, 2015. **14**(3): p. 658-73.

38. Pearson, H., et al., *MHC class I-associated peptides derive from selective regions of the human genome.* J Clin Invest, 2016. **126**(12): p. 4690-4701.
39. Fortier, M.H., et al., *The MHC class I peptide repertoire is molded by the transcriptome.* J Exp Med, 2008. **205**(3): p. 595-610.
40. Floor, S.N. and J.A. Doudna, *Tunable protein synthesis by transcript isoforms in human cells.* Elife, 2016. **5**: p. e10921.
41. Calvo, S.E., D.J. Pagliarini, and V.K. Mootha, *Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans.* Proc Natl Acad Sci U S A, 2009. **106**(18): p. 7507-12.
42. Ravid, T. and M. Hochstrasser, *Diversity of degradation signals in the ubiquitin-proteasome system.* Nat Rev Mol Cell Biol, 2008. **9**(9): p. 679-90.
43. Rechsteiner, M. and S.W. Rogers, *PEST sequences and regulation by proteolysis.* Trends Biochem Sci, 1996. **21**(7): p. 267-71.
44. van der Lee, R., et al., *Intrinsically disordered segments affect protein half-life in the cell and during evolution.* Cell Rep, 2014. **8**(6): p. 1832-44.
45. Gu, W., et al., *Both treated and untreated tumors are eliminated by short hairpin RNA-based induction of target-specific immune responses.* Proc Natl Acad Sci U S A, 2009. **106**(20): p. 8314-9.
46. Apcher, S., et al., *Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation.* Proc Natl Acad Sci U S A, 2011. **108**(28): p. 11572-7.
47. Eisenlohr, L.C., L. Huang, and T.N. Golovina, *Rethinking peptide supply to MHC class I molecules.* Nat Rev Immunol, 2007. **7**(5): p. 403-10.
48. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
49. Ingolia, N.T., et al., *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.
50. Johnstone, T.G., A.A. Bazzini, and A.J. Giraldez, *Upstream ORFs are prevalent translational repressors in vertebrates.* EMBO J, 2016. **35**(7): p. 706-23.

51. Arribere, J.A. and W.V. Gilbert, *Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing.* Genome Res, 2013. **23**(6): p. 977-87.
52. Kim, M.S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575-81.
53. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.
54. Vanderperre, B., et al., *Direct detection of alternative open reading frames translation products in human significantly expands the proteome.* PLoS One, 2013. **8**(8): p. e70698.
55. Lubec, G. and L. Afjehi-Sadat, *Limitations and pitfalls in protein identification by mass spectrometry.* Chem Rev, 2007. **107**(8): p. 3568-84.
56. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.
57. Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**(3): p. 193-204.
58. D'Lima, N.G., et al., *A human microprotein that interacts with the mRNA decapping complex.* Nat Chem Biol, 2017. **13**(2): p. 174-180.
59. Kearse, M.G., et al., *CGG Repeat-Associated Non-AUG Translation Utilizes a Cap-Dependent Scanning Mechanism of Initiation to Produce Toxic Proteins.* Mol Cell, 2016. **62**(2): p. 314-22.
60. Boon, T., et al., *Cloning and characterization of genes coding for tum- transplantation antigens.* J Autoimmun, 1989. **2 Suppl**: p. 109-14.
61. Boon, T. and A. Van Pel, *T cell-recognized antigenic peptides derived from the cellular genome are not protein degradation products but can be generated directly by transcription and translation of short subgenic regions. A hypothesis.* Immunogenetics, 1989. **29**(2): p. 75-9.
62. Fetten, J.V., N. Roy, and E. Gilboa, *A frameshift mutation at the NH2 terminus of the nucleoprotein gene does not affect generation of cytotoxic T lymphocyte epitopes.* J Immunol, 1991. **147**(8): p. 2697-705.

63. Shastri, N. and F. Gonzalez, *Endogenous generation and presentation of the ovalbumin peptide/Kb complex to T cells.* J Immunol, 1993. **150**(7): p. 2724-36.
64. Uenaka, A., et al., *Identification of a unique antigen peptide pRL1 on BALB/c RL male 1 leukemia recognized by cytotoxic T lymphocytes and its relation to the Akt oncogene.* J Exp Med, 1994. **180**(5): p. 1599-607.
65. Coulie, P.G., et al., *A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma.* Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7976-80.
66. Guilloux, Y., et al., *A peptide recognized by human cytolytic T lymphocytes on HLA-A2 melanomas is encoded by an intron sequence of the N-acetylglucosaminyltransferase V gene.* J Exp Med, 1996. **183**(3): p. 1173-83.
67. Robbins, P.F., et al., *The intronic region of an incompletely spliced gp100 gene transcript encodes an epitope recognized by melanoma-reactive tumor-infiltrating lymphocytes.* J Immunol, 1997. **159**(1): p. 303-8.
68. Lupetti, R., et al., *Translation of a retained intron in tyrosinase-related protein (TRP) 2 mRNA generates a new cytotoxic T lymphocyte (CTL)-defined and shared human melanoma antigen not expressed in normal cells of the melanocytic lineage.* J Exp Med, 1998. **188**(6): p. 1005-16.
69. Wang, R.F., et al., *Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen.* J Exp Med, 1996. **183**(3): p. 1131-40.
70. Mayrand, S.M., D.A. Schwarz, and W.R. Green, *An alternative translational reading frame encodes an immunodominant retroviral CTL determinant expressed by an immunodeficiency-causing retrovirus.* J Immunol, 1998. **160**(1): p. 39-50.
71. Shichijo, S., et al., *A gene encoding antigenic peptides of human squamous cell carcinoma recognized by cytotoxic T lymphocytes.* J Exp Med, 1998. **187**(3): p. 277-88.
72. Wang, R.F., et al., *A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames.* J Immunol, 1998. **161**(7): p. 3598-606.

73. Aarnoudse, C.A., et al., *Interleukin-2-induced, melanoma-specific T cells recognize CAMEL, an unexpected translation product of LAGE-1.* Int J Cancer, 1999. **82**(3): p. 442-8.
74. Ronsin, C., et al., *A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ.* J Immunol, 1999. **163**(1): p. 483-90.
75. Probst-Kepper, M., et al., *An alternative open reading frame of the human macrophage colony-stimulating factor gene is independently translated and codes for an antigenic peptide of 14 amino acids recognized by tumor-infiltrating CD8 T lymphocytes.* J Exp Med, 2001. **193**(10): p. 1189-98.
76. Rosenberg, S.A., et al., *Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy.* J Immunol, 2002. **168**(5): p. 2402-7.
77. Saulquin, X., et al., *+1 Frameshifting as a novel mechanism to generate a cryptic cytotoxic T lymphocyte epitope derived from human interleukin 10.* J Exp Med, 2002. **195**(3): p. 353-8.
78. Cardinaud, S., et al., *Identification of cryptic MHC I-restricted epitopes encoded by HIV-1 alternative reading frames.* J Exp Med, 2004. **199**(8): p. 1053-63.
79. Ho, O. and W.R. Green, *Cytolytic CD8+ T cells directed against a cryptic epitope derived from a retroviral alternative reading frame confer disease protection.* J Immunol, 2006. **176**(4): p. 2470-5.
80. Dolstra, H., et al., *A human minor histocompatibility antigen specific for B cell acute lymphoblastic leukemia.* J Exp Med, 1999. **189**(2): p. 301-8.
81. Van Den Eynde, B.J., et al., *A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription.* J Exp Med, 1999. **190**(12): p. 1793-800.
82. Schwab, S.R., et al., *Constitutive display of cryptic translation products by MHC class I molecules.* Science, 2003. **301**(5638): p. 1367-71.

83. Charpentier, M., et al., *IRES-dependent translation of the long non coding RNA meloe in melanoma cells produces the most immunogenic MELOE antigens.* Oncotarget, 2016. **7**(37): p. 59704-59713.
84. van Bergen, C.A., et al., *Selective graft-versus-leukemia depends on magnitude and diversity of the alloreactive T cell response.* J Clin Invest, 2017. **127**(2): p. 517-529.
85. Kracht, M.J., et al., *Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes.* Nat Med, 2017. **23**(4): p. 501-507.
86. Laumont, C.M., et al., *Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames.* Nat Commun, 2016. **7**: p. 10238.
87. Jackson, R.J., C.U. Hellen, and T.V. Pestova, *The mechanism of eukaryotic translation initiation and principles of its regulation.* Nat Rev Mol Cell Biol, 2010. **11**(2): p. 113-27.
88. Kozak, M., *Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes.* Cell, 1986. **44**(2): p. 283-92.
89. Liu, J., et al., *Initiation of translation from a downstream in-frame AUG codon on BRCA1 can generate the novel isoform protein DeltaBRCA1(17aa).* Oncogene, 2000. **19**(23): p. 2767-73.
90. Calviello, L., et al., *Detecting actively translated open reading frames in ribosome profiling data.* Nat Methods, 2016. **13**(2): p. 165-70.
91. Liang, H., et al., *PTENalpha, a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism.* Cell Metab, 2014. **19**(5): p. 836-48.
92. Liang, H., et al., *PTENbeta is an alternatively translated isoform of PTEN that regulates rDNA transcription.* Nat Commun, 2017. **8**: p. 14771.
93. Ivanov, I.P., et al., *Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences.* Nucleic Acids Res, 2011. **39**(10): p. 4220-34.

94. Lee, S., et al., *Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution.* Proc Natl Acad Sci U S A, 2012. **109**(37): p. E2424-32.
95. Bullock, T.N. and L.C. Eisenlohr, *Ribosomal scanning past the primary initiation codon as a mechanism for expression of CTL epitopes encoded in alternative reading frames.* J Exp Med, 1996. **184**(4): p. 1319-29.
96. Starck, S.R., et al., *Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I.* Science, 2012. **336**(6089): p. 1719-23.
97. Yang, N., et al., *Defining Viral Defective Ribosomal Products: Standard and Alternative Translation Initiation Events Generate a Common Peptide from Influenza A Virus M2 and M1 mRNAs.* J Immunol, 2016. **196**(9): p. 3608-17.
98. Quiros, P.M., A. Mottis, and J. Auwerx, *Mitonuclear communication in homeostasis and stress.* Nat Rev Mol Cell Biol, 2016. **17**(4): p. 213-26.
99. Starck, S.R., et al., *Translation from the 5′ untranslated region shapes the integrated stress response.* Science, 2016. **351**(6272): p. aad3867.
100. Hecht, A., et al., *Measurements of translation initiation from all 64 codons in E. coli.* Nucleic Acids Res, 2017. **45**(7): p. 3615-3626.
101. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.
102. Malarkannan, S., et al., *Presentation of out-of-frame peptide/MHC class I complexes by a novel translation initiation mechanism.* Immunity, 1999. **10**(6): p. 681-90.
103. Golovko, A., et al., *The eIF2A knockout mouse.* Cell Cycle, 2016. **15**(22): p. 3115-3120.
104. Carbonnelle, D., et al., *The melanoma antigens MELOE-1 and MELOE-2 are translated from a bona fide polycistronic mRNA containing functional IRES sequences.* PLoS One, 2013. **8**(9): p. e75233.
105. Walsh, D. and I. Mohr, *Viral subversion of the host protein synthesis machinery.* Nat Rev Microbiol, 2011. **9**(12): p. 860-75.

106. Namy, O., et al., *A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting.* Nature, 2006. **441**(7090): p. 244-7.
107. Zook, M.B., et al., *Epitopes derived by incidental translational frameshifting give rise to a protective CTL response.* J Immunol, 2006. **176**(11): p. 6928-34.
108. Belew, A.T., V.M. Advani, and J.D. Dinman, *Endogenous ribosomal frameshift signals operate as mRNA destabilizing elements through at least two molecular pathways in yeast.* Nucleic Acids Res, 2011. **39**(7): p. 2799-808.
109. Michel, A.M., et al., *Observation of dually decoded regions of the human genome using ribosome profiling data.* Genome Res, 2012. **22**(11): p. 2219-29.
110. Matsufuji, S., et al., *Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme.* Cell, 1995. **80**(1): p. 51-60.
111. Shigemoto, K., et al., *Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting.* Nucleic Acids Res, 2001. **29**(19): p. 4079-88.
112. Wills, N.M., et al., *A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene.* J Biol Chem, 2006. **281**(11): p. 7082-8.
113. Clark, M.B., et al., *Mammalian gene PEG10 expresses two reading frames by high efficiency -1 frameshifting in embryonic-associated tissues.* J Biol Chem, 2007. **282**(52): p. 37359-69.
114. Belew, A.T., et al., *Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway.* Nature, 2014. **512**(7514): p. 265-9.
115. Belew, A.T., et al., *PRFdb: a database of computationally predicted eukaryotic programmed -1 ribosomal frameshift signals.* BMC Genomics, 2008. **9**: p. 339.
116. Theis, C., J. Reeder, and R. Giegerich, *KnotInFrame: prediction of -1 ribosomal frameshift events.* Nucleic Acids Res, 2008. **36**(18): p. 6013-20.
117. Schueren, F., et al., *Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals.* Elife, 2014. **3**: p. e03640.
118. Blanchet, S., et al., *New insights into the incorporation of natural suppressor tRNAs at stop codons in Saccharomyces cerevisiae.* Nucleic Acids Res, 2014. **42**(15): p. 10061-72.

119. Beznoskova, P., et al., *Translation initiation factor eIF3 promotes programmed stop codon readthrough.* Nucleic Acids Res, 2015. **43**(10): p. 5099-111.
120. Howard, M.T., et al., *Sequence specificity of aminoglycoside-induced stop condon readthrough: potential implications for treatment of Duchenne muscular dystrophy.* Ann Neurol, 2000. **48**(2): p. 164-9.
121. Floquet, C., et al., *Statistical analysis of readthrough levels for nonsense mutations in mammalian cells reveals a major determinant of response to gentamicin.* PLoS Genet, 2012. **8**(3): p. e1002608.
122. Loughran, G., et al., *Evidence of efficient stop codon readthrough in four mammalian genes.* Nucleic Acids Res, 2014. **42**(14): p. 8928-38.
123. Carnes, J., et al., *Stop codon suppression via inhibition of eRF1 expression.* RNA, 2003. **9**(6): p. 648-53.
124. Chauvin, C., et al., *Involvement of human release factors eRF3a and eRF3b in translation termination and regulation of the termination complex formation.* Mol Cell Biol, 2005. **25**(14): p. 5801-11.
125. Eswarappa, S.M., et al., *Programmed translational readthrough generates antiangiogenic VEGF-Ax.* Cell, 2014. **157**(7): p. 1605-18.
126. Arribere, J.A., et al., *Translation readthrough mitigation.* Nature, 2016. **534**(7609): p. 719-23.
127. Lobanov, A.V., et al., *Position-dependent termination and widespread obligatory frameshifting in Euplotes translation.* Nat Struct Mol Biol, 2017. **24**(1): p. 61-68.
128. Bullock, T.N., et al., *Initiation codon scanthrough versus termination codon readthrough demonstrates strong potential for major histocompatibility complex class I-restricted cryptic epitope expression.* J Exp Med, 1997. **186**(7): p. 1051-8.
129. Baranov, P.V., J.F. Atkins, and M.M. Yordanova, *Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning.* Nat Rev Genet, 2015. **16**(9): p. 517-29.
130. Shibata, N., et al., *Degradation of Stop Codon Read-through Mutant Proteins via the Ubiquitin-Proteasome System Causes Hereditary Disorders.* J Biol Chem, 2015. **290**(47): p. 28428-37.

131. Topalian, S.L., et al., *Immunotherapy of patients with advanced cancer using tumor-infiltrating lymphocytes and recombinant interleukin-2: a pilot study.* J Clin Oncol, 1988. **6**(5): p. 839-53.
132. Fields, A.P., et al., *A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation.* Mol Cell, 2015. **60**(5): p. 816-27.
133. Yadav, M., et al., *Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.* Nature, 2014. **515**(7528): p. 572-6.
134. Kreiter, S., et al., *Mutant MHC class II epitopes drive therapeutic immune responses to cancer.* Nature, 2015. **520**(7549): p. 692-6.
135. Verdegaal, E.M., et al., *Neoantigen landscape dynamics during human melanoma-T cell interactions.* Nature, 2016. **536**(7614): p. 91-5.
136. Stevanovic, S., et al., *Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer.* Science, 2017. **356**(6334): p. 200-205.
137. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nat Rev Genet, 2016. **17**(2): p. 93-108.
138. Townsend, A., et al., *Source of unique tumour antigens.* Nature, 1994. **371**(6499): p. 662.
139. Schwitalle, Y., et al., *Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers.* Gastroenterology, 2008. **134**(4): p. 988-97.
140. Inderberg, E.M., et al., *T cell therapy targeting a public neoantigen in microsatellite instable colon cancer reduces in vivo tumor growth.* Oncoimmunology, 2017. **6**(4): p. e1302631.
141. de Rijke, B., et al., *A frameshift polymorphism in P2X5 elicits an allogeneic cytotoxic T lymphocyte response associated with remission of chronic myeloid leukemia.* J Clin Invest, 2005. **115**(12): p. 3506-16.
142. Maletzki, C., et al., *Frameshift-derived neoantigens constitute immunotherapeutic targets for patients with microsatellite-instable*

haematological malignancies: frameshift peptides for treating MSI+ blood cancers. Eur J Cancer, 2013. **49**(11): p. 2587-95.

143. Hayward, N.K., et al., *Whole-genome landscapes of major melanoma subtypes.* Nature, 2017. **545**(7653): p. 175-180.
144. Baca, S.C., et al., *Punctuated evolution of prostate cancer genomes.* Cell, 2013. **153**(3): p. 666-77.
145. Mertens, F., et al., *The emerging complexity of gene fusions in cancer.* Nat Rev Cancer, 2015. **15**(6): p. 371-81.
146. Yoshihara, K., et al., *The landscape and therapeutic relevance of cancer-associated transcript fusions.* Oncogene, 2015. **34**(37): p. 4845-54.
147. Philippe, N., et al., *CRAC: an integrated approach to the analysis of RNA-seq reads.* Genome Biol, 2013. **14**(3): p. R30.
148. Li, Y., et al., *ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data.* Nucleic Acids Res, 2017.
149. Andreev, D.E., et al., *Translation of 5' leaders is pervasive in genes resistant to eIF2 repression.* Elife, 2015. **4**: p. e03971.
150. Gerashchenko, M.V., A.V. Lobanov, and V.N. Gladyshev, *Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress.* Proc Natl Acad Sci U S A, 2012. **109**(43): p. 17394-9.
151. Prasad, S., S.R. Starck, and N. Shastri, *Presentation of Cryptic Peptides by MHC Class I Is Enhanced by Inflammatory Stimuli.* J Immunol, 2016. **197**(8): p. 2981-2991.
152. Goodenough, E., et al., *Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR.* Proc Natl Acad Sci U S A, 2014. **111**(15): p. 5670-5.
153. Apcher, S., et al., *Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17951-6.
154. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.

155. Crappe, J., et al., *PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration.* Nucleic Acids Res, 2015. **43**(5): p. e29.
156. Wiita, A.P., et al., *Global cellular response to chemotherapy-induced apoptosis.* Elife, 2013. **2**: p. e01236.

CHAPTER 4

4 Non-coding regions are the main source of targetable tumor-specific antigens

Céline M. Laumont[1,2,*], Krystel Vincent[1,2,*], Leslie Hesnard[1,2], Éric Audemard[1], Éric Bonneil[1], Jean-Philippe Laverdure[1], Patrick Gendron[1], Mathieu Courcelles[1], Marie-Pierre Hardy[1], Caroline Côté[1], Chantal Durette[1], Charles St-Pierre[1,2], Mohamed Benhammadi[1,2], Joël Lanoix[1], Suzanne Vobecky[3], Élie Haddad[3], Sébastien Lemieux[1,4], Pierre Thibault[1,5, †], and Claude Perreault[1,2, †,‡]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[2]Department of Medicine, Faculty of Medicine, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[3]CHU Sainte-Justine, Université de Montréal, 3175 Chemin de la Côte Ste-Catherine Montreal, Quebec, Canada, H3T 1C5.

[4]Department of Computer Science and Operations Research, Faculty of Arts and Sciences, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[5]Department of Chemistry, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7.

[*] These authors contributed equally to this work

[†] These authors jointly supervised this work

[‡] Corresponding author. E-mail: claude.perreault@umontreal.ca

Science Translational Medicine, Volume 10, Issue 470, 10.1126/scitranslmed.aau5516 (December 5, 2018)

4.1 Context

We showed that, in normal cells, cryptic MAPs represent a significant portion of the MAP repertoire. These MAPs, generated by reproducible non-canonical translation events, derive from genomic regions that are polymorphic at the population level. This last observation lead us to hypothesize that cryptic MAPs could be an important source of TSAs in cancer cells.

Since no resources were available to identify cryptic TSAs, we developed an alignment-free strategy, called k-mer profiling, that allows to fish out non-tolerogenic sequences from the transcriptome of cancer cells by comparing it to the one of TEC/ mTEC. Those sequences are further translated in order to generate DBs suitable for MS. Using this approach we were able to characterize the TSA landscape of nine tumors and demonstrated that non-coding regions are the main source of TSAs. Moreover, most of these TSAs are aeTSAs often derived from EREs. Because those TSAs are likely to be shared between patients, we do believe that this discovery will push forward the development of cancer immunotherapy, especially for cancer types with a low mutational load.

4.2 Authors' contributions

Céline M. Laumont*: designed the study, performed bioinformatics analyses (Figures 4.1, 4.2, 4.5 and 4.6), analyzed all data and wrote the first draft of the manuscript.

Krystel Vincent*: designed the study, performed experiments (Figures 4.3, 4.4, 4.5 and 4.6), analyzed all data and wrote the first draft of the manuscript.

Leslie Hesnard: performed experiments (Figures 4.4 and 4.5), analyzed data and contributed to the writing of first draft of the manuscript.

Éric Audemard: developed NEKTAR and general discussion on k-mers.

Éric Bonneil: acquired mass spectrometry data and validated all tumor-specific antigens identified in this project (Supplementary Figures 4.3, 4.8 and 4.9)

Jean-Philippe Laverdure: performed bioinformatics analyses (kallisto, freeBayes) for all analyzed samples.

Patrick Gendron: downloaded publicly available RNA-sequencing data (ENCODE, GTEx) and generated all resulting k-mer databases, submitted RNA-sequencing data produced by our lab to GEO.

Mathieu Courcelles: processed raw mass spectrometry data to allow further analyses and submitted all mass spectrometry data produced by our laboratory to PRIDE.

Marie-Pierre Hardy and **Caroline Côté**: prepared B-ALLs samples for mass spectrometry.

Chantal Durette: prepared B-ALLs samples and acquired mass spectrometry data.

Charles St-Pierre and Mohamed Benhammadi: prepared $mTEC^{hi}$ samples prior to RNA-sequencing.

Joël Lanoix: prepared samples for mass spectrometry (immunoprecipitation).

Suzanne Vobecky and **Élie Haddad**: provided human thymi for TEC/mTEC extraction.

Sébastien Lemieux: contributed to the analysis presented in Supplementary Figure 4.9 and general discussion on k-mers.

Pierre Thibault: analyzed data and general discussion on mass spectrometry.

Claude Perreault: designed the study, analyzed data and wrote the first draft of the manuscript.

All authors edited and approved the final version of the manuscript.

*These authors equally contributed to this work.

4.3 One sentence summary

A proteogenomic method identifies potentially actionable tumor-specific antigens and shows that most of them are not coded by classic exons.

4.4 Abstract

Tumor-specific antigens (TSAs) represent ideal targets for cancer immunotherapy, but few have been identified thus far. We therefore developed a proteogenomic approach to enable the high-throughput discovery of TSAs coded by potentially all genomic regions. In two murine cancer cell lines and seven human primary tumors, we identified a total of 40 TSAs, about 90% of which derived from allegedly non-coding regions and would have been missed by standard exome-based approaches. Moreover, the majority of these TSAs derived from non-mutated yet aberrantly expressed transcripts (such as endogenous retroelements) that could be shared by multiple tumor types. In mice, the efficacy of TSA vaccination was influenced by two parameters that can be estimated in humans and could serve for TSA prioritization in clinical studies: TSA expression and the frequency of TSA-responsive T cells in the pre-immune repertoire. In conclusion, the strategy reported herein could considerably facilitate the identification and prioritization of actionable human TSAs.

4.5 Introduction

$CD8^+$ T cells are the main mediators of naturally occurring and therapeutically induced immune responses to cancer. Accordingly, the abundance of $CD8^+$ tumor-infiltrating lymphocytes (TILs) positively correlates with response to immune checkpoint inhibitors and favorable prognosis[1-3]. As $CD8^+$ T cells recognize major histocompatibility complex class I (MHC I)-associated peptides, the most important unanswered question is the nature of the specific peptides recognized by $CD8^+$ TILs[4]. Knowing that the abundance of $CD8^+$ TILs correlates with the mutation load of tumors, the dominant paradigm holds that $CD8^+$ TILs recognize mutated tumor-specific antigens (mTSAs), commonly referred to as neoantigens[2,5,6]. The superior immunogenicity of mTSAs is ascribed to their selective expression on tumors which minimizes the risk of immune tolerance[7]. Nonetheless, some TILs have been shown to recognize cancer-restricted non-mutated MHC peptides[8] that we will refer to as aberrantly expressed TSAs (aeTSAs). aeTSAs can derive from a variety of cis- or trans-acting genetic and epigenetic changes that lead to the transcription and translation of genomic sequences normally not expressed in cells, such as endogenous retroelements (EREs)[9-11].

Considerable efforts are being devoted to discovering actionable TSAs that can be used in therapeutic cancer vaccines. The most common strategy hinges on reverse immunology, in which exome sequencing is performed on tumor cells to identify mutations, and MHC-binding prediction software tools are used to identify which mutated peptides might be good MHC binders[12,13]. Although reverse immunology can enrich for TSA candidates, at least 90% of these candidates are false positives[6,14] because available computational methods may predict MHC binding, but they cannot predict other steps involved in MHC peptide processing[15,16]. To overcome this limitation, a few studies have included mass spectrometry (MS) analyses in their TSA discovery pipeline[17], thereby providing a rigorous molecular definition of several TSAs[18,19]. However, the yield of these approaches has been meager: in melanoma, one of the most mutated tumor types, an average of two TSAs per individual tumors have been validated by MS[20], while only a handful of TSAs has been found for other

cancer types[15]. The paucity of TSAs is puzzling, because injection of TILs or immune checkpoint inhibitors would not cause tumor regression if tumors did not express immunogenic antigens[21]. We surmised that approaches based on exonic mutations have failed to identify TSAs because they did not take into account two crucial elements. First, these approaches focus only on mTSAs and neglect aeTSAs, essentially because there is currently no method for high-throughput identification of aeTSAs. This represents a major shortcoming because, whereas mTSAs are private antigens (that is, unique to a given tumor), aeTSAs would be preferred targets for vaccine development as they can be shared by multiple tumors[8,10]. Second, focusing on the exome as the only source of TSAs is very restrictive. Indeed, and of particular relevance to TSA discovery, 99% of cancer mutations are located in non-coding regions[22]. Moreover, the exome (all protein-coding sequences) represents only 2% of the human genome, whereas up to 75% of the genome can be transcribed and potentially translated[23]. Hence, many allegedly non-coding regions are in fact protein-coding, and translation of non-coding regions has been shown to generate numerous MHC peptides[24,25], some of which were retrospectively identified as targets of TILs and autoreactive T cells[26,27].

With these considerations in mind, we developed a proteogenomic strategy designed to discover mTSAs and aeTSAs coded by all genomic regions. We used this approach to study two well-characterized murine cancer cell lines, CT26 and EL4, as well as seven primary human samples comprising four B-lineage acute lymphoblastic leukemias (B-ALLs) and three lung cancers. Our main objectives were to determine whether non-coding regions contribute to the TSA landscape and which parameters may influence TSA immunogenicity.

4.6 Results

4.6.1 Rationale and design of a proteogenomic method for TSA discovery

Attempts to computationally predict TSAs using various algorithms are fraught with exceedingly high false discovery rates[28]. Hence, a systems-level molecular definition of the MHC peptide repertoire may only be achievable by high-throughput MS studies[4]. Current approaches use tandem MS (MS/MS) software tools, such as Peaks[29], which rely on a user-defined protein database to match each acquired MS/MS spectrum to a peptide sequence. As the reference proteome does not contain TSAs, MS-based TSA discovery workflows must use proteogenomic strategies to build customized databases derived from tumor RNA-sequencing (RNA-Seq) data[30] that should ideally contain all proteins, even unannotated ones, expressed in the considered tumor sample. As current MS/MS software tools cannot deal with the large search space created by translating all RNA-Seq reads in all reading frames[31,32], we devised a proteogenomic strategy enriching for cancer-specific sequences to comprehensively characterize the landscape of TSAs coded by all genomic regions. The resulting database, termed a global cancer database, is composed of two customizable parts. The first part, the canonical cancer proteome (**Figure 4.1A**), was obtained by *in silico* translation of expressed protein-coding transcripts in their canonical frame; it therefore contains proteins coded by exonic sequences that are normal or contain single-base mutations. The second part, the cancer-specific proteome **(Figure 4.1B**), was generated using an alignment-free RNA-Seq workflow called k-mer profiling, because current mappers and variant callers poorly identify structural variants. This second dataset enabled the detection of peptides encoded by any reading frame of any genomic origin (including structural variants), as long as they were cancer-specific (that is, absent from normal cells). Here, we elected to use MHC II^{hi} medullary thymic epithelial cells (mTEC^{hi}) cells as a “normal control” because they express most known genes and orchestrate T cell selection to induce central tolerance to MHC peptides coded by their vast transcriptome (**Supplementary Figure 4.1A**)[33]. Thus, to identify

RNA sequences that were cancer-specific, we chopped cancer RNA-Seq reads into 33-nucleotide-long sequences, called k-mers[34], from which we removed k-mers present in syngeneic mTEC^{hi} cells (**Supplementary Figure 4.2, A and B**). Redundancy inherent to the k-mer space was removed by assembling overlapping cancer-specific k-mers into longer sequences, called contigs, which were 3-frame translated in silico (**Figure 4.1B** and **Supplementary Figure 4.2, C and D**). We then concatenated the canonical and cancer-specific proteomes to create a global cancer database, one for each analyzed sample (**Supplementary Table 4.1A**). Using these optimized databases, we identified MHC peptides eluted from two well-characterized mouse tumor cell lines that we sequenced by MS, namely CT26, a colorectal carcinoma from a Balb/c mouse and EL4, a T-lymphoblastic lymphoma from a C57BL/6 mouse[35,36](**Figure 4.1C** and **Supplementary Table 4.2A**).

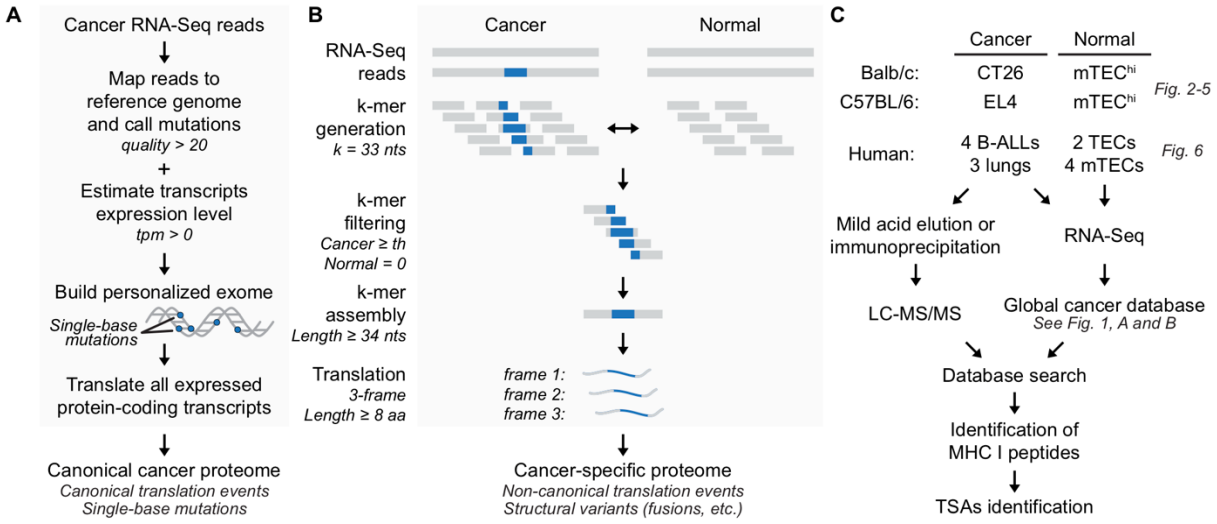


Figure 4.1 | Proteogenomic workflow for the identification of TSAs. (**A** and **B**) Schematic detailing how the canonical cancer proteome (**A**) and cancer-specific proteome (**B**) were built for each analyzed sample. In panel (**A**), 'quality' refers to the Phred score; a score >20 means that the accuracy of the nucleobase call is at least 99%. (**C**) The combination of the above two proteomes, termed the global cancer database, was then used to identify MHC peptides, and more specifically TSAs, sequenced by liquid chromatography-MS/MS (LC-MS/MS). We analyzed two well-characterized murine cell lines, CT26 and EL4, and seven human primary samples, namely four B-ALLs and three lung tumor biopsies (n = 2–4 per sample). Statistics regarding each part of the global cancer database can be found in **Supplementary Table 4.1**, and implementation details of building the cancer-specific proteome by k-mer profiling are presented in **Supplementary Figure 4.2**. aa: amino acids, nts: nucleotides, th: sample-specific threshold for k-mer occurrence (see **section 4.8.11** for details), tpm: transcripts per million.

4.6.2 Non-coding regions as a major source of TSAs

We identified 1,875 MHC peptides on CT26 cells and 783 on EL4 cells (**Supplementary Tables 4.3** and **4.4**). Among these, peptides absent from the mTEC^{hi} proteome were considered TSA candidates if (i) their 33-nucleotide-long peptide-coding sequence derived from a full cancer-restricted 33-nucleotide-long k-mer and was absent from the mTEC^{hi} transcriptome, or if (ii) their 24-to-30-nucleotide-long peptide-coding sequence, derived from a truncated version of a cancer-restricted 33-nucleotide-long k-mer, was overexpressed by at least 10-fold in the transcriptome of cancer vs. mTEC^{hi} cells (**Supplementary Figure 4.3A**). As no error estimation was used in our study, we manually validated the MS spectra of our TSA candidates. Before assigning peptides a genomic location, we also removed any indistinguishable isoleucine/leucine variants (**Supplementary Figures 4.3, B and C** and **4.4**) and ended up with a total of 6 mTSAs and 15 aeTSA candidates: 14 presented by CT26 cells and 7 by EL4 cells (**Figure 4.2, A and B**). MHC peptides that were both mutated and aberrantly expressed were included in the mTSA category. All of these peptides were absent from the Immune Epitope Database[37], except for one: the AH1 peptide (SPSYVYHQF), the sole aeTSA previously identified on CT26 cells using reverse immunology[10,38].

In order to assess the stringency of our database-building strategy based on the removal of mTEC^{hi} k-mers from cancer k-mers, we evaluated the peripheral expression of RNAs coding for aeTSAs across a panel of 22 tissues[39,40] (**Supplementary Table 4.5**). Four of the 15 aeTSA candidates had an expression profile similar to that of previously reported "overexpressed" tumor-associated antigens[41,42], as their peptide-coding sequences were expressed in most or all tissues **(Figure 4.2C)**. These four peptides were therefore excluded from the TSA list. In contrast, 11 peptides were considered genuine aeTSAs as their source transcripts were either totally absent or present at trace amounts in a few tissues **(Figure 4.2C)**. We note that detection of low transcript amounts is less relevant as MHC peptides preferentially derive from highly abundant transcripts[43,44]. This concept is illustrated by the AH1 TSA, which elicits strong antitumor responses devoid of adverse effects[10,38], despite the weak expression

of its peptide-coding sequence in the liver, thymus and urinary bladder (**Figure 4.2C**). These results demonstrate that subtracting mRNA sequences found in mTEC^{hi} strongly enriches for cancer-restricted peptide-coding sequences. When we consider our entire murine TSA dataset (6 mTSAs and 11 aeTSAs), we find that most of them derive from atypical translation events: the out-of-frame translation of a coding exon or the translation of non-coding regions (**Figure 4.2D**). We also noticed that any type of non-coding region can generate TSAs (**Supplementary Table 4.6**): intergenic and intronic sequences, non-coding exons, untranslated regions (UTRs)/exon junctions, as well as EREs, which appear to be a particularly rich source of TSAs (8 aeTSAs and 1 mTSA). Finally, our approach efficiently captured at least one structural variant as we identified an antigen, VTPVYQHL, derived from a very large intergenic deletion (~7,500 bp) in EL4 cells (**Supplementary Table 4.6B**). Altogether, these observations confirm that non-coding regions are the main source of TSAs and that they have the potential to considerably expand the TSA landscape of tumors.

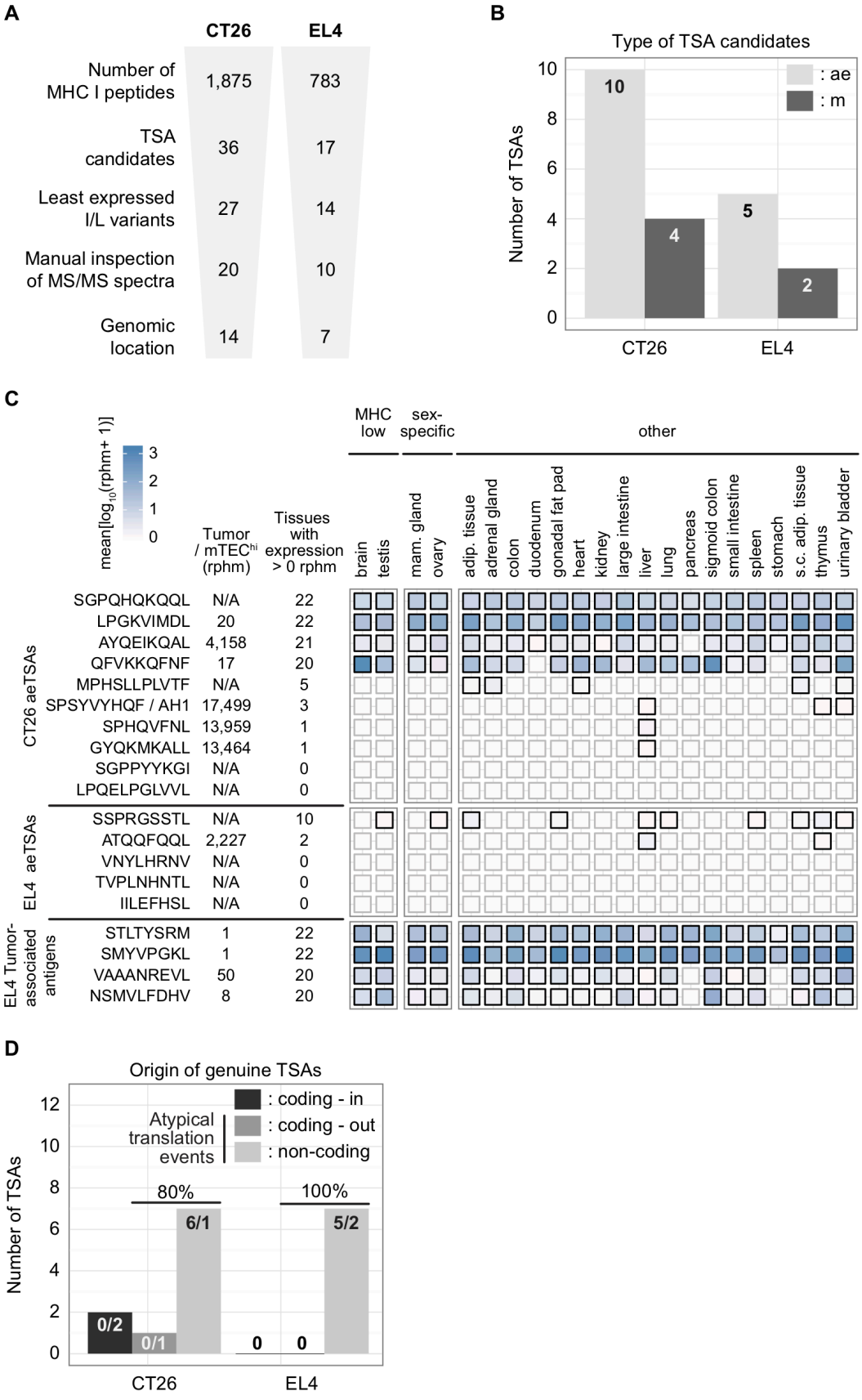

A
CT26
EL4
Number of MHC I peptides
1,875
783
TSA candidates
36
17
Least expressed I/L variants
27
14
Manual inspection of MS/MS spectra
20
10
Genomic location
14
7
B
Type of TSA candidates
Number of TSAs
: ae
: m
10
4
5
2
CT26
EL4
C
mean[log10(rphm+ 1)]
Tumor / mTEChi (rphm)
Tissues with expression > 0 rphm
MHC low
sex-specific
other
brain
testis
mam. gland
ovary
adip. tissue
adrenal gland
colon
duodenum
gonadal fat pad
heart
kidney
large intestine
liver
lung
pancreas
sigmoid colon
small intestine
spleen
stomach
s.c. adip. tissue
thymus
urinary bladder
CT26 aeTSAs
SGPQHQKQQL N/A 22
LPGKVIMDL 20 22
AYQEIKQAL 4,158 21
QFVKKQFNF 17 20
MPHSLLPLVTF N/A 5
SPSYVYHQF / AH1 17,499 3
SPHQVFNL 13,959 1
GYQKMKALL 13,464 1
SGPPYYKGI N/A 0
LPQELPGLVVL N/A 0
EL4 aeTSAs
SSPRGSSTL N/A 10
ATQQFQQL 2,227 2
VNYLHRNV N/A 0
TVPLNHNTL N/A 0
IILEFHSL N/A 0
EL4 Tumor-associated antigens
STLTYSRM 1 22
SMYVPGKL 1 22
VAAANREVL 50 20
NSMVLFDHV 8 20
D
Origin of genuine TSAs
Number of TSAs
Atypical translation events
: coding - in
: coding - out
: non-coding
80%
100%
6/1
5/2
0/2
0/1
0
0
CT26
EL4

Figure 4.2 | Most TSAs derive from the translation of non-coding regions. (**A**) Flowcharts indicating key steps involved in TSA discovery (see **Supplementary Figure 4.3, A to C** for details). I/L: isoleucine/leucine. (**B**) Barplot showing the number of mTSAs (m) and aeTSA candidates (ae) in CT26 and EL4 cells. (**C**) Heatmap showing the average expression of peptide-coding sequences, in reads per hundred million reads sequenced (*rphm*), for aeTSA candidates and EL4 tumor-associated antigens[41,42] in 22 tissues/organs (see **Supplementary Table 4.5**). For each peptide-coding sequence, the expression fold change and the number of positive tissues (*rphm* > 0, bold squares) are presented on the left-hand side of the heatmap. For fold changes, N/A indicates that the corresponding peptide-coding sequence was not expressed in syngeneic mTEC^{hi}. Adip. tissue: adipose tissue, mam. gland: mammary gland and s.c. adip. tissue: subcutaneous adipose tissue. (**D**) Barplots depicting the number of TSAs derived from the translation of non-coding regions (non-coding) and of coding exons in-frame (coding – in) or out-of-frame (coding – out). The number of aeTSAs/mTSAs is reported within bars. The proportion of TSAs derived from atypical translation events is shown above bars. Features of CT26 and EL4 TSAs can be found in **Supplementary Tables 4.6A and B**, respectively.

4.6.3 Protection against EL4 cells following immunization against individual TSAs

We then performed detailed studies on some of the TSAs that seemed most therapeutically promising: those presented by EL4 cells and whose peptide-coding sequence was not expressed by any normal tissue (**Figure 4.2C** and **Supplementary Tables 4.6B** and **4.7**). To assess immunogenicity, C57BL/6 mice were immunized twice with either unpulsed (control group) or TSA-pulsed DCs before being challenged with live EL4 cells. Priming against IILEFHSL or TVPLNHNTL prolonged survival for 10% of mice, with only one TVPLNHNTL-immunized mouse surviving up to day 150 (**Figure 4.3A**). In contrast, the other three TSAs showed superior efficacy, with day-150 survival rates of 20% (VNYIHRNV), 30% (VTPVYQHL) and 100% (VNYLHRNV) (**Figure 4.3, B and C**). To evaluate the long-term efficacy of TSA vaccination, surviving mice were rechallenged with live EL4 cells at day 150 and monitored for signs of disease. The two VNYIHRNV-immunized survivors died of leukemia within 50 days, whereas all others (immunized against TVPLNHNTL, VTPVYQHL or VNYLHRNV) survived the rechallenge (**Figure 4.3**). We conclude that immunization against individual TSAs confers different degrees of protection against EL4 cells (0%-100%), and that in most cases, this protection is long-lasting.

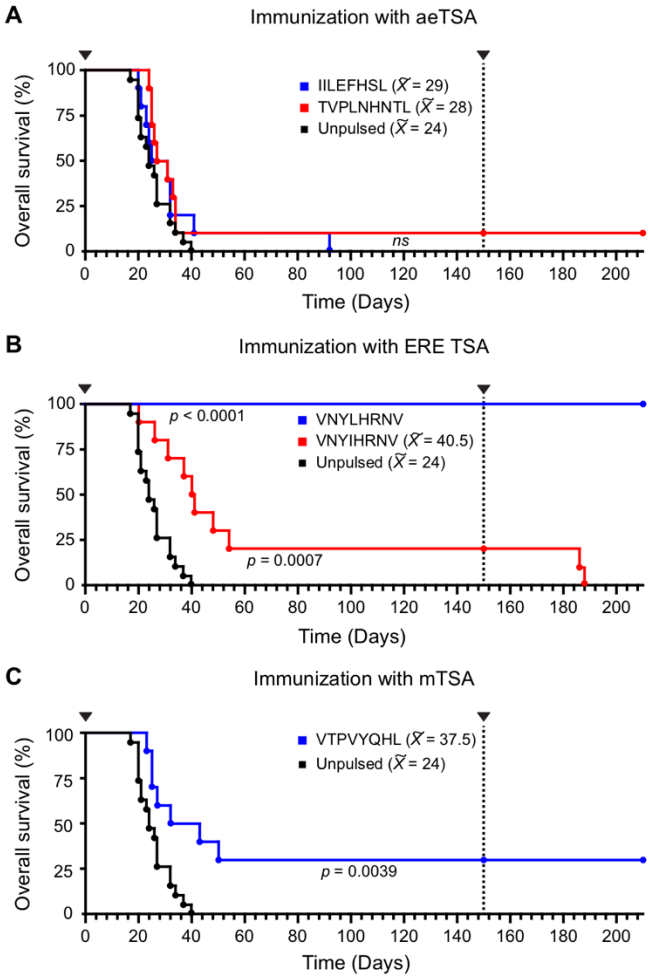


Figure 4.3 | Immunization against individual TSAs confers different degrees of protection against EL4 cells. C57BL/6 mice were immunized twice with DCs pulsed with individual TSAs: (**A**) two aeTSAs, (**B**) two ERE TSAs (one aeTSA or one mTSA), or (**C**) one mTSA. Mice were injected i.v. with 5 x 10^5 live EL4 cells (arrowheads) on day 0 and all surviving mice were rechallenged on day 150. Control groups were immunized with unpulsed DCs (solid black line). $\tilde{X}$ represents the median survival. Statistical significance of immunized vs. control groups was calculated using a log-rank test, where ns stands for not significant ($p > 0.05$). n=10 mice per group for peptide-specific immunization, n=19 mice for control group.

4.6.4 Frequency of TSA-responsive T cells in naive and immunized mice

In various models, the strength of *in vivo* immune responses is regulated by the number of antigen-reactive T cells[45,46]. We therefore assessed the frequency of TSA-responsive T cells in naive and immunized mice using a tetramer-based enrichment protocol[47,48], for which the gating strategy and one representative experiment can be found in **Supplementary Figure 4.5, A to C**. As positive controls, we used tetramers to detect $CD8^+$ T cells specific for three immunodominant viral epitopes (gp-33, M45 and B8R). We confirmed that that these T cells had a high abundance and that their frequency was similar to that observed in previous studies[46] (**Figure 4.4A**). In naive mice, $CD8^+$ T cells specific for TVPLNHNTL, VTPVYQHL and IILEFHSL were rare (less than one $tetramer^+$ cell per 10^6 $CD8^+$ T cells), whereas $CD8^+$ T cells specific for the ERE TSAs (VNYIHRNV and VNYLHRNV) displayed frequencies similar to those of our viral controls (**Figure 4.4A** and **Supplementary Figure 4.6A**). Accordingly, in mice immunized with TSA-pulsed DCs, we found that the T cell frequencies against the two ERE TSAs, as assessed by tetramer staining or IFN-γ ELISpot assays (**Supplementary Figures 4.1B**, **4.5, C and D** and **4.6A**), were higher than that of TVPLNHNTL, VTPVYQHL and IILEFHSL (**Figure 4.4, B and C**). Moreover, in both naive and immunized mice, results obtained with tetramer staining and IFN-γ ELISpot correlated with each other (**Supplementary Figure 4.7**). Finally, we estimated that the functional avidity of T cells specific for VNYIHRNV and VNYLHRNV was similar to that of T cells specific for two highly immunogenic non-self antigens: the minor histocompatibility antigens $H7^a$ and $H13^a$ (**Figure 4.4D**). Hence, these TSAs, derived from allegedly non-coding regions, were recognized by highly abundant T cells with a high functional avidity. This is particularly noteworthy for the VNYLHRNV aeTSA as it has an unmutated germline sequence.

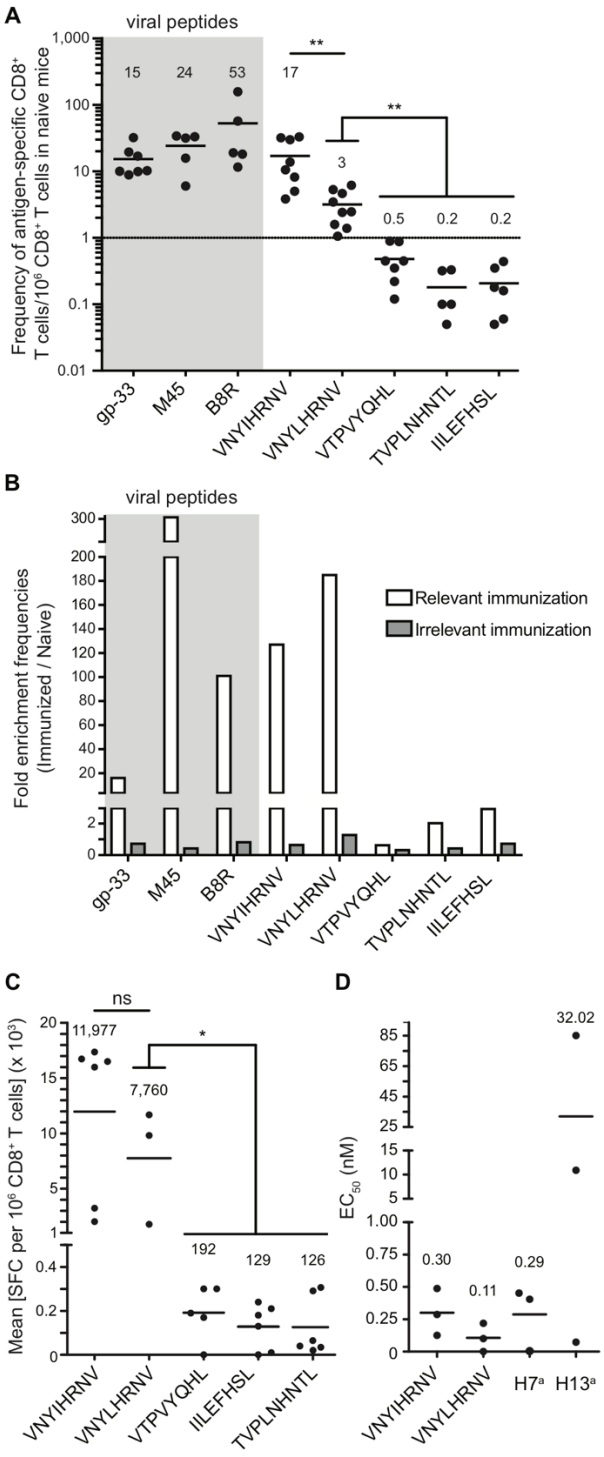


Figure 4.4 | Frequency of and IFN-γ secretion by TSA-responsive T cells in naive and immunized mice. (**A**) Number of tetramer$^+$ CD8$^+$ T cells per 10^6 CD8$^+$ T cells in naive mice. Circle: one mouse (n = 5 to 9 mice per group), dotted line: frequency of 1 tetramer$^+$ T cell per 10^6 CD8$^+$ T cells. (**B**) Fold enrichment of tetramer$^+$ CD8$^+$ T cells after immunization with relevant (white bars) or irrelevant (gray bars) peptides

$\left({Mean\ frequency_{immunized}}/{Mean\ frequency_{naive}}\right)$. (**C**) The number of spot-forming cells (SFCs), measured by an IFN-γ ELISpot assay, averaged across technical replicates (circles) after being converted to SFCs per 10^6 $CD8^+$ T cells: $\left[{(SFCs_{imuunized} - SFCs_{naive})}/{Number\ of\ T\ cells\ plated}\right] \times 10^6$. (**D**) The functional avidity of T cells recognizing specific TSAs and two previously reported positive controls (H7[a] and H13[a] [42]) was estimated by calculating an EC_{50}, corresponding to the peptide concentration were half of plated antigen-specific T cells secreted IFN-γ. (**B-D**) Three independent experiments. On relevant panels, full horizontal lines and numbers above each condition represent mean values. Viral peptides used as control are highlighted in gray. ns: not significant, * = $p \leq 0.05$ and ** = $p \leq 0.01$ (two-sided Wilcoxon rank sum test with the Benjamini-Hochberg correction).

Taken together, our results show that the frequency of TSA-responsive T cells was a crucial parameter for TSA immunogenicity. However, VTPVYQHL was an outsider: it afforded the second-best protection against EL4 challenge even though its cognate T cells were present at a very low frequency (**Figures 4.3** and **4.4, A to C**). In order to better evaluate the importance of T cell expansion in leukemia protection, we estimated the frequency of tetramer$^+$ $CD8^+$ T cells in long-term survivors following rechallenge with EL4 cells on day 150 (**Figure 4.3**). These analyses were performed on day 210 or at the time of sacrifice (in the case of VNYIHRNV-primed mice). Two points can be made from these analyses (**Supplementary Figure 4.6, B and C**). First, all long-term survivors, including VTPVYQHL-immunized mice, showed a discernable population of TSA-responsive (tetramer$^+$) $CD8^+$ T cells. Second, although VNYIHRNV was recognized by a particularly large population of tetramer$^+$ cells, it was the only TSA that did not protect mice upon rechallenge. Altogether, our results suggest that expansion of TSA-responsive T cells was necessary for protection against EL4 cells, but was insufficient in the case of VNYIHRNV.

4.6.5 The importance of antigen expression for protection against EL4 cells

Next we evaluated the impact of antigen expression on immunogenicity by assessing the abundance of TSAs at the RNA level in the EL4 cell population that was injected on day 0 (**Figure 4.3**). The sequence encoding the TSA conferring the best protection against EL4 cells (VNYLHRNV) was expressed more than the other TSA-coding sequences (**Figure 4.5A**). This suggests that VNYLHRNV is likely "clonal" (expressed by all EL4 cells) and highly expressed, whereas the other TSAs are sub-clonal and/or expressed at low amounts. Next, using parallel reaction monitoring (PRM) MS, we analyzed the TSA copy number per cell in the EL4 cell population used for rechallenge (day 150, **Figure 4.5B**). As expected[41], there was no linear relationship between TSA abundance at the RNA and peptide levels (**Figure 4.5, A and B**). Notably, the most protective TSA, VNYLHRNV, was one of the two most abundant TSAs (> 500 copies per cell), whereas VNYIHRNV, which offered no protection upon rechallenge (**Figure 4.3B**), was no longer detected on EL4 cells. This observation suggests that VNYIHRNV was a sub-clonal TSA and that antigen loss most likely explained the lack of protection upon rechallenge. Finally, we noted that TSAs were immunogenic when presented by DCs but not when presented by EL4 cells: injection of live EL4 cells without prior immunization did not induce an expansion of TSA-responsive T cells (**Figure 4.5C** and **Supplementary Figure 4.6D)**, and immunization with irradiated EL4 cells did not confer any protection against live EL4 cells (**Figure 4.5D**). This suggests that, in the absence of immunization, highly immunogenic TSAs (such as VNYLHRNV) were ignored because they were not efficiently cross-presented by DCs, highlighting the importance of efficient T-cell priming in cancer immunotherapy.

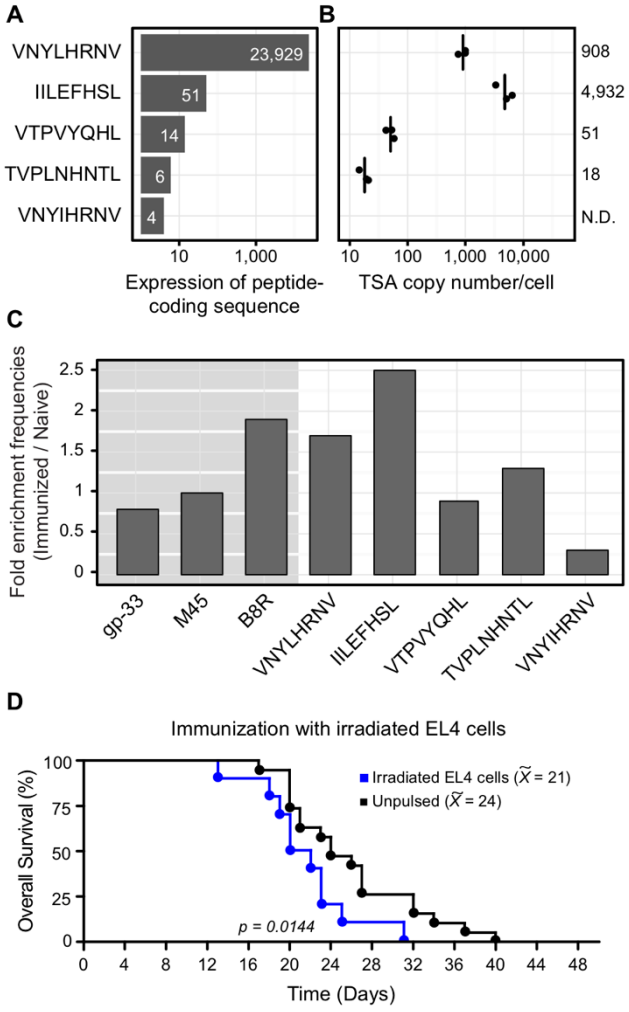


Figure 4.5 | High expression of EL4 TSAs is necessary but not sufficient to induce antileukemic responses. (**A** and **B**) Analysis of TSA expression at the RNA and peptide levels was performed on EL4 cells injected into mice at day 0 or 150, respectively. (**A**) The number of RNA-Seq reads fully overlapping the RNA sequences encoding each TSA. (**B**) TSA copy number per cell was estimated by PRM-MS using ^{13}C-synthetic peptide analogs of the TSAs (three replicates). Black lines represent the mean TSA copy number per cell (also indicated on the left-hand side of the graph). N.D.: not detected. (**C**) Fold enrichment for tetramer$^+$ CD8$^+$ T cells after injection with live EL4 cells without prior immunization $\left(Mean\ frequency_{EL4-injected} \middle/ Mean\ frequency_{naive}\right)$. Fold enrichment for T cells recognizing viral peptides are shown as negative controls and are highlighted in gray. Three independent experiments were performed. (**D**) Overall survival of C57BL/6 female mice immunized twice with irradiated (10,000 cGy) EL4 cells (blue line, n=10 mice) or unpulsed DCs (black line, n=19 mice), then injected i.v. with 5 x 10^5 live EL4 cells. $\tilde{X}$ represents the median survival. Statistical significance of immunized vs. control groups was calculated using a log-rank test.

4.6.6 Impact of non-coding regions on the TSA landscape of human primary tumors

Having established that non-coding regions are a major source of TSAs in two murine cell lines, we applied our proteogenomic approach to identify MHC I peptides from seven human primary tumor samples: four B-lineage ALLs and three lung cancers (**Supplementary Figure 4.8** and **Supplementary Tables 4.8 to 4.15**). Rather than using RNA-Seq data from murine syngeneic mTEC^{hi}, we sequenced the transcriptome of total TECs (n = 2) and purified mTECs (n = 4) from six unrelated donors undergoing corrective cardiovascular surgery (**Supplementary Table 4.2B**). Notably, we found minimal inter-individual differences and demonstrated that this cohort size was sufficient to cover almost the full breadth of the mTEC transcriptomic landscape. Indeed, computing the cumulative number of detected transcripts showed that minimal gains would be achieved by adding more samples (**Supplementary Figure 4.9**). Using these RNA-Seq data as the repertoire of normal k-mers to generate global cancer databases (**Supplementary Table 4.1B**) as described in **Figure 4.1,** we identified two mTSAs and 27 aeTSA candidates (**Figure 4.6A**). After validating their assignment to a single genomic location and the identity of their MS/MS spectrum, we also ensured that mTSAs did not intersect with known germline polymorphisms (**Supplementary Figures 4.3, 4.10** and **4.11**). In order to further validate the status of aeTSA candidates, we did as for murine aeTSAs (**Figure 4.2C**), we analyzed the expression of aeTSA coding sequences in RNA-Seq data from 28 tissues (6-50 individuals per tissue, **Figure 4.6B** and **Supplementary Table 4.16**). Based on these data, we excluded six aeTSA candidates: three that were widely expressed, like most previously reported overexpressed tumor-associated antigens[49], and three that were expressed at substantial amounts in a single organ, the liver (**Figure 4.6B**). We therefore ended up with a total of two mTSAs and 20 non-redundant aeTSAs (**Figure 4.6C** and **Supplementary Tables 4.17** and **4.18**). Of note, the SLTALVFHV aeTSA was shared by our two HLA-A*02:01-positive ALLs (**Supplementary Table 4.17**). This aeTSA derives from the 3'UTR of *TCL1A*, a gene implicated in lymphoid malignancies.

Altogether, our results show that our proteogenomic approach can characterize the repertoire of mTSAs and aeTSAs on individual tumors in about two weeks.

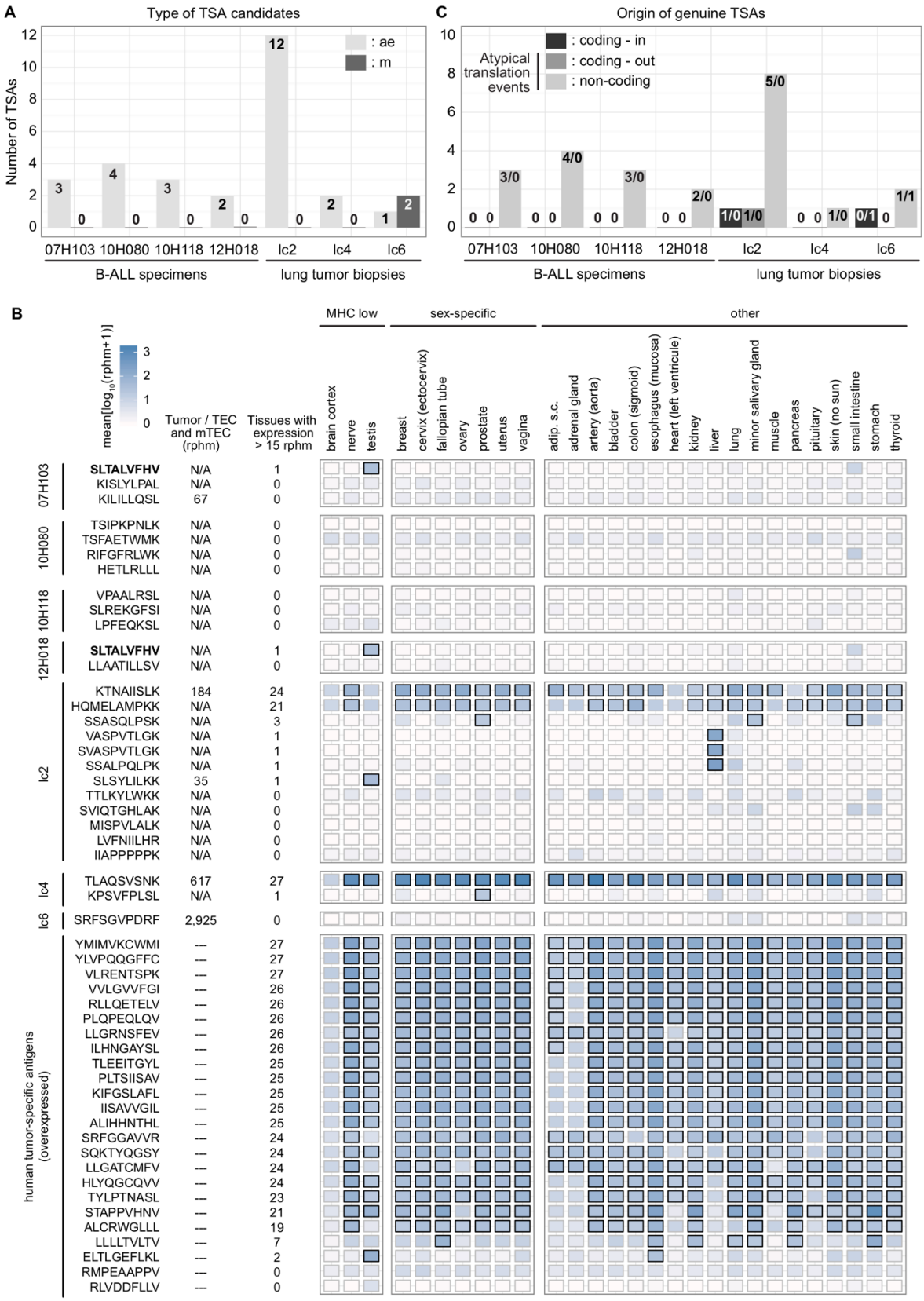

A
Type of TSA candidates
Number of TSAs
: ae
: m
07H103 10H080 10H118 12H018 lc2 lc4 lc6
B-ALL specimens
lung tumor biopsies
C
Origin of genuine TSAs
Atypical translation events
: coding - in
: coding - out
: non-coding
B
mean[log10(rphm+1)]
Tumor / TEC and mTEC (rphm)
Tissues with expression > 15 rphm
MHC low
sex-specific
other
brain cortex
nerve
testis
breast
cervix (ectocervix)
fallopian tube
ovary
prostate
uterus
vagina
adip. s.c.
adrenal gland
artery (aorta)
bladder
colon (sigmoid)
esophagus (mucosa)
heart (left ventricule)
kidney
liver
lung
minor salivary gland
muscle
pancreas
pituitary
skin (no sun)
small intestine
stomach
thyroid
human tumor-specific antigens (overexpressed)

Figure 4.6 | Most TSAs detected in human primary tumors derive from the translation of non-coding regions. (**A**) Barplot showing the number of aeTSAs candidates (ae) and mTSAs (m) in each primary sample analyzed. (**B**) Heatmap showing the average expression of peptide-coding sequences, in *rphm,* for aeTSAs and overexpressed tumor-associated antigens obtained from Cancer Immunity Peptide database[49] across a panel of 28 tissues (see **Supplementary Table 4.16**). For each peptide-coding sequence, the expression fold change (Tumor/TEC and mTEC) and the number of positive tissues (*rphm* > 15, bold squares) are shown on the left-hand side of the heatmap. For fold changes, N/A and --- indicate that the corresponding peptide-coding sequence was not detected in TEC/mTEC samples or not computed, respectively. Adip. s.c.: adipose subcutaneous. (**C**) Barplots depicting the number of TSAs derived from the translation of non-coding regions (non-coding) or from coding exons translated in-frame (coding – in) or out-of-frame (coding – out). Within bars is shown the number of aeTSAs/mTSAs. Features of human TSAs identified in each sample can be found in **Supplementary Tables 4.17** and **4.18**.

4.7 Discussion

In order to explore the global landscape of TSAs, we developed a proteogenomic approach that incorporates two features in the construction of databases for MS analyses: alignment-free k-mer profiling of RNA-Seq data and subtraction of mTEC k-mers. In a context where TSA discovery is a critical unmet medical need, our approach led us to discover that the TSA landscape is much larger than previously anticipated. Indeed, 35 out of the 39 non-redundant TSAs reported here derived from atypical translation events: two from the out-of-frame translation of coding exons and 33 from allegedly non-coding regions. Hence, ~ 90% of our TSAs would have been missed by standard approaches focusing on exonic mutations. In addition to MHC peptides derived from RNAs containing single-base mutations, our approach efficiently captured peptides generated by complex structural variants, as exemplified by VTPVYQHL, which derived from a large intergenic deletion (~7,500 bp) in EL4 cells. Subtraction of mTEC k-mers was critical for the high-throughput identification of immunogenic aeTSAs including unmutated peptides that are not constitutively presented by $mTEC^{hi}$ to thymocytes during the establishment of central tolerance. This is well-illustrated by VNYLHRNV, an unmutated TSA absent from $mTEC^{hi}$ and other peripheral tissues, although strongly expressed in EL4 cells, that is recognized by highly abundant $CD8^+$ T cells with a high functional avidity. Nonetheless, a few peptide-coding transcripts undetectable in $mTEC^{hi}$ were detected in peripheral tissues, suggesting that k-mers from both mTECs and peripheral tissues should be used in order to identify genuine TSAs. In the present study, we chose to use peripheral expression as an *a posteriori* validation step. An alternative approach would be to remove all k-mers expressed in peripheral tissues when building the database for MS.

TSAs derived from non-coding regions present a number of peculiar and highly relevant features. First, it is evident that EREs are a rich source of TSAs; they generated 9 of the 17 TSAs found in murine cell lines and 4 of the 23 TSAs in human tumors. The difference in the proportion of ERE TSAs we identified in murine cell lines vs. primary human tumors might be ascribed to the fact that in vitro culture conditions do not recapitulate the immune pressure exerted on developing tumors, that ERE

expression greatly varies across tumor types[50] or that human EREs are more degenerate and therefore less likely to be translated than murine EREs[9]. Nonetheless, ERE TSAs remain particularly relevant to the development of cancer vaccines, as both oncogenic viruses and viral-like sequences in the human genome appear to be particularly immunogenic[51,52]. Second, most TSAs derived from non-coding regions do not overlap mutations and are therefore, by definition, aeTSAs. Such aeTSAs, which include EREs, present a major advantage over mTSAs (of both coding or non-coding origin): whereas mTSAs are private antigens, aeTSAs can be shared by multiple tumors[8]. Indeed, we were able to identify such shared aeTSA (SLTALVFHV) in humans, while Probst *et al.* showed that mice immunized against the AH1 aeTSA (SPSYVYHQF), that we identified by MS on CT26 cells, survived the challenge with three different cancer cell lines: the WEHI-164 fibrosarcoma as well as the C51 and CT26 colorectal cancers[10].

Altogether, our MS-based discovery of TSAs in primary human B-ALL and lung cancer demonstrates that the impetus to develop TSA vaccines should not be limited to cancers with a high mutational load. Indeed, because ALLs harbor very few exonic mutations, it was presumed that they might not express any TSA (*5).* Our data show that TSAs can be found in ALL provided that the search strategy encompasses aeTSAs. Moreover, two elements argue that TSAs derived from atypical translation events represent promising targets for T cell-based cancer immunotherapy: i) they outnumber TSAs derived from coding regions and ii) they are mostly unmutated, which increases their potential to be shared between patients.

We acknowledge that our study presents several limitations for which solutions can be envisioned. First, as our approach is not compatible with the computation of classical false discovery rates, TSAs must undergo meticulous validation by manual inspection of MS spectra and, ideally, confirmation using synthetic peptide analogs. Second, our approach relies on shotgun mass spectrometry, which suffers from a limited dynamic range. Consequently, we only detected the most abundant TSAs and are likely underestimating the extent of aeTSA sharing between patients. Because shared aeTSAs may represent promising actionable targets for cancer

immunotherapy[8], aeTSA frequency across patients and/or tumor types could be further evaluated by targeted MS analyses which have a more limited scope, but are 10-times more sensitive than shotgun MS and can provide quantitative data such as the number of TSA copies per cell[53]. Third, because TSA immunogenicity cannot be predicted[54], it has to be tested experimentally for each TSA. This issue is being addressed by several research groups that are currently developing artificial platforms requiring less material than the conventional IFN-γ ELISpot assays used for such testing[8,55,56].

In practice, how could we prioritize TSAs for clinical trials? In our EL4 tumor model, the efficacy of TSA immunization was largely determined by two criteria: TSA abundance and the frequency of TSA-responsive T cells. TSA abundance can be assessed by targeted MS analyses, and the frequency of TSA-responsive T cells in PBMCs of a cohort of subjects could be estimated using MHC-peptide multimers or functional assays. Widely shared, highly abundant TSAs recognized by high frequency T cells could then be selected for clinical trials. These optimal aeTSAs could then potentially be combined in a single vaccine using already available synthesis and delivery platforms[13,57].

4.8 Methods

4.8.1 Study design

The purpose of this study was to develop a proteogenomic approach that would enable the identification of TSAs derived from any region of the genome, and to identify features that influenced TSA immunogenicity. To do so, we first characterized the TSA landscape of two murine cell lines, the EL4 T-lymphoblastic lymphoma cell line and the CT26 colorectal cancer cell line that were both obtained from the American Type Culture Collection. As a normal control, we used thymi isolated from 5–8 week-old C57BL/6 or Balb/c mice obtained from The Jackson Laboratory. Mice were housed under specific pathogen-free conditions and all experimental protocols were approved by the Comité de Déontologie de l'Expérimentation sur des Animaux of Université de Montréal. We also applied our approach to seven human primary cancer samples from treatment-naive patients. These included three lung tumor biopsies (lc2, lc4 and lc6) purchased from Tissue Solutions and four primary leukemic samples (B-ALL specimens 07H103, 10H080, 10H118 and 12H018) that were collected and cryopreserved at the Banque de Cellules Leucémiques du Québec at Hôpital Maisonneuve-Rosemont. The project was approved by the Comité d'éthique de la recherche de l'Hôpital Maisonneuve-Rosemont (CÉR 12100). NOD Cg-Prkdc^{scid}Il2rg^{tm1Wjl}/SzJ (NSG) mice were used to expand our B-ALL specimens. These mice were purchased and housed as described for C57BL/6 and Balb/c mice. As a normal control, we used thymi obtained from 3-month-old to 7-year-old individuals undergoing corrective cardiovascular surgery (CHU Saint Justine Research Ethic Board, protocol and biobank #2126). No statistical method was used to predetermine sample size. One replicate was sequenced for all RNA-Seq experiments. For mass spectrometry, at least two replicates were analyzed. To assess the immunogenicity of EL4-derived TSAs, we measured the frequency and antigen avidity of T cells recognizing TSAs. In addition, we estimated the survival of 8-12 week old C57BL/6 female mice that were immunized or not with individual TSAs. Investigators were not blinded during sample preparation or during data collection and analysis. For all *in vitro*

and *in vivo* experiments described in this manuscript, at least three replicates were analyzed and found to be concordant with each other. No data were excluded from the analyses and values are reported in **Supplementary Table 4.7**. The number of mice used, numbers of replicates and statistical values (where applicable) are provided in the figure legends. For information regarding original RNA-Seq and MS data, see **section 4.9.2.4** and **Supplementary Table 4.2**.

4.8.2 Statistical analysis

Procedures to evaluate statistical significance are described in the relevant figure legends. Overall, a log-rank test was used for survival curves, a Wilcoxon rank sum test with the Benjamini-Hochberg correction for multiple testing was used to compare T-cell frequencies as estimated by tetramer and ELISpot and a one-sided Wilcoxon rank sum test was used to compare T-cell frequencies between immunized and rechallenged mice. $P \leq 0.05$ was considered significant.

4.8.3 Cell lines

The EL4 T-lymphoblastic lymphoma cell line, the CT26 colorectal cancer cell line and the B-cell hybridoma HB-124 were obtained from the American Type Culture Collection. EL4 and CT26 cells were cultured in RPMI 1640/HEPES supplemented with 10% heat-inactivated fetal bovine serum, 1% L-glutamine and 1% penicillin-streptomycin. Cell culture media were further supplemented with 1% non-essential amino acids and 1% sodium-pyruvate or 1% sodium-pyruvate only for EL4 and CT26 cells, respectively. To produce the anti-CDR2 antibody, HB-124 cells were cultured in IMDM supplemented with 10% heat-inactivated fetal bovine serum. Unless stated otherwise, all reagents were purchased from Gibco.

4.8.4 Human primary samples

Primary leukemic samples were expanded in vivo after transplantation in NSG mice as previously described[59]. Briefly, 1-2 x 10^6 B-ALL cells were thawed and transplanted via i.v. injection into 8–12 week-old sub-lethally irradiated (250 cGy, 137

Cs-gamma source) NSG mice. Mice were sacrificed at signs of disease and cell suspensions were prepared from mechanically disrupted spleens or, for 07H103, from a mix of splenocytes, bone marrow and peritoneal ascites. From there, Ficoll gradients were used to enrich for B-ALL cells prior to isolation of MHC peptides (see **section 4.8.12**). After Ficoll gradient, purity and viability of each B-ALL sample were assessed using flow cytometry and one representative experiment is shown in **Supplementary Figure 4.10**. Briefly, 0.5 x 10^6 cells were stained with Pacific Blue anti-human CD45 (BioLegend), PE-Cy7 anti-human CD19 (BD Bioscience), APC-eFluor780 anti-mouse CD45.1 (eBioscience) and 7-aminoactinomycin D (7-AAD, BD Biosciences). B-ALL cells were defined as 7-AAD$^-$huCD45$^+$huCD19$^+$. Data acquisition was performed on a BD Canto II cytometer (BD Bioscience). The analysis was done with BD FACSDiva 4.1 software. For all samples, HLA typing was obtained using OptiType version 1.0, running with default parameters for RNA-Seq data (see **section 4.8.8**).

4.8.5 Peptides

Native and ^{13}C-labelled versions of tumor-specific antigens (TSAs) were synthesized by GenScript. For the ^{13}C-analogs, the labelled amino acids are underlined in the following list: VNY<u>I</u>HRNV, VNY<u>L</u>HRNV, TVP<u>L</u>NHNTL, VTPVYQH<u>L</u>, II<u>L</u>EFHSL. Purity, as determined by the manufacturer, was greater than 95% and 75% for native and ^{13}C-labelled peptides, respectively.

4.8.6 Murine mTEC^{hi} extraction

Thymi were isolated from 5–8 week-old C57BL/6 or Balb/c mice and mechanically disrupted to extract thymocytes. Stromal cell enrichment was performed as previously described[60]. Thymic stromal cells were stained with biotinylated Ulex europaeus lectin 1 (UEA1; Vector Laboratories), PE-Cy7–conjugated streptavidin (BD Biosciences), and the following antibodies: Alexa Fluor 700 or FITC anti-CD45, PE anti–I-A^b (BD Biosciences), Alexa Fluor 700 I-A/I-E, APC-Cy7 anti-EpCAM (BioLegend). Cell viability was assessed using 7-AAD (BD Biosciences). Live mature medullary thymic epithelial cells (mTEC^{hi}) were gated as 7-AAD$^-$ CD45$^-$ EpCAM$^+$ UEA1$^+$

MHC II^{hi} and sorted on a FACS AriaIIIu (BD Biosciences, **Supplementary Figure 4.1A**).

4.8.7 Human TEC and mTEC extraction

Thymi obtained from 3-month-old to 7-year-old individuals were kept at 4ºC in 50 ml conical tubes containing media and cut in 2-5 mm cubes within hours following their surgical resection. For long-term preservation, thymic cubes were frozen in cryovials containing heat-inactivated human serum / 10% DMSO and kept in liquid nitrogen for a maximum of 3 years. Cryopreserved thymic samples were transferred to our laboratory on dry ice and used to isolate human TEC and mTEC following a protocol adapted from C. Stoeckle *et al.*[61]. Thymic tissue was cut into small fragments, then digested at 37°C using a solution of 2 mg/mL Collagenase A (Roche) and 0.1 mg DNase I/ml (Sigma-Aldrich) in RPMI-1640 (Gibco) for three to five periods of 40 min. After the second digestion, a solution of Trypsin/EDTA (Gibco) was added, for which the activity was neutralized by adding FBS (Invitrogen) 15 min before the end of incubation. For TEC and mTEC sorting (**Supplementary Figure 4.1B**), cell suspensions were stained with Pacific blue-conjugated anti-CD45 (BioLegend), PE-conjugated anti-HLA-DR (BioLegend), APC-conjugated anti-EpCAM (BioLegend), Alexa 488-conjugated anti-CDR2 (produced in our lab with the HB-124 hybridoma – see **section 4.8.3** – and conjugated with the Dylight 488 Fast conjugation kit from abcam, only for mTEC samples) and cell viability was assessed using 7-AAD (BD Biosciences).

4.8.8 RNA extraction, library preparation and sequencing

For EL4 and CT26 cells, one replicate of 5 x 10^6 cells was used to perform RNA-Seq. For C57BL/6 and Balb/c mTEC^{hi}, RNA-Seq was performed in triplicate on a minimum of 31,686 or 16,338 FACS-sorted cells extracted from 2 females and 2 males. For primary leukemic cells, RNA-Seq was performed on a single replicate of 2.0 to 4.0 x 10^6 cells. For human TEC and mTEC, we performed one RNA-Seq replicate per donor with 33,076 to 84,198 FACS-sorted TECs or 50,058 to 100,719 mTECs. In all

cases, total RNA was isolated using TRIzol (Invitrogen), further purified using the RNeasy kit or RNeasy micro kit (Qiagen) as recommended by each manufacturer. For each lung tumor biopsy (three in total), total RNA was isolated from ~30 mg of tissues using the AllPrep DNA/RNA/miRNA Universal kit (Qiagen) as recommended by the manufacturer and was used to perform one replicate of RNA-Seq per sample. Each murine sample (EL4, CT26 and murine mTEC^{hi}) were quantified on a Nanodrop 2000 (Thermo Fisher Scientific) and RNA quality was assessed on a 2100 Bioanalyzer (Agilent Genomics) in order to select samples with an RNA integrity number ≥ 9. For human samples (B-ALLs, lung tumor biopsies and human TEC/mTEC), quantification of total RNA was made by QuBit (ABI) and quality of total RNA was assessed with the 2100 BioAnalyzer (Agilent Genomics) in order to select samples with an RNA integrity number ≥ 7. cDNA libraries were prepared from 2-4 µg for EL4 and CT26 cells, 50-100 ng for murine mTEC^{hi}, 500 ng for B-ALLs specimens, 4 µg for lung tumor biopsies, 8-13 ng for human TECs or 41-68 ng for human mTECs of total RNA using the TruSeq Stranded Total RNA Library Prep Kit (EL4 cells), KAPA Stranded mRNA-Seq Kit (CT26 cells, C57BL/6 mTEC^{hi}, human mTEC, lung tumors and B-ALL specimens) or KAPA RNA HyperPrep Kit (Balb/c mTEC^{hi}, human TEC). These libraries were further amplified by 9-16 cycles of PCR before sequencing. Paired-end RNA-Seq was performed on an Illumina NextSeq 500 (Balb/c mTEC^{hi}, human TEC and mTEC) or HiSeq 2000 (any other sample) and yielded an average of 175 and 199 x 10^6 reads per murine and human sample, respectively.

4.8.9 Generation of canonical cancer and normal proteomes

For all samples, RNA-Seq reads were trimmed for sequencing adapters and low quality 3' bases using Trimmomatic version 0.35 and then aligned to the reference genome, GRCm38.87 for murine samples and GRCh38.88 for human samples, using STAR version 2.5.1b[62] running with default parameters except for --alignSJoverhangMin, --alignMatesGapMax, --alignIntronMax, and --alignSJstitchMismatchNmax parameters for which default values were replaced by 10, 200,000, 200,000 and “5 -1 5 5”, respectively. Single-base mutations with a minimum alternate count setting of 5 were identified using freeBayes version 1.0.2-16-gd466dde

(arXiv:1207.3907) and exported in a VCF, which was converted to an agnostic single-nucleotide polymorphism file format compatible with pyGeno[63]. Finally, transcript expression was quantified in transcripts per million (tpm) with kallisto version 0.43.0 (https://pachterlab.github.io/kallisto/about) running with default parameters. Of note, kallisto index was constructed using the index functionality and using the appropriate *.cdna.all.fa.gz files downloaded from Ensembl[64]. To build each sample's canonical proteome, we used pyGeno to (i) insert high-quality sample-specific single-base mutations (freeBayes quality > 20) in the reference exome, thereby creating a personalized exome, and to (ii) export sample-specific sequence(s) of known proteins generated by expressed transcripts (tpm > 0). These protein sequences were written to a fasta file that was subsequently used for mass spectrometry (MS) database searches (Cancer canonical proteome) and/or MHC peptide classification (Cancer and normal canonical proteome). See **Figure 4.1A** for a schematic and **Supplementary Table 4.1** for statistics.

4.8.10 Generation of cancer and normal k-mer databases

For all cancer and normal samples, both R1 and R2 fastq files were independently downloaded and trimmed for sequencing adapters and low quality 5' and 3' bases using Trimmomatic version 0.35. To ensure that all reads were on the transcript-encoding strand, R1 reads were reverse complemented using the **fastx_reverse_complement** function of the FASTX-Toolkit version 0.0.14. Using Jellyfish version 2.2.3[34], we then generated 33- and 24-nucleotide-long k-mer databases, required for k-mer profiling and MHC peptide classification, respectively. See **Supplementary Figure 4.2A** for details. Of note, when multiple biological replicates (murine mTEC^{hi}) or when multiple samples from unrelated donors (human TEC and mTEC) were available, fastq files were concatenated to generate a single normal k-mer database per condition (C57BL/6, Balb/c or human).

4.8.11 k-mer filtering and generation of cancer-specific proteomes

To extract 33-nucleotide-long k-mers that could give rise to TSAs, we applied a sample-specific threshold on k-mer occurrence in order to exclude sequencing errors: at least 4 times in EL4 or CT26 cells, 7 times in lung tumor biopsies and 10 times in primary leukemic samples. Cancer-specific k-mers were then obtained by excluding those expressed in the relevant murine $mTEC^{hi}$ or human TEC/mTEC k-mer database (**Supplementary Figure 4.2B**). This cancer-specific k-mer set was further assembled into longer sequences, called contigs. Briefly, one of the submitted 33-nucleotide-long k-mer is randomly selected to be used as a seed that is then extended from both ends with consecutive k-mers overlapping by 32 nucleotides on the same strand (-r option disabled, as we were working with stranded sets of k-mers). The assembly process stops when either no k-mers can be assembled, meaning that no 32-nucleotide-overlapping k-mer can be found, or when more than one k-mer fits (–a 1 option for linear assembly). If so, a new seed is selected and the assembly process resumes until all k-mers from the submitted list have been used once (**Supplementary Figure 4.2C**). This step is done by the **kmer_assembly** tool from NEKTAR, an in-house developed software. To obtain amino acid sequences, we 3-frame translated contigs that were at least 34 nucleotides long using an in-house python script. Cancer-specific amino acid sequences were then split at internal stop codons and resulting subsequences of at least 8 amino acids long were given a unique ID before being included in the relevant cancer-specific proteome (**Supplementary Figure 4.2D**). See **Figure 4.1B** for schematic and **Supplementary Table 4.1** for statistics.

4.8.12 Isolation of MHC peptides

For EL4 and CT26 cells, three biological replicates of 250 x 10^6 cells were prepared from exponentially growing cells. For all primary leukemic samples, three biological replicates of ~450 to 700 x 10^6 cells were prepared from freshly harvested leukemic cells (see **section 4.8.4**). MHC peptides were obtained as previously described[65], with minor modifications: following mild acid elution (MAE), peptides were

desalted on an Oasis HLB cartridge (30 mg, Waters) and filtered on a 3 kDa molecular weight cut-off (Amicon Ultra-4, Millipore) to remove β2m proteins. For one of our primary leukemic samples (specimen 10H080), we prepared four additional replicates of 100 x 10^6 cells and isolated MHC peptides by immunoprecipitation (IP) as previously described[59]. Finally, lung tumor biopsies (wet weight ranging from 771 to 1,825 mg, see **section 4.8.1**) were cut in small pieces (cubes of ~3 mm in size) and 5 ml of ice-cold PBS containing protein inhibitor cocktail (Sigma) was added to each tissue sample. Tissues were first homogenized twice using an Ultra Turrax T25 homogenizer (20 seconds at 20,000 rpm, IKA-Labortechnik) and then once using an Ultra Turrax T8 homogenizer (20 seconds at 25,000 rpm, IKA-Labortechnik). Then, 550 µl of ice-cold 10X lysis buffer (10% w/v CHAPS) was added to each sample and MHC peptides were immunoprecipitated as previously described[59] using 1 mg (1 ml) of covalently cross-linked W6/32 antibody to protein A magnetic beads per sample. Regardless of the isolation technique, peptide extracts were all dried using a Speed-Vac and kept frozen prior to MS analyses.

4.8.13 Mass spectrometry analyses

Dried peptide extracts were re-suspended in 0.2 % formic acid. For EL4 and CT26, peptide extracts were loaded on a home-made C_{18} pre-column (5 mm x 360 μm i.d. packed with C_{18} Jupiter Phenomenex) and separated on a home-made C_{18} analytical column (15 cm x 150 µm i.d. packed with C_{18} Jupiter Phenomenex) with a 56-min gradient from 0–40 % acetonitrile (0.2 % formic acid) and a 600 nl.min^{-1} flow rate on a nEasy-LC II system. For all human samples, peptide extracts were loaded on a home-made C_{18} analytical column (15 cm x 150 µm i.d. packed with C_{18} Jupiter Phenomenex) with a 56-min gradient from 0–40 % acetonitrile (0.2 % formic acid, 07H103, 10H080-MAE, 10H118 and 12H018) or with a 100-min gradient from 5–28 % acetonitrile (0.2 % formic acid, lung tumor biopsies and 10H080-IP) and a 600 nl.min-1 flow rate on a nEasy-LC II system. Samples were analyzed with a Q-Exactive Plus (EL4, Thermo Fisher Scientific) or HF (all other samples, Thermo Fisher Scientific). For the Q-Exactive Plus, each full MS spectrum, acquired with a 70,000 resolution, was

followed by 12 MS/MS spectra, where the most abundant multiply charged ions were selected for MS/MS sequencing with a resolution of 17,500, an automatic gain control target of 1e6, an injection time of 50 ms and a collision energy of 25 %. For the Q-Exactive HF, each full MS spectrum, acquired with a 60,000 resolution, was followed by 20 MS/MS spectra, where the most abundant multiply charged ions were selected for MS/MS sequencing with a resolution of 15,000 (CT26, 07H103, 10H080-MAE, 10H118, 12H018) or 30,000 (lung tumor biopsies, 10H080-IP), an automatic gain control target of 5e4, an injection time of 100 ms and a collision energy of 25 %. Peptides were identified using Peaks 8.5 (Bioinformatics Solution Inc.) and peptide sequences were searched against the relevant global cancer database, obtained by concatenating the canonical cancer proteome and cancer-specific proteome (see **sections 4.8.9** and **4.8.11**). For peptide identification, tolerance was set at 10 ppm and 0.01 Da for precursor and fragment ions, respectively. Occurrence of oxidation (M) and deamidation (NQ) were considered as post-translational modifications.

4.8.14 Identification of MHC peptides

To select for MHC peptides, lists of unique identifications obtained from Peaks were filtered to include 8–11-amino-acid-long peptides that had a percentile rank ≤ 2% for at least one on the relevant MHC I molecules, as predicted by NetMHC 4.0[66]. We did not compute false discovery rates on these lists since we wanted to identify as many TSAs as possible. However, we applied a sample-specific threshold on the Peaks score to guaranty that our list of MHC peptides only included 5% of decoy identifications $\left(\%_{decoy} = \left({}^{\#_{decoys}}\!/_{\#_{targets}}\right) \times 100\right)$. MHC peptides identified in each samples are reported in **Supplementary Tables 4.3, 4.4** and **4.8 to 4.15**.

4.8.15 Identification and validation of TSA candidates

To identify TSA candidates, each MHC peptide and its coding sequence were queried to the relevant cancer and normal canonical proteomes or cancer and normal 24-nucleotide-long k-mer databases, respectively. Here, the normal canonical proteome and normal 24-nucleotide-long k-mer database were built using (i) RNA-Seq

data from syngeneic mTEC^{hi} for EL4 and CT26 cells and (ii) RNA-Seq data from two TECs and four mTECs samples for all human tumor samples. MHC peptides detected in the normal canonical proteome were excluded regardless of their coding sequence detection status, as they are likely to be tolerogenic. MHC peptides that were truly cancer-specific, in other words, which were neither detected in the normal canonical proteome nor in normal k-mers, were flagged as TSA candidates. MHC peptides absent from both canonical proteomes but present in both k-mer databases needed to have their RNA coding sequence overexpressed by at least 10-fold in cancer cells compared to normal cells in order to be flagged as such (**Supplementary Figure 4.3A**). Finally, MHC peptides corresponding to several RNA sequences (derived from different proteins) could only be flagged as TSA candidate if their respective coding sequences were concordant, that is, if those consistently flagged the relevant peptide as a TSA candidate. MS/MS spectra of all TSA candidates were manually inspected to remove any spurious identifications. Besides, sequences presenting with multiple genomically possible I/L variants were further inspected to report both variants when they were distinguishable by MS, or only the most expressed variant when they were not (**Supplementary Figure 4.3B**). Finally, we assigned a genomic location to all those MS-validated TSA candidates by mapping reads containing MHC peptide-coding sequences on the reference genome (GRCm38.87 or GRCh38.88) using BLAT (tool from the UCSC genome browser). TSA candidates for which reads did not match to a concordant genomic location or which matched to hypervariable regions (such as the MHC, Ig or TCR genes) or multiple genes were excluded. For those with a concordant genomic location, we used IGV[67] to exclude TSA candidates with a MHC peptide-coding sequence overlapping synonymous mutations with regard to their relevant normal counterpart or, for human TSA candidates, those overlapping a known germline polymorphism (listed in dbSNP v. 149, **Supplementary Figure 4.3C**). Remaining peptides were classified as mTSAs or aeTSA candidates, depending if their coding sequence overlapped a cancer-specific mutation or not.

4.8.16 Peripheral expression of MHC peptide-coding sequences

To assess the peripheral expression of tumor-associated antigens' and aeTSA candidates' peptide-coding sequences, we used RNA-Seq data from 22 murine tissues, which had been sequenced by the ENCODE consortium[39,40] (**Supplementary Table 4.5**) or from 28 peripheral human tissues (~50 donors per tissue), which had been sequenced by the GTEx consortium and downloaded from the GTEx Portal on 04/16/2018 (phs000424.v7.p2, **Supplementary Table 4.16**). Briefly, RNA-Seq data from each tissue were transformed into 24-nucleotide-long k-mer databases with Jellyfish 2.2.3 (using the –C option) and used to query each peptide-coding sequence's 24-nucleotide-long k-mer set. For each RNA-Seq experiment, the number of reads fully overlapping a given MHC peptide-coding sequence ($r_{overlap}$) was estimated using the k-mer set's minimum occurrence (k_{min}). Indeed, we hypothesized that k_{min} ~ $r_{overlap}$ because, except for low complexity RNA-Seq reads that might generate the same k-mer multiple times, one k-mer always originate from a single RNA-Seq read. Thus, to compare the peptide-coding sequence expression across all tissues, we transformed this $r_{overlap}$ value into a number of reads detected per 10^8 reads sequenced ($rphm$) using the following formula: $rphm = \frac{(r_{overlap} \times 10^8)}{r_{tot}}$, with r_{tot} representing the total number of reads sequenced in a given RNA-Seq experiment. These values were then log-transformed ($\log_{10}(rphm + 1)$) and averaged across all RNA-Seq experiments of a given tissue. aeTSA candidates exhibiting a peripheral expression in 10 or less tissues (at rphm > 0) or in less than 5 tissues other than the liver (at rphm > 15) for murine and human candidates respectively, were considered as genuine aeTSAs. Features of those aeTSAs, as well as mTSAs are reported in **Supplementary Tables 4.6, 4.17** and **4.18**.

4.8.17 MS validation of TSA candidates

For CT26 TSA candidates and two EL4 TSA candidates (ATQQFQQL and SSPRGSSTL), we compared the previously acquired MS/MS spectra to the relevant ^{12}C-analog. For the other five EL4 TSA candidates tested in vivo (IILEFHSL,

TVPLNHNTL, VNYIHRNV, VNYLHRNV, VTPVYQHL), we eluted MHC peptides from six additional EL4 replicates (~450 to 1,400 x 10^6 cells per replicate), which were all processed as previously described (see **sections 4.8.12** and **4.8.13**). For absolute quantification, three of the six EL4 replicates were spiked with 500 fmol of each ^{13}C-labelled TSA. For sequence validation, MS/MS spectrum of ^{12}C TSA candidates were acquired prior to sample analysis by PRM-MS. Briefly, the PRM acquisition, which monitored five peptides as scheduled (each peptide is only monitored in a 10-minute window centered on its elution time), consisted of one MS1 scan followed by the targeted MS/MS scans in HCD mode. Automatic gain controls and injection times for the survey scan and the tandem mass spectra were 3e6 – 50 ms and 2e5 – 100 ms, respectively. In all cases, Skyline[68] was used to extract the endogenous MS/MS spectrum of each TSA candidate and compare it to the relevant ^{12}C MS/MS spectrum (sequence validation) or to extract the intensity of the endogenous and the relevant synthetic ^{13}C-labelled peptide (absolute quantification). Using the following formula, these intensities were further used to compute the number of TSA copy per cell for each replicate: $\left(n_{synthetic} \times I_{endogenous} \times N_A/I_{synthetic}\right) \times (1/N_{cells})$ with $n_{synthetic}$, initial number of moles spiked for the considered synthetic ^{13}C-labelled TSA; $I_{endogenous}$ and $I_{synthetic}$, intensity of the relevant endogenous and ^{13}C-labelled TSA, respectively; N_A, Avogadro's number; N_{cells}, initial number of cells used for mild acid elution.

4.8.18 Cumulative number of transcripts detected in human TEC and mTEC samples

Restricting our analysis to transcripts expressed at a tpm > 1 in at least one of our six samples (2 TECs and 6 mTECs), we first computed Spearman's rank correlation coefficient for each 1-to-1 TEC/mTEC comparison. Then, using those same sets of expressed transcripts, we computed the cumulative numbers of transcripts (cT) detected as each additional sample are analyzed. Because the order in which samples are introduced in the analysis can influence cT values, we averaged the cT values across all sample permutations and used those average data points to fit the following predictive curve (with the R's 'nls' function): $cT = \frac{a \times (nS-1)}{[b+(nS-1)]} + c$, with cT, the cumulative

numbers of transcripts and nS, the number of analyzed samples. This equation was then used to extrapolate the number of transcripts that would have been detected by studying up to 20 samples and which can be estimated by simply computing $\lim_{nS\to\infty}(cT)$.

4.8.19 Generation of bone marrow-derived dendritic cells (DCs), mouse immunization and EL4 cell injection

Bone marrow-derived DCs were generated as previously described[42]. For mouse immunization, DCs from male C57BL/6 mice were pulsed with 2 µM of the selected peptide for 3 hours, then washed. 8–12 week-old female C57BL/6 mice were injected i.v. with 10^6 DCs pulsed with one of the TSA or with irradiated EL4 cells (10,000 cGy) at day −14 and −7. As negative control, C57BL/6 female mice were immunized with unpulsed DCs. At day 0 and day 150, mice were injected i.v. with 5 x 10^5 EL4 cells and were monitored for weight loss, paralysis, or tumor outgrowth.

4.8.20 IFN-γ ELISpot and avidity assays

IFN-γ ELISpot and avidity assays were performed as previously described[42]. Briefly, Millipore MultiScreen PVDF plates were permeabilized with 35% ethanol, washed, and coated overnight using the Mouse IFN-γ ELISpot Ready-SET-Go! reagent set (eBioscience). At day 0 following mice immunization, splenocytes were harvested from immunized or naive mice. 30 x 10^6 splenocyte/mL were stained with FITC-conjugated anti-CD8a (BD Biosciences) for 30 minutes at 4°C, washed, and sorted using a FACS AriaIIu or a FACS AriaIIIu apparatus (BD Biosciences, **Supplementary Figure 4.1B**). Sorted $CD8^+$ T cells were plated and incubated at 37°C for 48 hours in the presence of irradiated splenocytes (4,000 cGy) from syngeneic mice pulsed with the relevant peptide (4 µM for the ELISpot assay and 10^{-4} to 10^{-14} M for the avidity assay). As a negative control, $CD8^+$ T cells from naive mice were incubated with peptide-pulsed splenocytes. Spots were revealed using the reagent set manufacturer protocol and were enumerated using an ImmunoSpot S5 UV Analyzer (Cellular Technology Ltd). IFN-γ production was expressed as the number of spot-forming cells per 10^6 $CD8^+$ T cells and the EC_{50} was calculated using a dose-response curve.

4.8.21 Cell isolation from lymphoid tissue and tetramer-based enrichment protocol

Spleen and lymph nodes (inguinal, axillary, brachial, cervical and mesenteric) were harvested from naive C57BL/6 mice or at day 0 for immunized mice, at signs of disease or day 21 for non-immunized EL4-injected mice and at signs of disease or day 210 for rechallenged mice. Single-cell suspensions were stained with Fc block and 10 nM of PE- or APC-labeled peptide/MHC tetramers (NIH Tetramer Core Facility) for 30 minutes at 4ºC. After washing with ice-cold sorting buffer (PBS with 2% FBS), cells were resuspended in 200 µL of sorting buffer and 50 µL of anti-PE and/or anti-APC antibody conjugated magnetic microbeads (Miltenyi Biotech), then incubated for 20 minutes at 4ºC. Cells were then washed and tetramer+ cells were magnetically enriched as previously described[47,48]. The resulting tetramer$^+$-enriched fractions were stained with APC Fire 750-conjugated anti-B220, F4/80, CD19, CD11b, CD11c (BioLegend), PerCP-conjugated anti-CD4 (BioLegend), BV421-conjugated anti-CD3 (BD Biosciences), BB515-conjugated anti-CD8$^+$ (BD Biosciences), BV510-conjugated anti-CD44 (BD Biosciences) antibodies and Zombie NIR Fixable Viability Kit (BioLegend). Anti-CD11b and CD11c were left out for the analysis of post-immunization repertoires because these markers may be expressed by some activated T cells. The entire stained sample was then analyzed on a FACS CantoII cytometer (BD Biosciences) and fluorescent counting beads (Thermo Fisher Scientific) were used to normalize results. As negative control, we enriched the antigen-specific CD8$^+$ T-cell repertoires targeting three viral epitopes: gp-33 from the lymphocytic choriomeningitis virus protein gp-33 (KAVYNFATC; H-2-D^b), M45 from the murine cytomegalovirus protein M45 (HGIRNASFI; H-2-D^b) and B8R from the vaccinia virus protein B8R (TSYKFESV; H-2-K^b).

4.9 References and Notes

4.9.1 References

1. Mlecnik, B., et al., *The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis.* Sci Transl Med, 2016. **8**(327): p. 327ra26.
2. Charoentong, P., et al., *Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade.* Cell Rep, 2017. **18**(1): p. 248-262.
3. Wei, S.C., et al., *Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade.* Cell, 2017. **170**(6): p. 1120-1133 e17.
4. Shao, W., et al., *The SysteMHC Atlas project.* Nucleic Acids Res, 2018. **46**(D1): p. D1237-D1247.
5. Martin, S.D., et al., *Targeting the undruggable: immunotherapy meets personalized oncology in the genomic era.* Ann Oncol, 2015. **26**(12): p. 2367-74.
6. Marty, R., et al., *MHC-I Genotype Restricts the Oncogenic Mutational Landscape.* Cell, 2017. **171**(6): p. 1272-1283 e15.
7. Yarchoan, M., et al., *Targeting neoantigens to augment antitumour immunity.* Nat Rev Cancer, 2017. **17**(4): p. 209-222.
8. Gee, M.H., et al., *Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes.* Cell, 2018. **172**(3): p. 549-563 e16.
9. Kassiotis, G. and J.P. Stoye, *Immune responses to endogenous retroelements: taking the bad with the good.* Nat Rev Immunol, 2016. **16**(4): p. 207-19.
10. Probst, P., et al., *Sarcoma Eradication by Doxorubicin and Targeted TNF Relies upon CD8+ T-cell Recognition of a Retroviral Antigen.* Cancer Res, 2017. **77**(13): p. 3644-3654.
11. Goel, S., et al., *CDK4/6 inhibition triggers anti-tumour immunity.* Nature, 2017. **548**(7668): p. 471-475.
12. Ott, P.A., et al., *An immunogenic personal neoantigen vaccine for patients with melanoma.* Nature, 2017. **547**(7662): p. 217-221.

13. Sahin, U., et al., *Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer.* Nature, 2017. **547**(7662): p. 222-226.
14. Capietto, A.H., S. Jhunjhunwala, and L. Delamarre, *Characterizing neoantigens for personalized cancer immunotherapy.* Curr Opin Immunol, 2017. **46**: p. 58-65.
15. *The problem with neoantigen prediction.* Nat Biotechnol, 2017. **35**(2): p. 97.
16. Pearson, H., et al., *MHC class I-associated peptides derive from selective regions of the human genome.* J Clin Invest, 2016. **126**(12): p. 4690-4701.
17. Di Marco, M., J.K. Peper, and H.G. Rammensee, *Identification of Immunogenic Epitopes by MS/MS.* Cancer J, 2017. **23**(2): p. 102-107.
18. Yadav, M., et al., *Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.* Nature, 2014. **515**(7528): p. 572-6.
19. Bassani-Sternberg, M. and G. Coukos, *Mass spectrometry-based antigen discovery for cancer immunotherapy.* Curr Opin Immunol, 2016. **41**: p. 9-17.
20. Bassani-Sternberg, M., et al., *Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.* Nat Commun, 2016. **7**: p. 13404.
21. Gubin, M.M., et al., *Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens.* Nature, 2014. **515**(7528): p. 577-81.
22. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nat Rev Genet, 2016. **17**(2): p. 93-108.
23. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
24. Laumont, C.M., et al., *Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames.* Nat Commun, 2016. **7**: p. 10238.
25. Laumont, C.M. and C. Perreault, *Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy.* Cell Mol Life Sci, 2018. **75**(4): p. 607-621.
26. Rosenberg, S.A., et al., *Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using*

lymphocytes from a patient with a durable complete regression following immunotherapy. J Immunol, 2002. **168**(5): p. 2402-7.

27. Kracht, M.J., et al., *Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes.* Nat Med, 2017. **23**(4): p. 501-507.

28. Simoni, Y., et al., *Bystander CD8(+) T cells are abundant and phenotypically distinct in human tumour infiltrates.* Nature, 2018. **557**(7706): p. 575-579.

29. Zhang, J., et al., *PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.* Mol Cell Proteomics, 2012. **11**(4): p. M111 010587.

30. Alfaro, J.A., et al., *Onco-proteogenomics: cancer proteomics joins forces with genomics.* Nat Methods, 2014. **11**(11): p. 1107-13.

31. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.

32. Noble, W.S., *Mass spectrometrists should search only for peptides they care about.* Nat Methods, 2015. **12**(7): p. 605-8.

33. Takahama, Y., et al., *Generation of diversity in thymic epithelial cells.* Nat Rev Immunol, 2017. **17**(5): p. 295-305.

34. Marcais, G. and C. Kingsford, *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.* Bioinformatics, 2011. **27**(6): p. 764-70.

35. Castle, J.C., et al., *Immunomic, genomic and transcriptomic characterization of CT26 colorectal carcinoma.* BMC Genomics, 2014. **15**: p. 190.

36. Fontaine, P., et al., *Adoptive transfer of minor histocompatibility antigen-specific T lymphocytes eradicates leukemia cells without causing graft-versus-host disease.* Nat Med, 2001. **7**(7): p. 789-94.

37. Vita, R., et al., *FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability.* Database (Oxford), 2018. **2018**.

38. Huang, A.Y., et al., *The immunodominant major histocompatibility complex class I-restricted antigen of a murine colon tumor derives from an endogenous retroviral gene product.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9730-5.

39. Yue, F., et al., *A comparative encyclopedia of DNA elements in the mouse genome.* Nature, 2014. **515**(7527): p. 355-64.
40. Sloan, C.A., et al., *ENCODE data at the ENCODE portal.* Nucleic Acids Res, 2016. **44**(D1): p. D726-32.
41. Fortier, M.H., et al., *The MHC class I peptide repertoire is molded by the transcriptome.* J Exp Med, 2008. **205**(3): p. 595-610.
42. Vincent, K., et al., *Rejection of leukemic cells requires antigen-specific T cells with high functional avidity.* Biol Blood Marrow Transplant, 2014. **20**(1): p. 37-45.
43. Bassani-Sternberg, M., et al., *Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.* Mol Cell Proteomics, 2015. **14**(3): p. 658-73.
44. Granados, D.P., et al., *The nature of self for T cells-a systems-level perspective.* Curr Opin Immunol, 2015. **34**: p. 1-8.
45. Chen, W., et al., *Dissecting the multifactorial causes of immunodominance in class I-restricted T cell responses to viruses.* Immunity, 2000. **12**(1): p. 83-93.
46. Jenkins, M.K. and J.J. Moon, *The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude.* J Immunol, 2012. **188**(9): p. 4135-40.
47. Moon, J.J., et al., *Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude.* Immunity, 2007. **27**(2): p. 203-13.
48. Legoux, F.P. and J.J. Moon, *Peptide:MHC tetramer-based enrichment of epitope-specific T cells.* J Vis Exp, 2012(68): p. 4420.
49. Vigneron, N., et al., *Database of T cell-defined human tumor antigens: the 2013 update.* Cancer Immun, 2013. **13**: p. 15.
50. Rooney, M.S., et al., *Molecular and genetic properties of tumors associated with local immune cytolytic activity.* Cell, 2015. **160**(1-2): p. 48-61.
51. Balachandran, V.P., et al., *Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer.* Nature, 2017. **551**(7681): p. 512-516.

52. Welters, M.J., et al., *Vaccination during myeloid cell depletion by cancer chemotherapy fosters robust T cell responses.* Sci Transl Med, 2016. **8**(334): p. 334ra52.

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

54. Gfeller, D. and M. Bassani-Sternberg, *Predicting Antigen Presentation-What Could We Learn From a Million Peptides?* Front Immunol, 2018. **9**: p. 1716.

55. Lu, Y.C., et al., *An Efficient Single-Cell RNA-Seq Approach to Identify Neoantigen-Specific T Cell Receptors.* Mol Ther, 2018. **26**(2): p. 379-389.

56. Hu, Z., et al., *A cloning and expression system to probe T cell receptor specificity and assess functional avidity to neoantigens.* Blood, 2018.

57. Kranz, L.M., et al., *Systemic RNA delivery to dendritic cells exploits antiviral defence for cancer immunotherapy.* Nature, 2016.

58. Vizcaino, J.A., et al., *2016 update of the PRIDE database and its related tools.* Nucleic Acids Res, 2016. **44**(22): p. 11033.

59. Lanoix, J., et al., *Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods.* Proteomics, 2018: p. e1700251.

60. Kim, M.J., et al., *Young, proliferative thymic epithelial cells engraft and function in aging thymuses.* J Immunol, 2015. **194**(10): p. 4784-95.

61. Stoeckle, C., et al., *Isolation of myeloid dendritic cells and epithelial cells from human thymus.* J Vis Exp, 2013(79): p. e50951.

62. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

63. Daouda, T., C. Perreault, and S. Lemieux, *pyGeno: A Python package for precision medicine and proteogenomics.* F1000Res, 2016. **5**: p. 381.

64. Zerbino, D.R., et al., *Ensembl 2018.* Nucleic Acids Res, 2018. **46**(D1): p. D754-D761.

65. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.

66. Andreatta, M. and M. Nielsen, *Gapped sequence alignment using artificial neural networks: application to the MHC class I system.* Bioinformatics, 2016. **32**(4): p. 511-7.
67. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
68. Bereman, M.S., et al., *An Automated Pipeline to Monitor System Performance in Liquid Chromatography-Tandem Mass Spectrometry Proteomic Experiments.* J Proteome Res, 2016. **15**(12): p. 4763-4769.

4.9.2 Notes

4.9.2.1 Acknowledgements

We thank the following members of IRIC core facilities for their sound advice and technical assistance: Jennifer Huber and Florent Guilloteau from the genomic platform, Simon Comtois-Marotte and Emilie Cossette from the proteomic platform, Gaël Dulude, Danièle Gagné and Annie Gosselin from the flow cytometry platform as well as Isabelle Caron from the animal care facility. We acknowledge the dedicated work of Claude Rondeau from the Banque de Cellules Leucémiques du Québec. We also thank the NIH Tetramer Core Facility for providing all the tetramers used in this study. Furthermore, we thank the ENCODE consortium, especially the laboratories of Dr. Thomas Gingeras (Cold Spring Harbor Laboratory) and Dr. Michael Snyder (Stanford University) for generating the murine tissue datasets used in this study. Finally, we thank the Genotype-Tissue Expression (GTEx) Project for providing RNA-Seq data from human tissues used in this study.

4.9.2.2 Funding

This work was supported by grants from the Canadian Cancer Society (Grant 701564 to C.P. and P.T.), the Terry Fox Research Institute (Grant TRP 1060/32-iTNT to C.P.), and the Quebec Breast Cancer Foundation (Grant 19579 to C.P.). C.M.L. is supported by a Cole Foundation fellowship. IRIC receives infrastructure support from Genome Canada, the Canadian Center of Excellence in Commercialization and

Research, the Canadian Foundation for Innovation, and the Fonds de Recherche du Québec-Santé (FRQS).

4.9.2.3 Competing interests

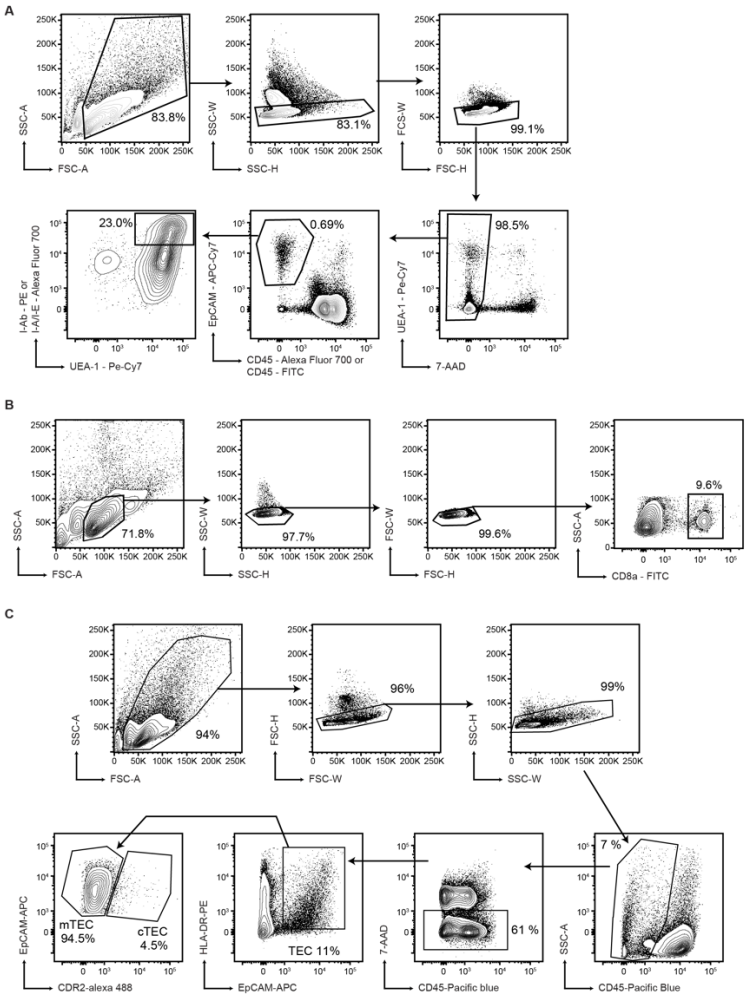
C.M.L., S.L., P.T. and C.P. are named inventors in the patent application 782-15691.134-US PROV APPLICATION (pending) filed by Université de Montréal on December 22, 2017. This patent application covers the method used for TSA discovery described in **Figure 4.1** and the TSAs listed in **Supplementary Tables 4.6, 4.17** and **4.18**. The remaining authors declare no competing interests.

4.9.2.4 Data and materials availability

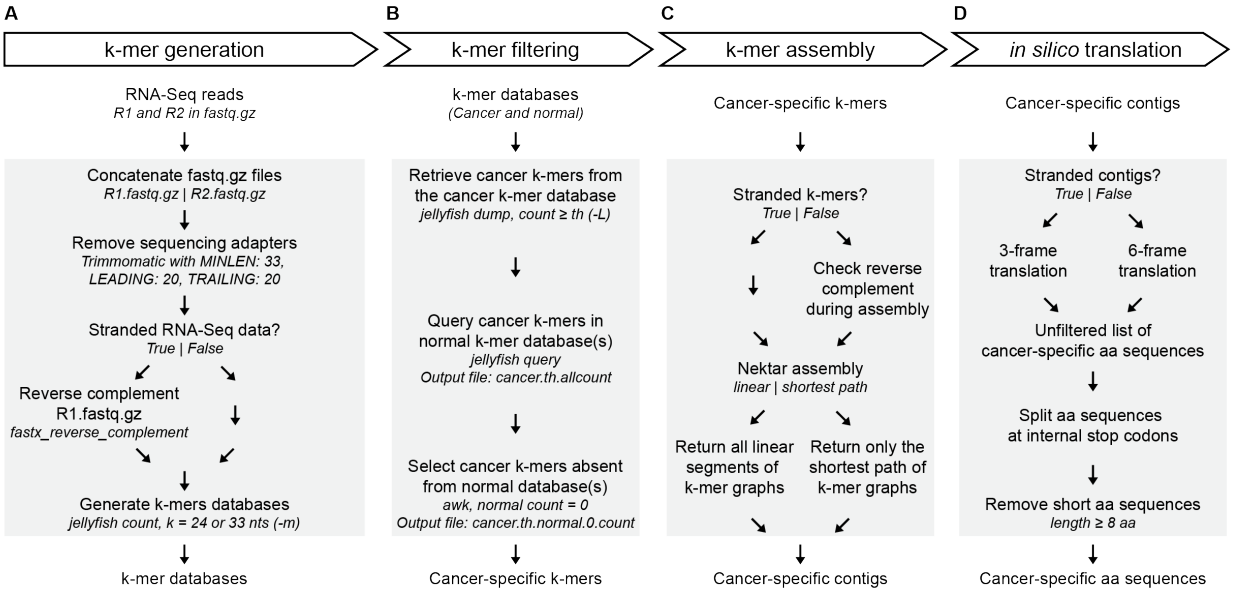
In-house scripts and the NEKTAR software used in this study are available on Zedono, using the following DOI: 10.5281/zenodo.1484486. pyGeno is available on GitHub (https://github.com/tariqdaouda/pyGeno). Information regarding all samples used in this study is listed in **Supplementary Table 4.2**. k-mer databases for human TECs and mTECs are available on Zenodo using the following DOIs: 10.5281/zenodo.1484261 (k=24 nts) and 10.5281/zenodo.1484490 (k=33 nts). For all other samples, RNA-seq and expression data have been deposited to the NCBI's Sequence Read Archive and GEO under accession code GSE113992, containing the GSE111092 and the GSE113972 sets of murine and human sequencing and expression data, respectively. MS raw data and associated databases have been deposited to the ProteomeXchange Consortium via the PRIDE[58] partner repository with the following dataset identifiers: PXD009065 and 10.6019/PXD009065 (CT26 cell line), PXD009064 and 10.6019/PXD009064 (EL4 cell line), PXD009749 and 10.6019/PXD009749 (07H103), PXD009753 and 10.6019/PXD009753 (10H080, mild acid elution), PXD007935 – assay # 81756 and 10.6019/PXD007935 (10H080, immunoprecipitation)[59], PXD009750 and 10.6019/PXD009750 (10H118), PXD009751 and 10.6019/PXD009751 (12H018), PXD009752 and 10.6019/PXD009752 (lc2), PXD009754 and 10.6019/PXD009754 (lc4) and PXD009755 and 10.6019/PXD009755 (lc6).

4.10 Supplementary Materials

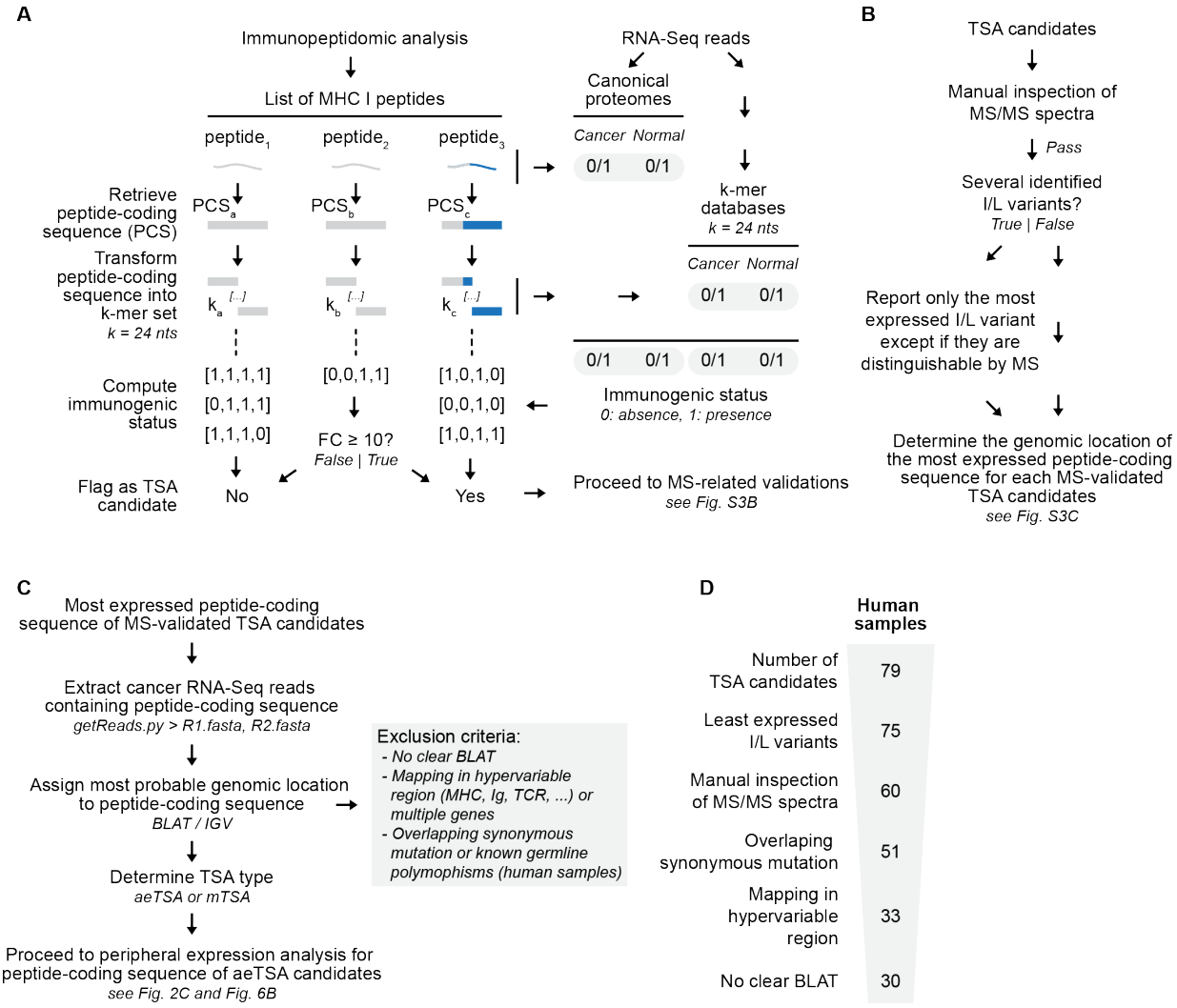
4.10.1 Supplementary Figures



Supplementary Figure 4.1 | Gating strategies for cells isolated by FACS. (**A**) Gating strategy for the isolation of murine mTEC^{hi}. mTEC^{hi} isolation was performed on single-cell suspensions isolated from thymi of C57BL/6 or Balb/c mice. After doublets exclusion, mTEC^{hi} cells were defined as 7-AAD^{-}, EpCAM^{+}, CD45^{-} (Alexa Fluor 700 for C57BL/6 or FITC for Balb/c mice), UEA-1^{+} and I-Ab^{+} (C57BL/6 mice) or I-A/I-E^{+} (Balb/c mice). (**B**) Gating strategy for the isolation of CD8^{+} T cells for IFN-γ ELISpot assays. CD8^{+} T cell isolation was performed on single-cell suspensions isolated from the spleen of naive or immunized C57BL/6 mice. After doublets exclusion, the CD8a marker was used to enrich for CD8^{+} T cells. (**C**) Gating strategy for the isolation of human TECs and mTECs. Cell sorting was performed on single-cell suspensions isolated from thymi that were obtained from 3-month-old to 7-year-old individuals undergoing corrective cardiovascular surgery. After doublets exclusion, TECs were defined as CD45^{-}, 7-AAD^{-}, EpCAM^{+} and HLA-DR^{+}. mTECs were further defined as CDR2^{-}.



Supplementary Figure 4.2 | Architecture of the codes used for our k-mer profiling workflow. (**A-D**) Details pertaining to the codes used to generate k-mers from RNA-Seq reads (**A**), filter k-mers (**B**), assemble k-mers into contigs (**C**) and translate contigs (**D**).

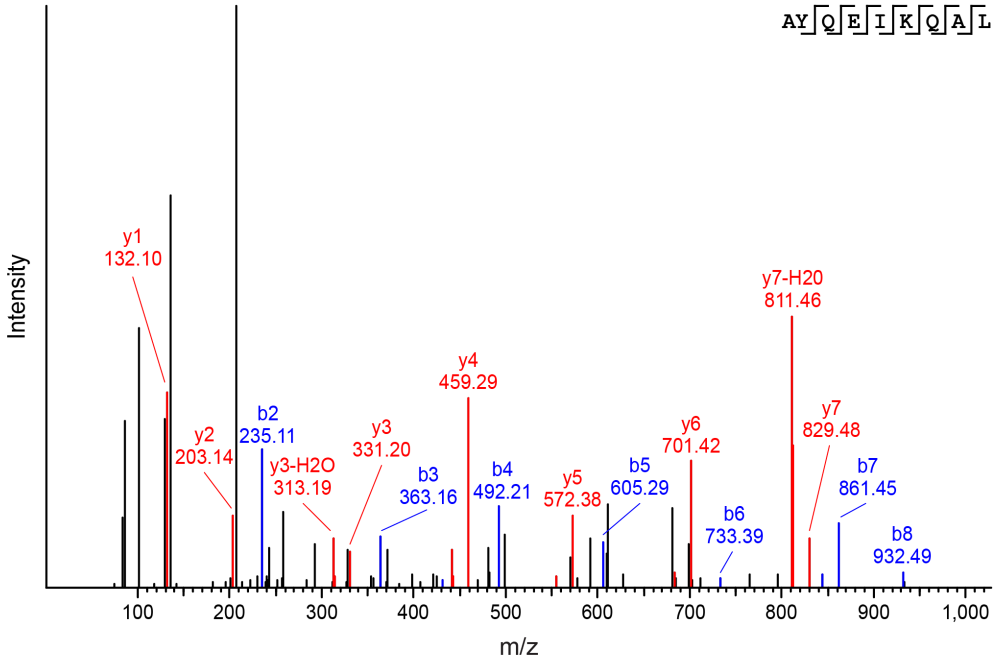


Supplementary Figure 4.3 | TSA validation process. (**A**) Schematic detailing the computational strategy used for identification of TSA candidates. FC: Tumor / syngeneic mTEC^{hi} (murine samples) or TEC/mTEC (human samples). (**B**) Strategy used to perform the MS-related validations of MHC peptides flagged as TSA candidates. (**C**) Schematic summarizing the strategy used to assign a genomic location to MS-validated murine TSA candidates (CT26 and EL4) as well as MS-validated human TSA candidates for B-ALL specimens and lung cancers. (**D**) Flowchart indicating key steps involved in human TSA validation, where the number of remaining peptides after each validation step is indicated (see **Supplementary Figure 4.3, A to C** for details).

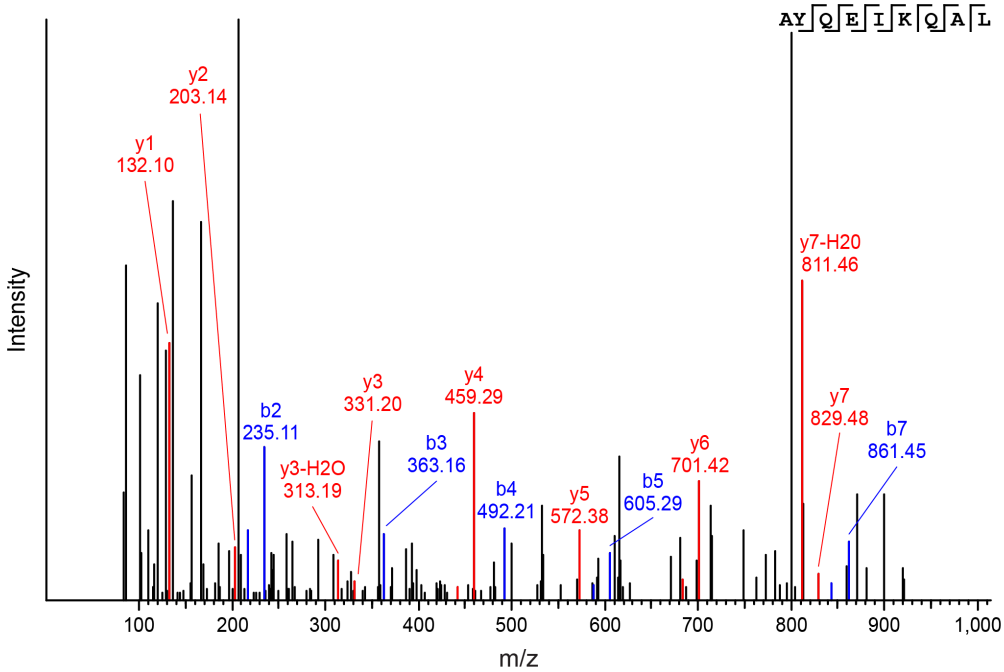
A

AYQEIKQAL - ERE aeTSA

Synthetic peptide



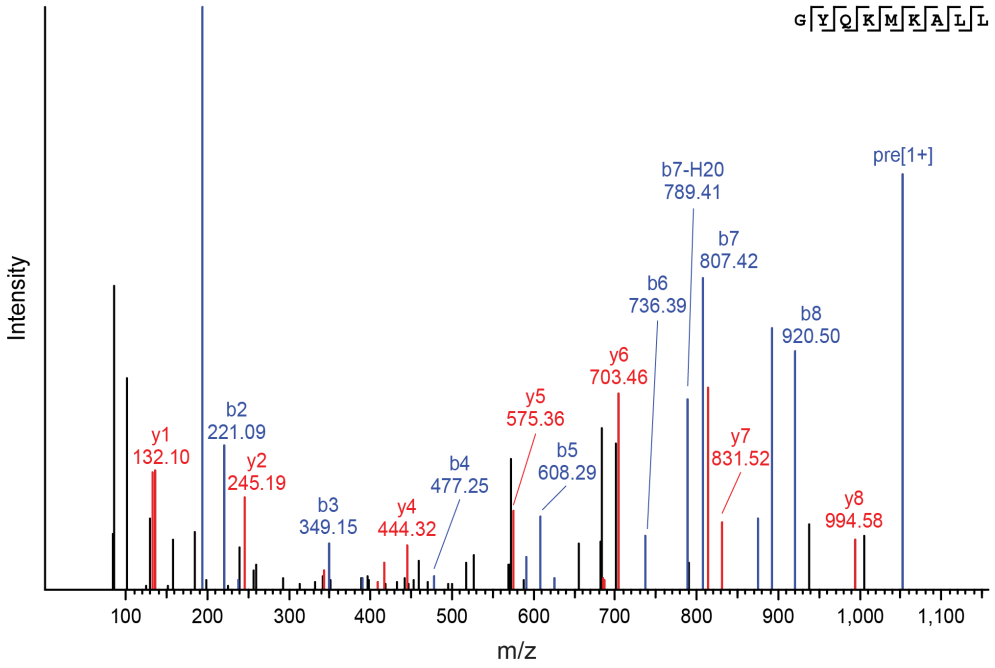
Endogenous peptide



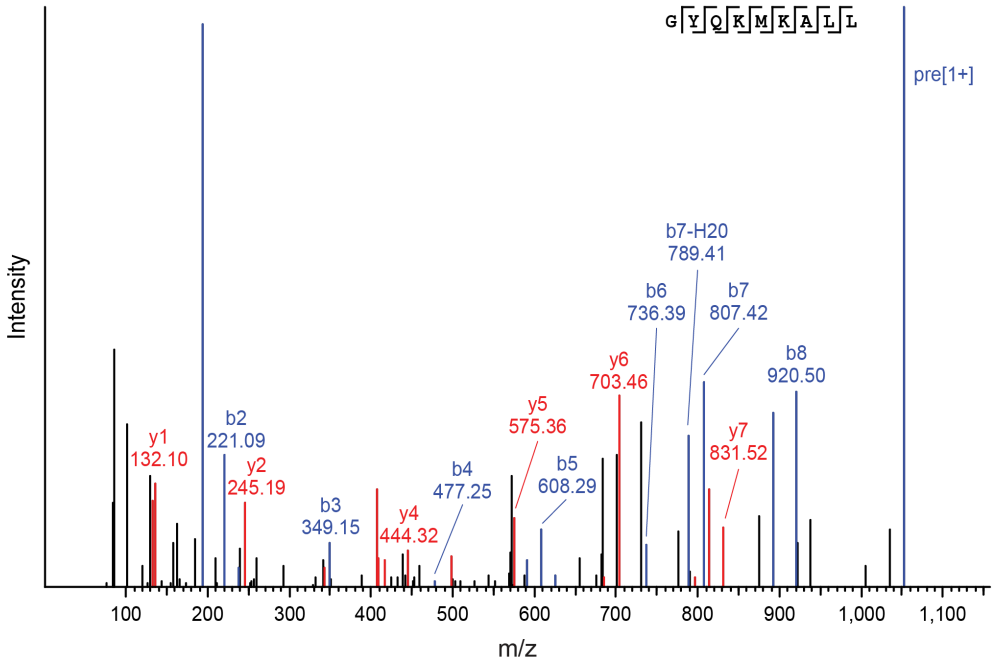
B

GYQKMKALL - ERE aeTSA

Synthetic peptide



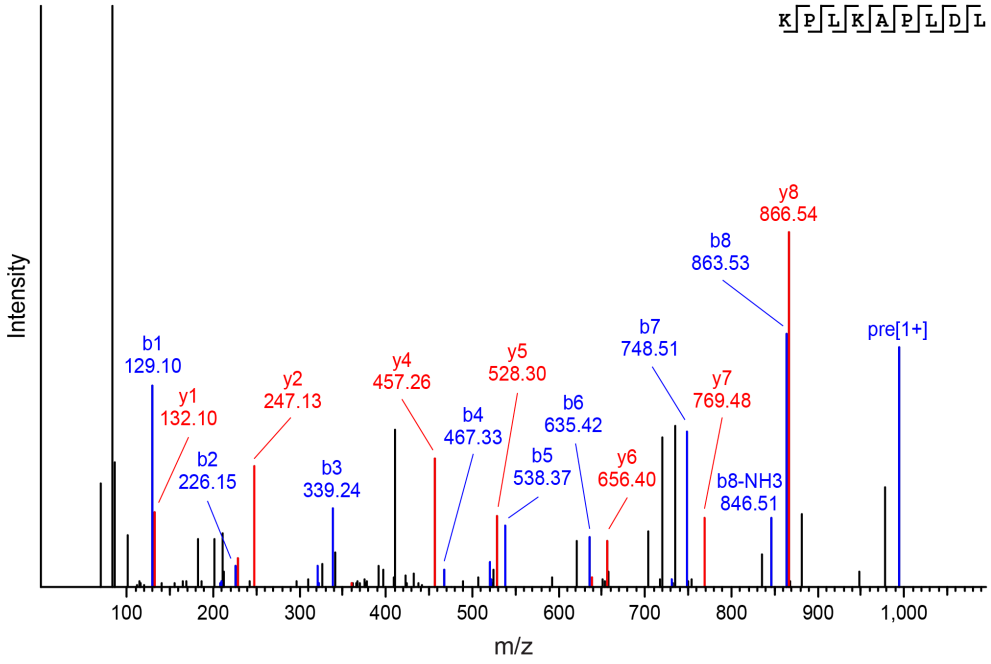
Endogenous peptide



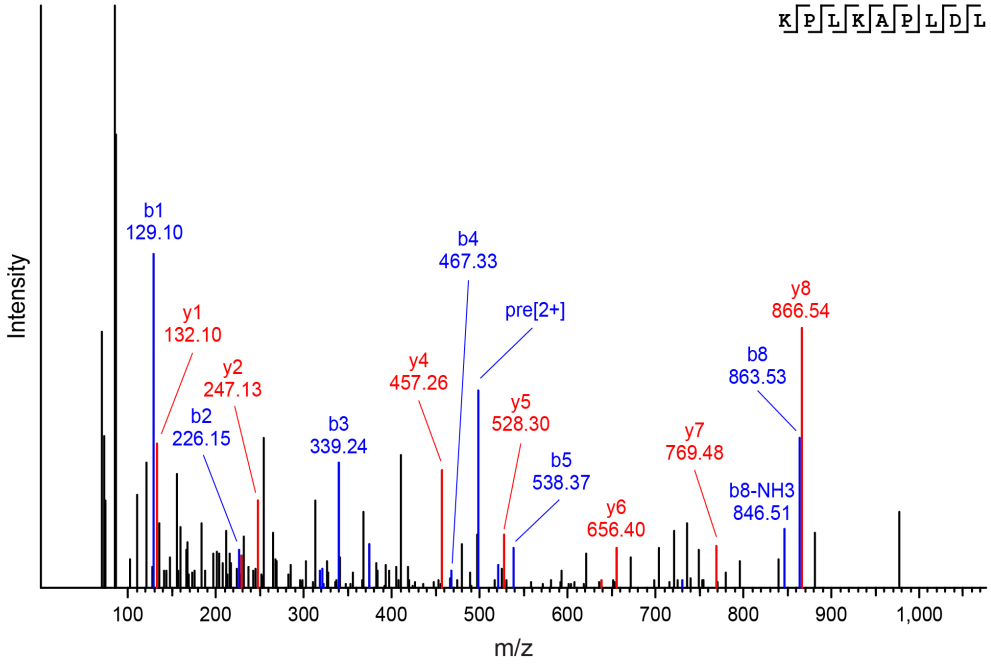
C

KPLKAPLDL - mTSA

Synthetic peptide



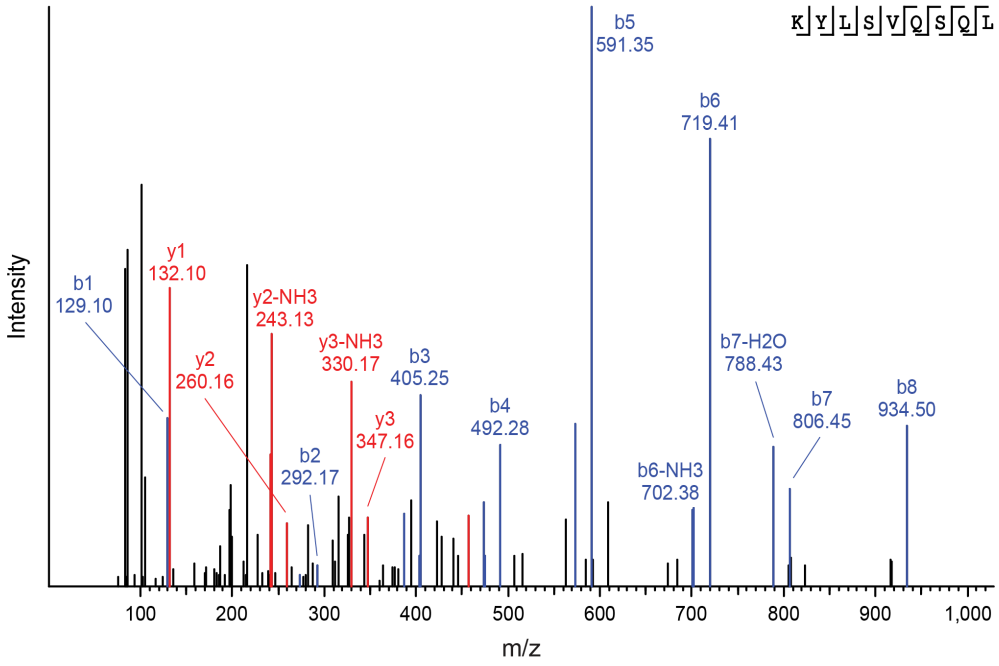
Endogenous peptide



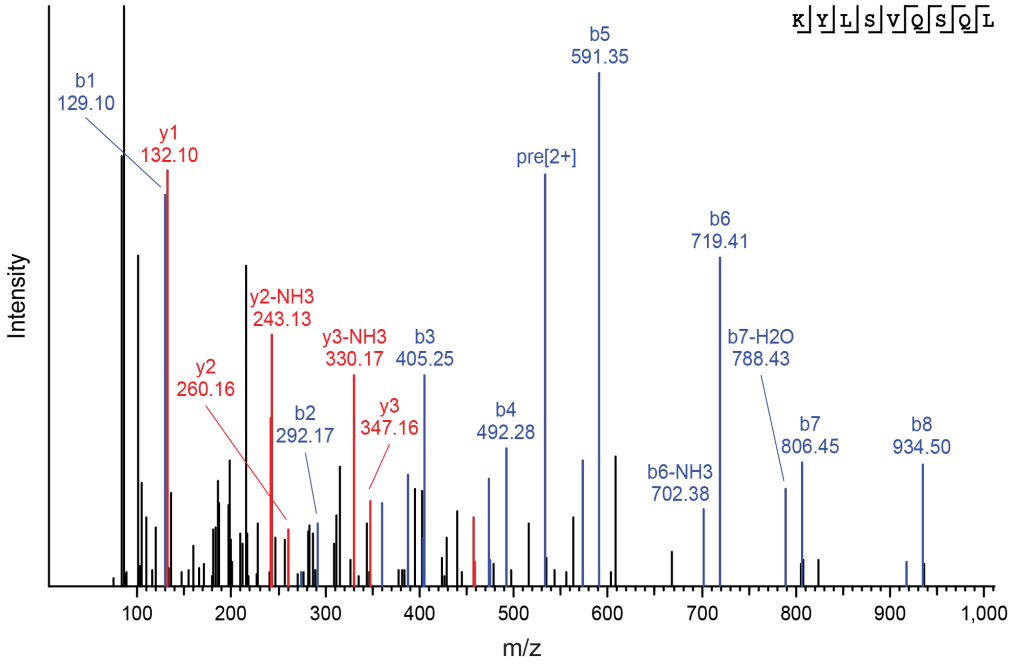
D

KYLSVQSQL - mTSA

Synthetic peptide



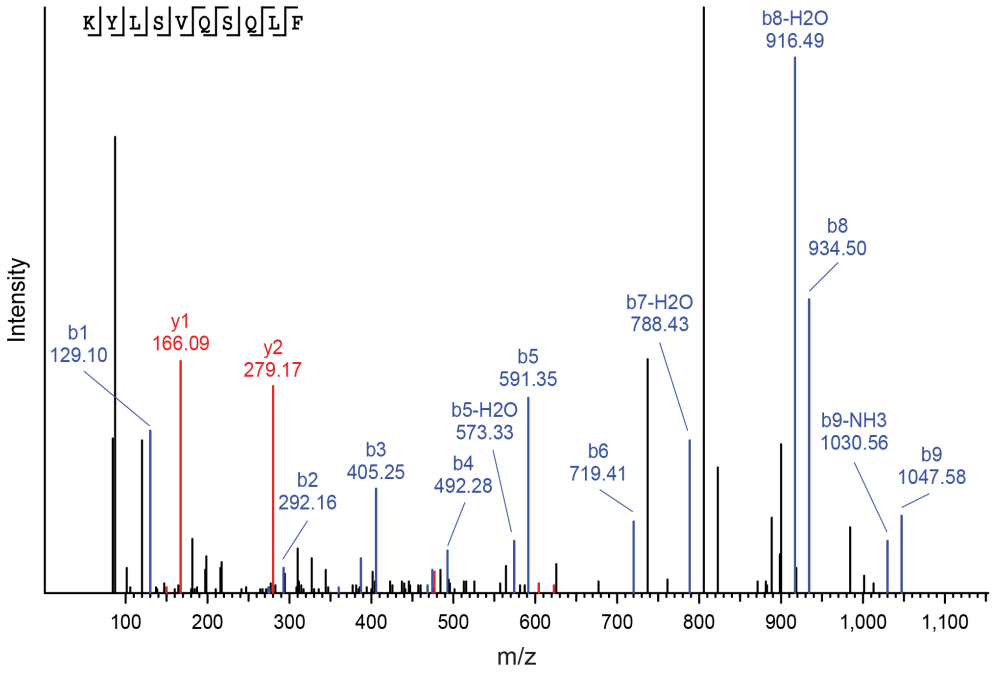
Endogenous peptide



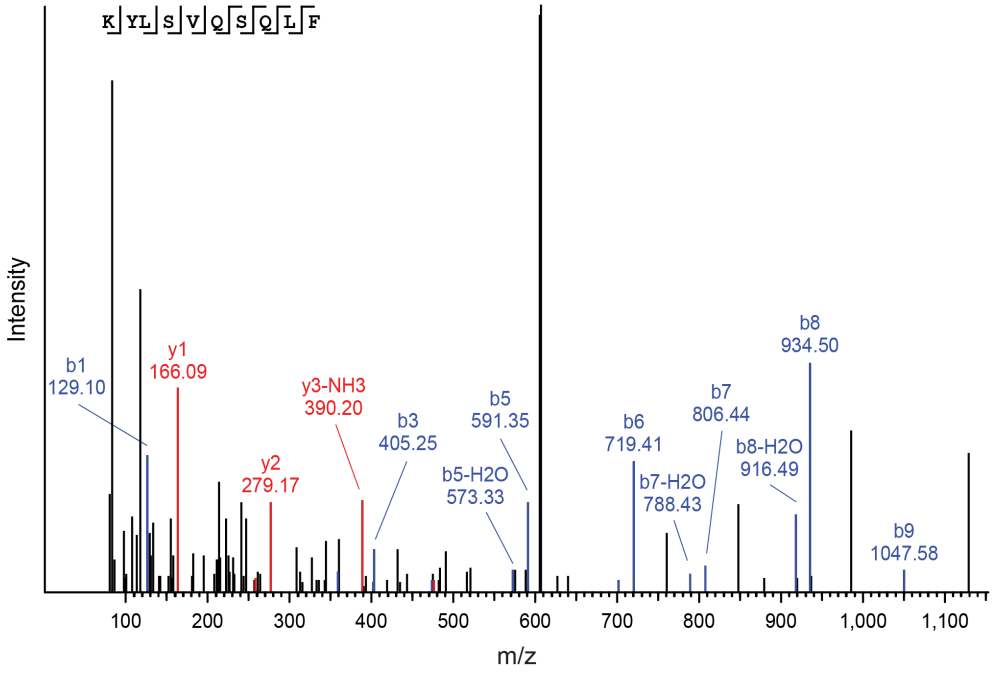
E

KYLSVQSQLF - mTSA

Synthetic peptide



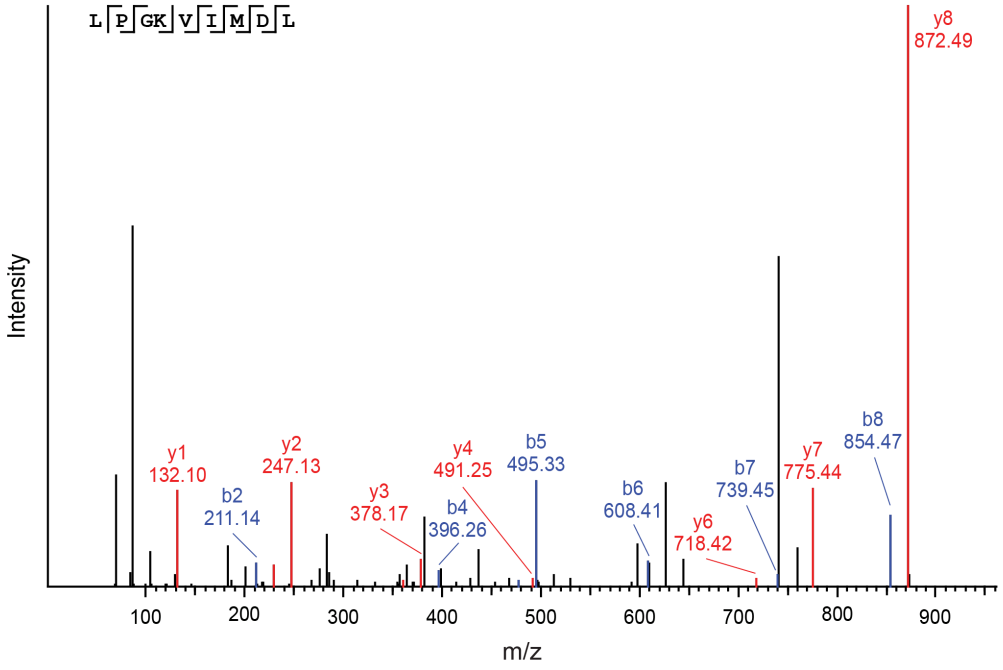
Endogenous peptide



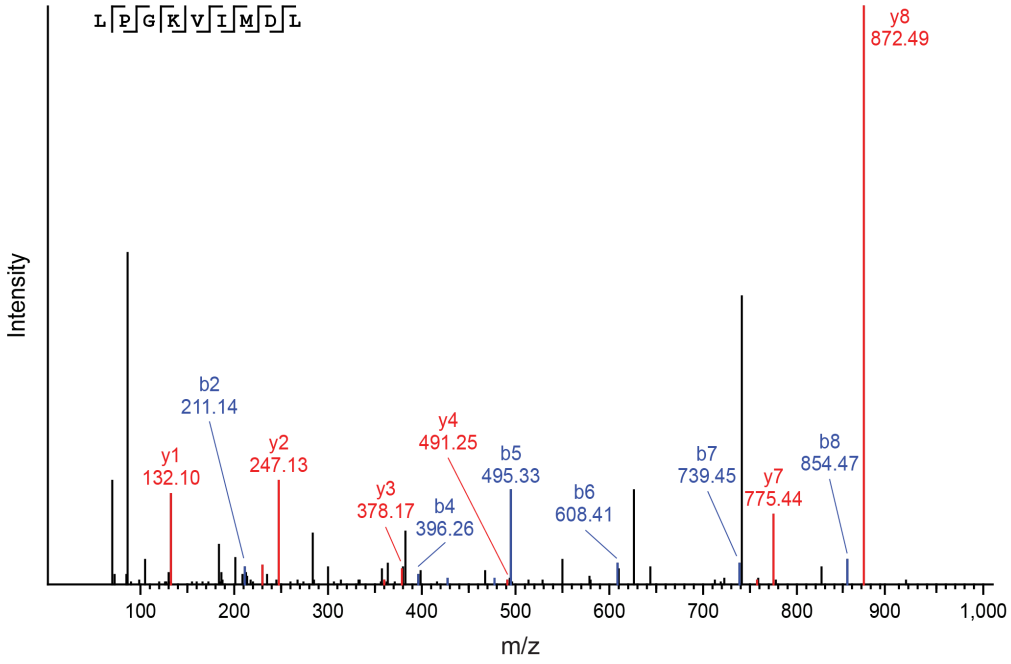
F

LPGKVIMDL - aeTSA

Synthetic peptide



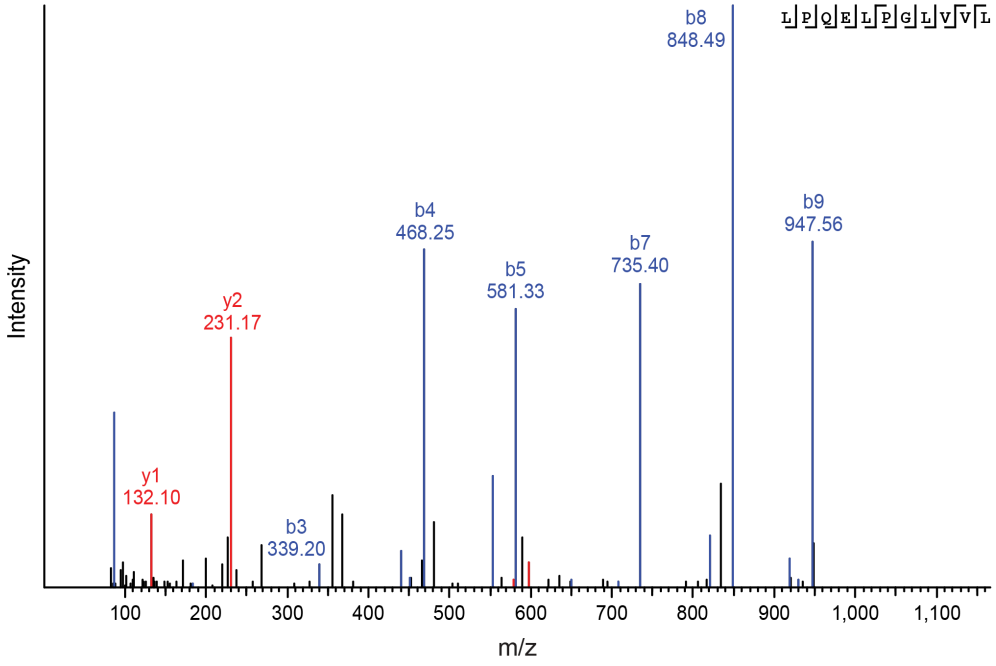
Endogenous peptide



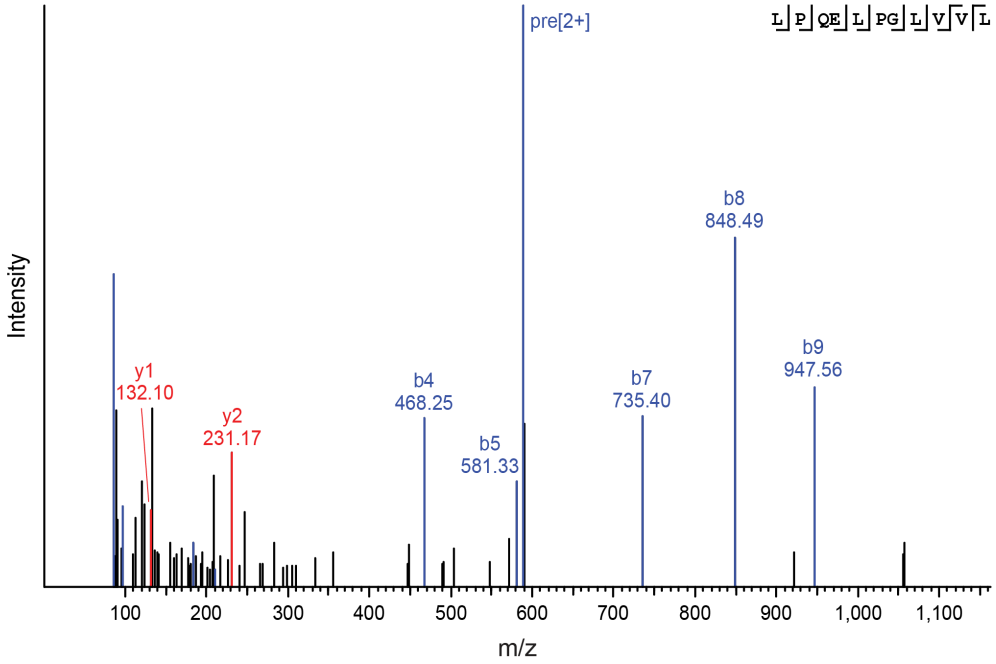
G

LPQELPGLVVL - ERE aeTSA

Synthetic peptide



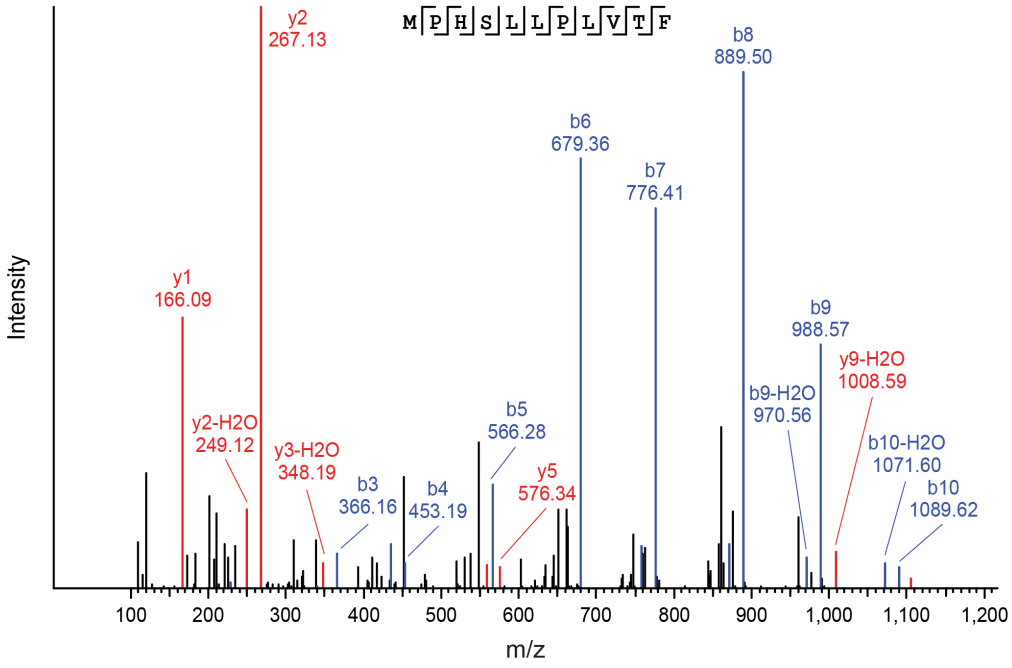
Endogenous peptide



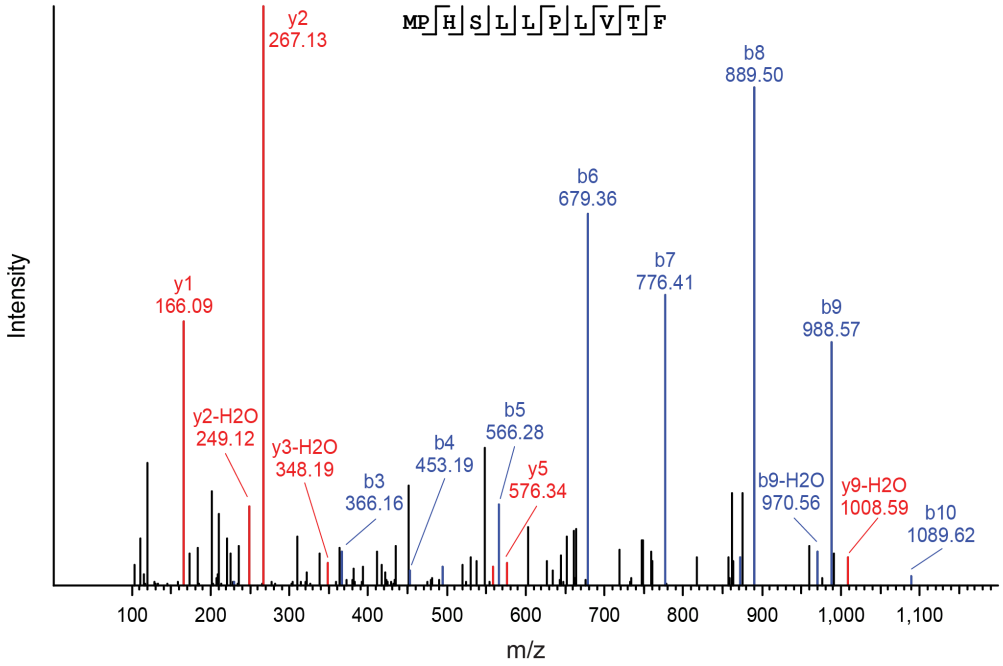
H

MPHSLLPLVTF - aeTSA

Synthetic peptide



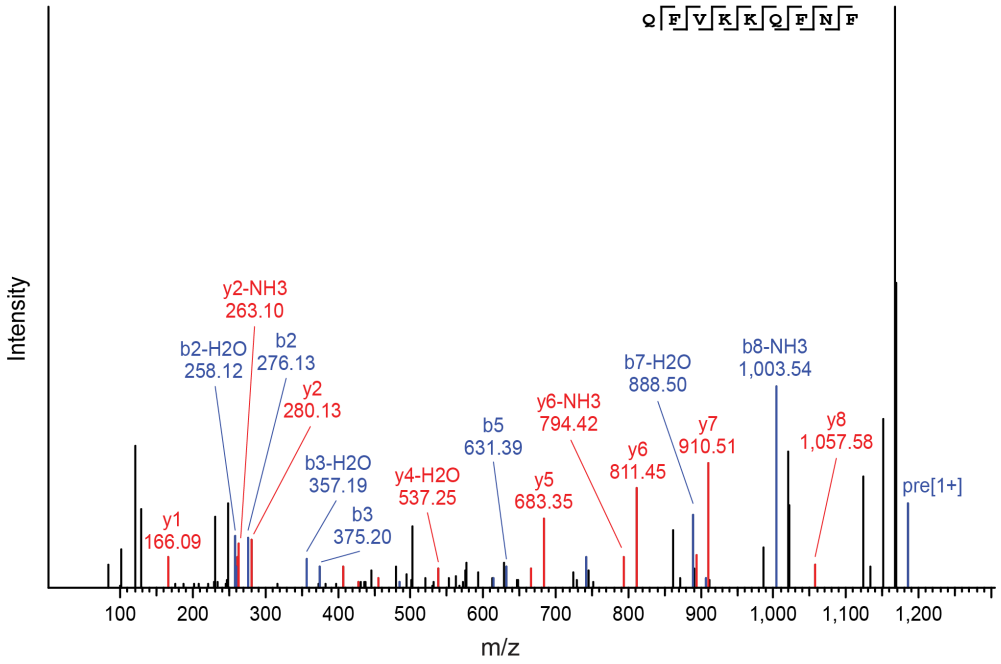
Endogenous peptide



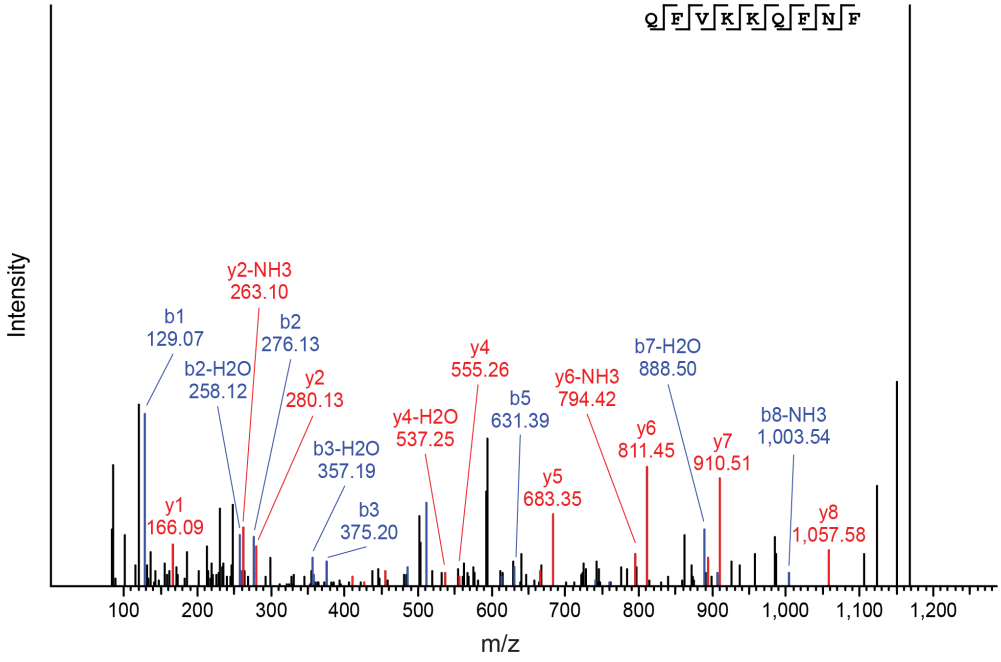
I

QFVKKQFNF - aeTSA

Synthetic peptide



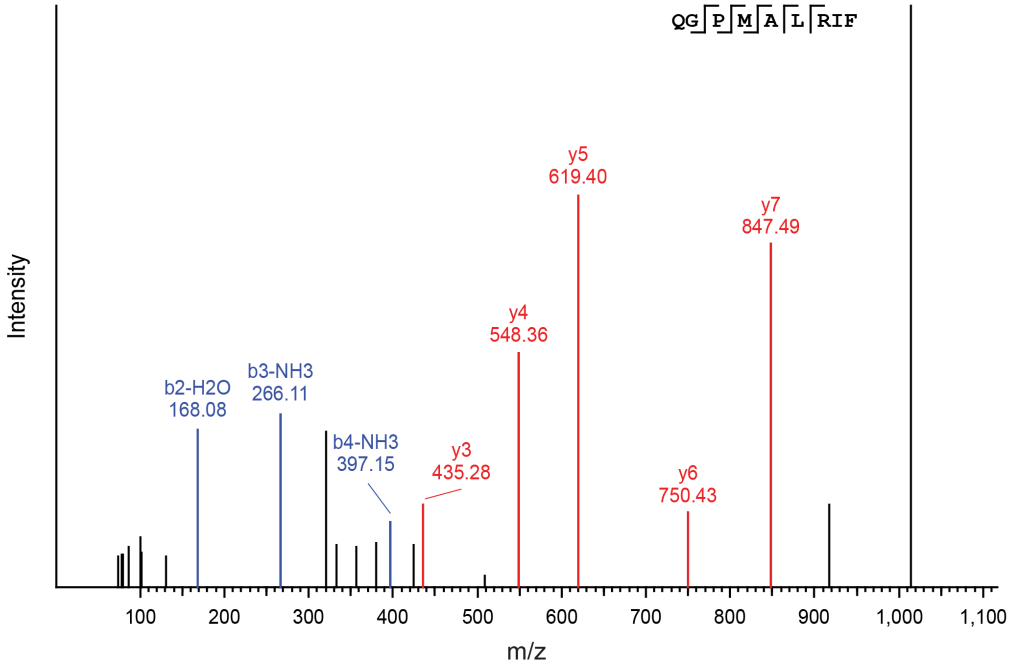
Endogenous peptide



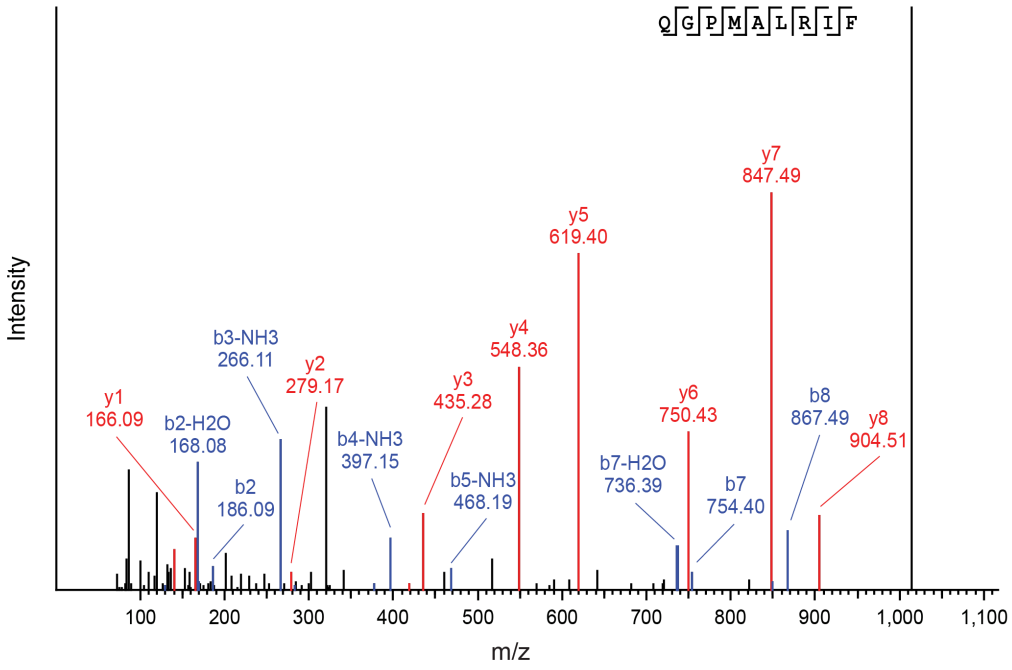
J

QGPMALRIF - mTSA

Synthetic peptide



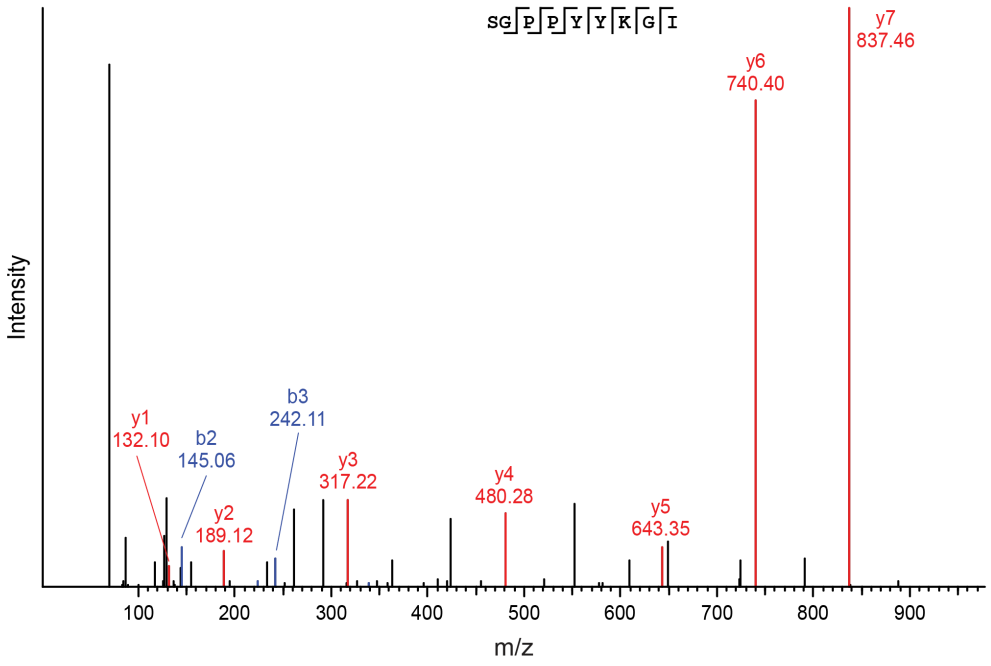
Endogenous peptide



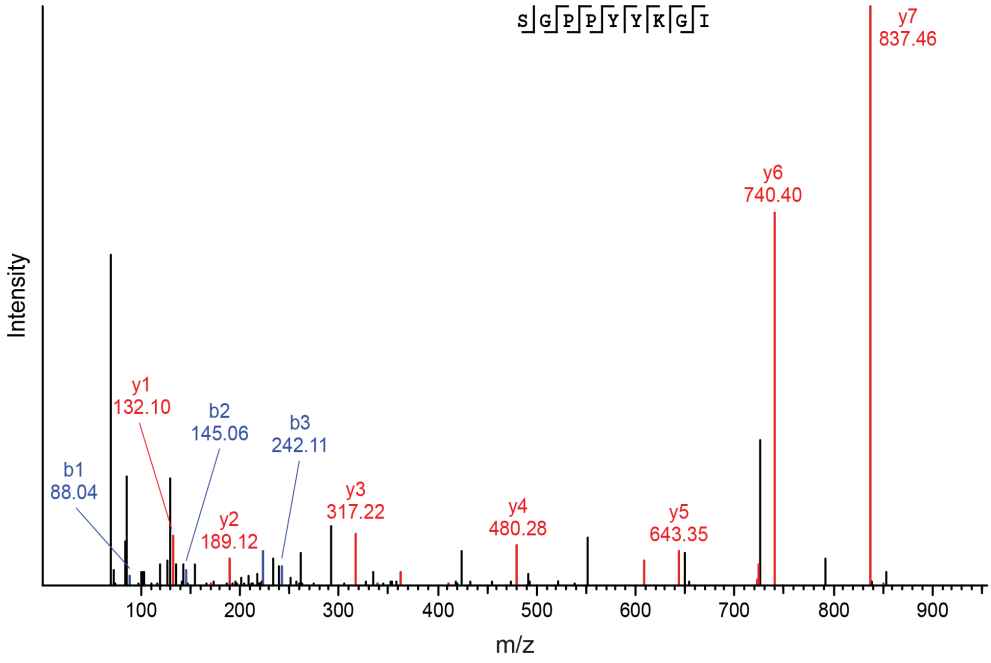
K

SGPPYYKGI - ERE aeTSA

Synthetic peptide



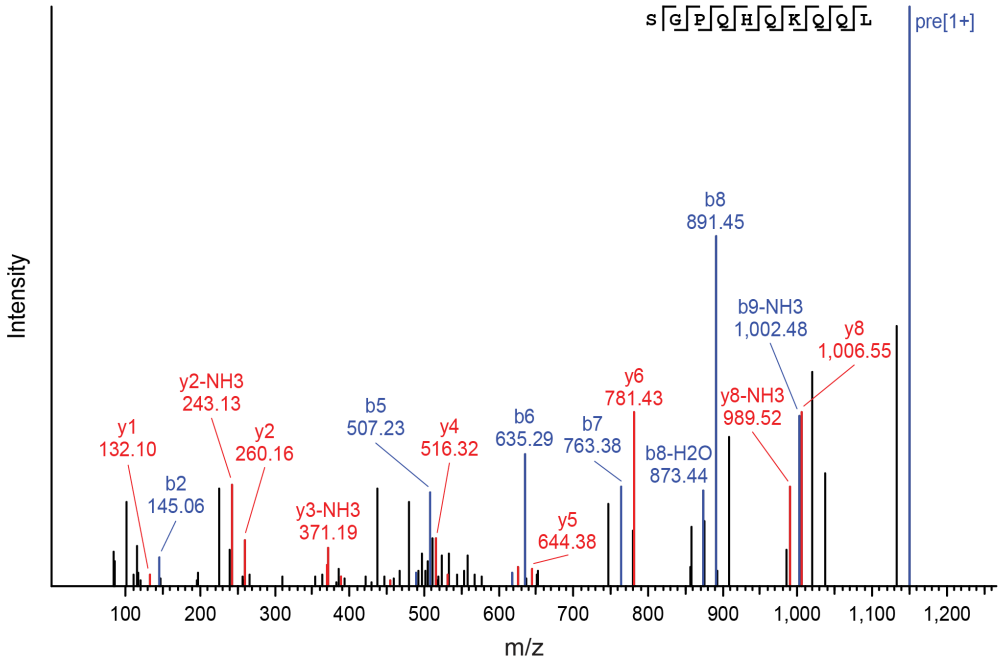
Endogenous peptide



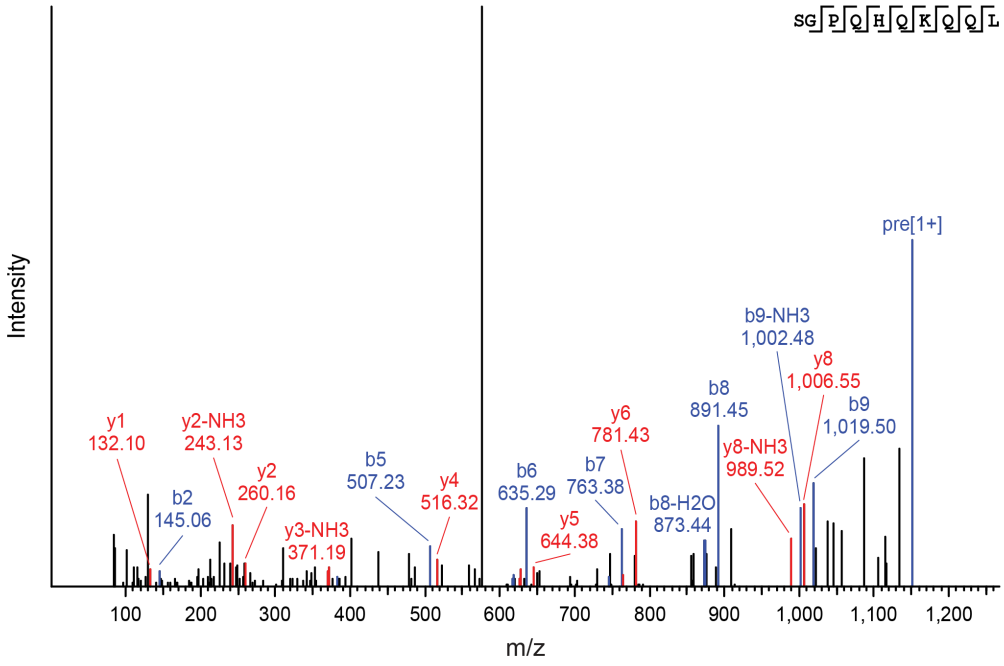
L

SGPQHQKQQL - aeTSA

Synthetic peptide



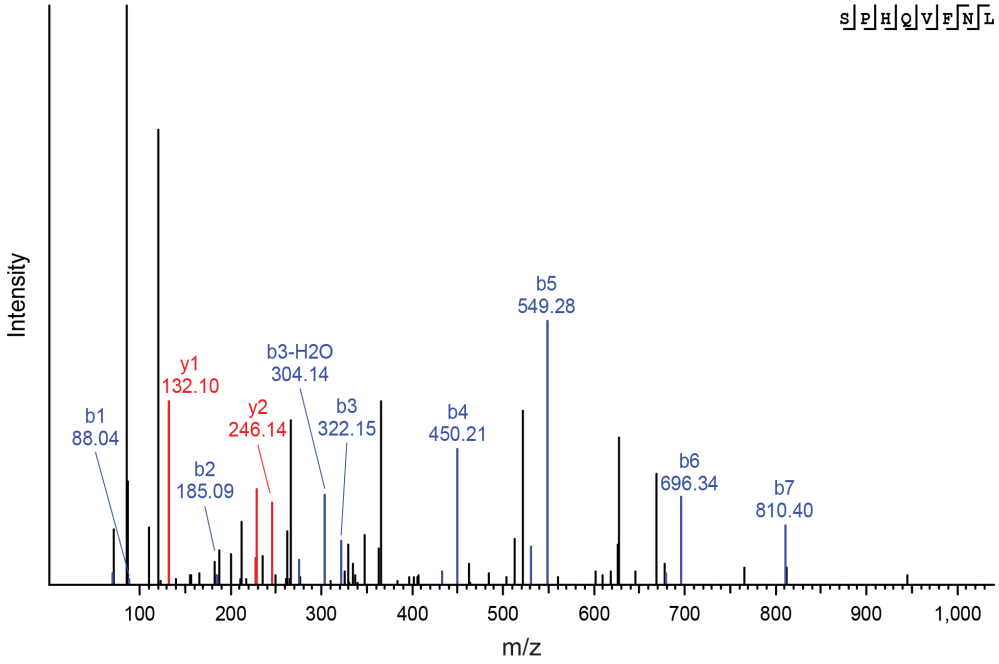
Endogenous peptide



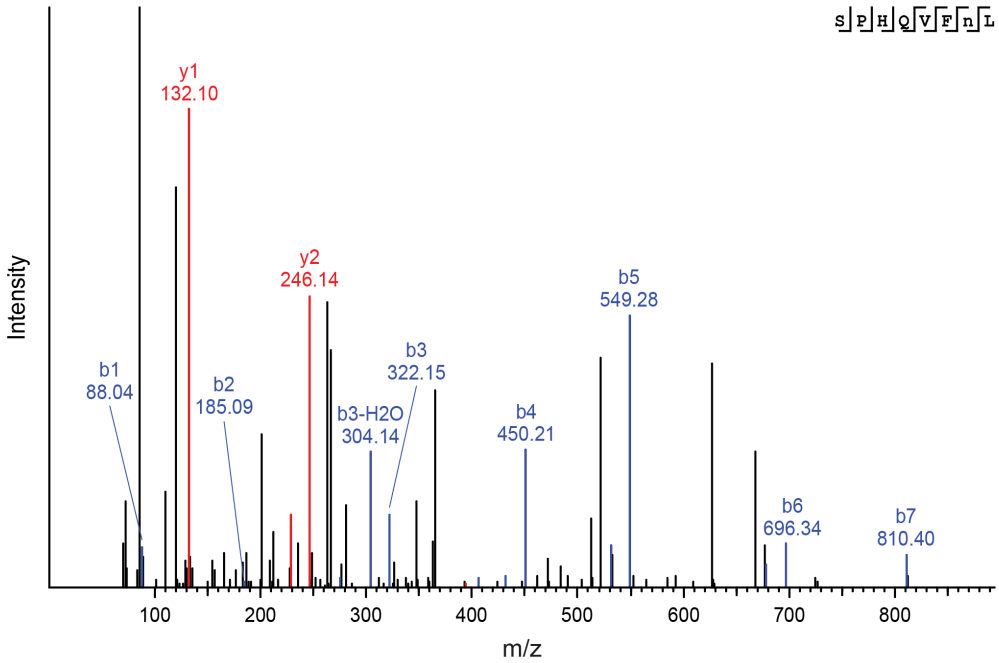
M

SPHQVFNL - ERE aeTSA

Synthetic peptide



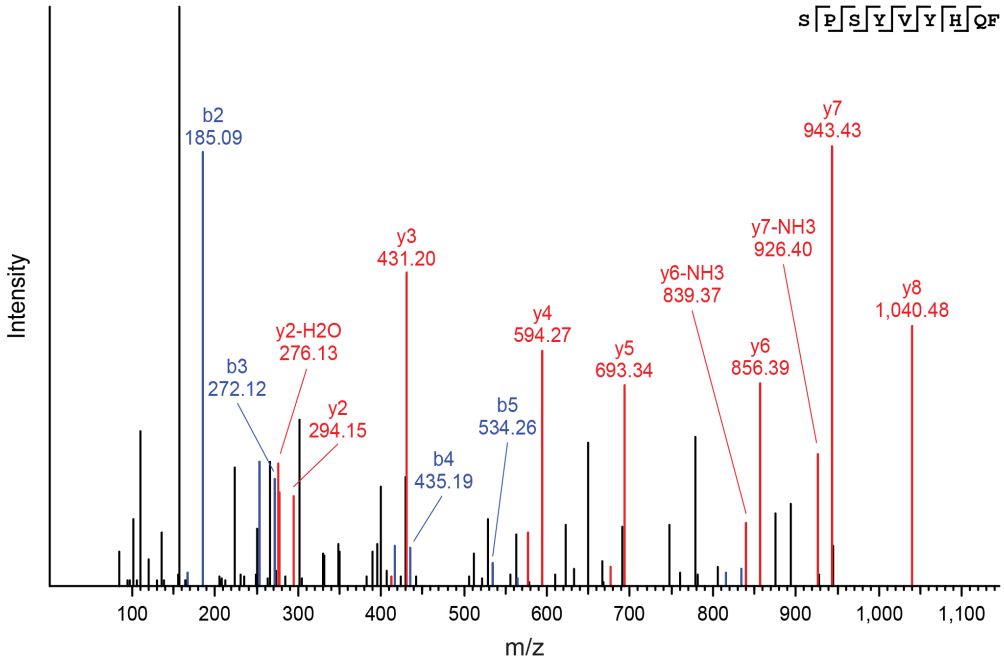
Endogenous peptide



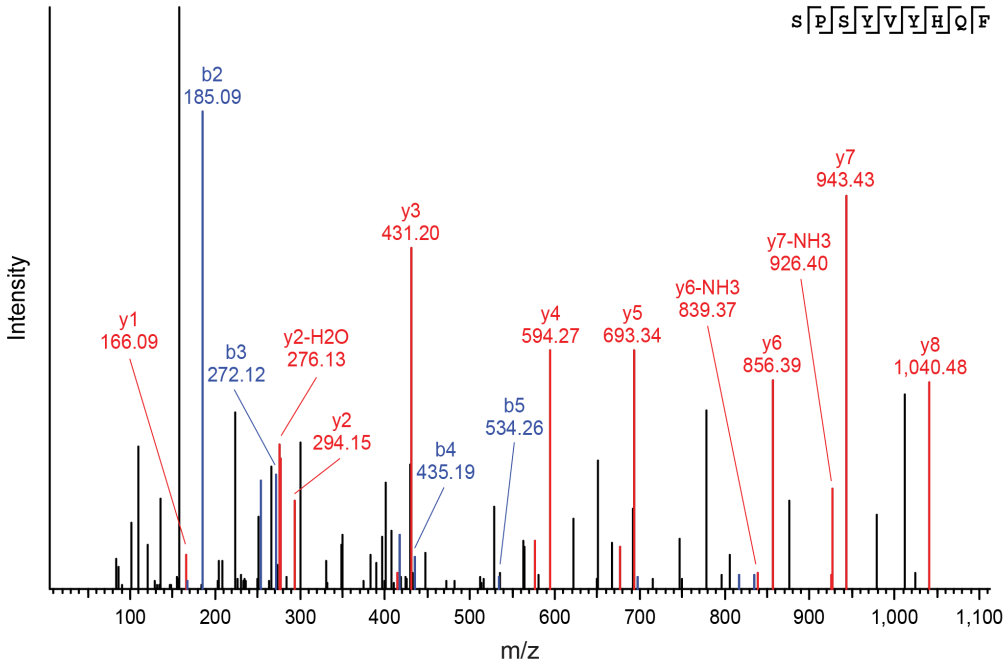
N

SPSYVYHQF - ERE aeTSA

Synthetic peptide



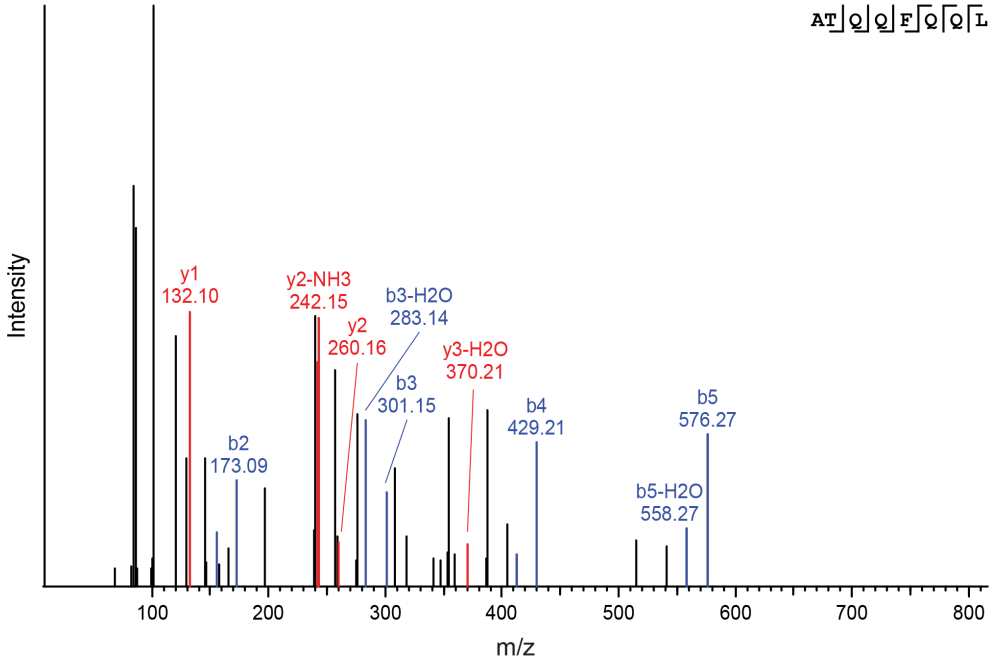
Endogenous peptide



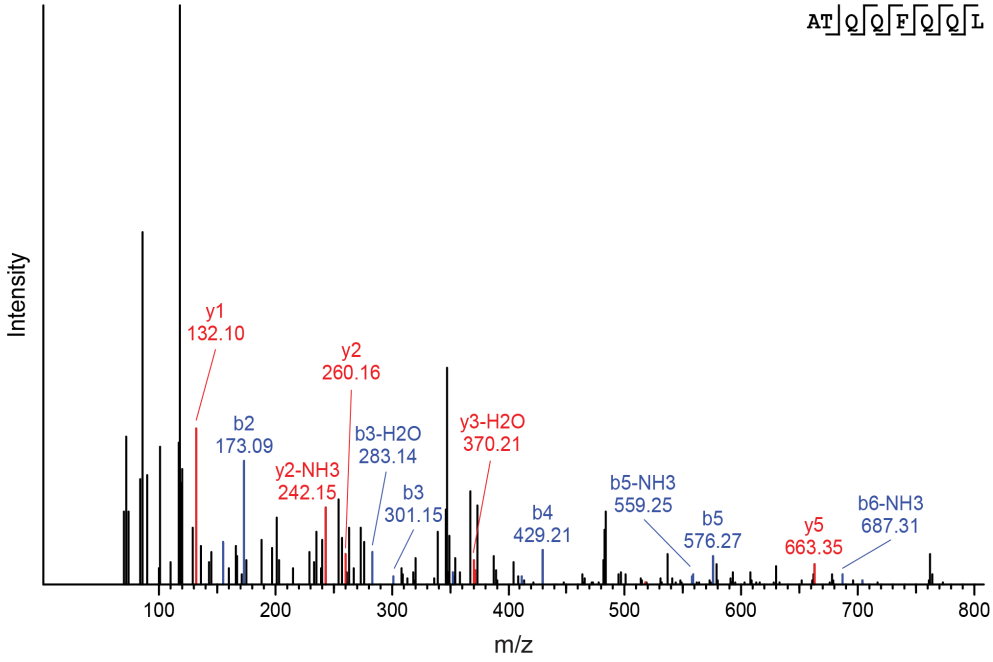
O

ATQQFQQL - ERE aeTSA

Synthetic peptide



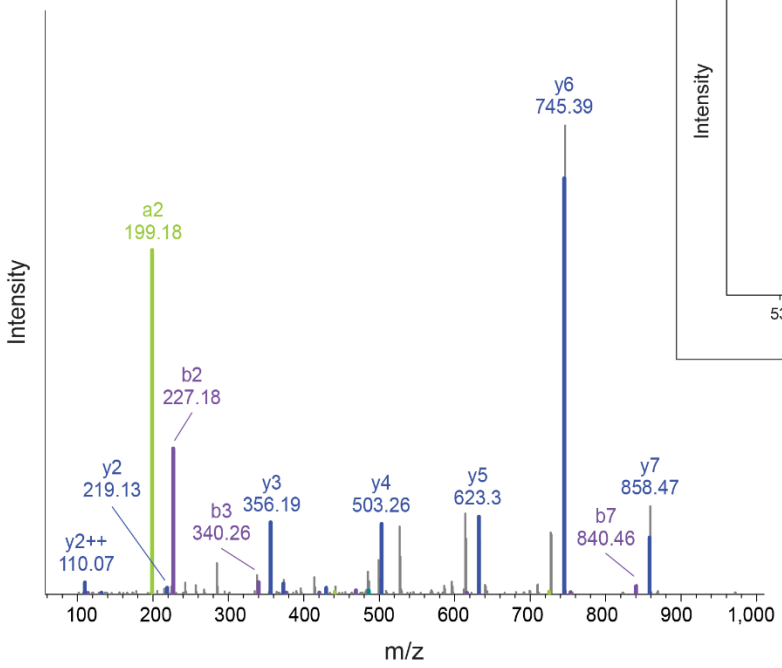
Endogenous peptide

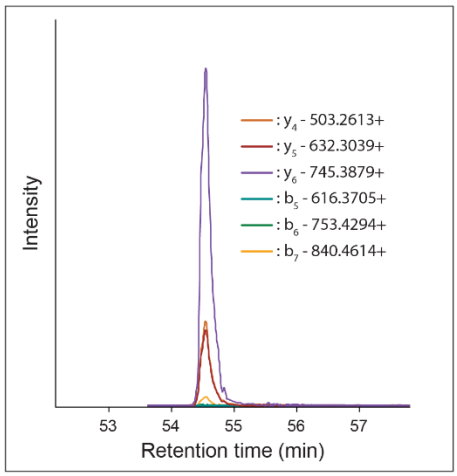


P

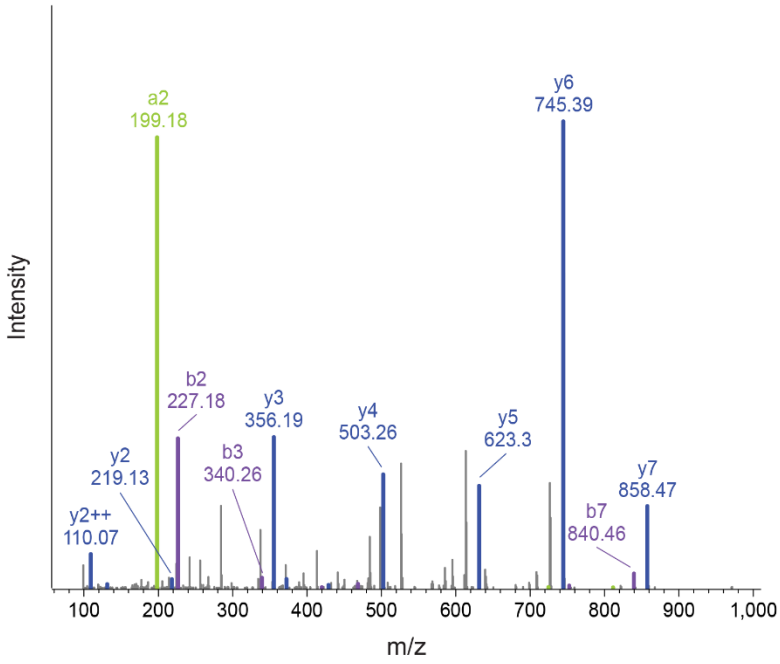
IILEFHSL - aeTSA

Synthetic peptide





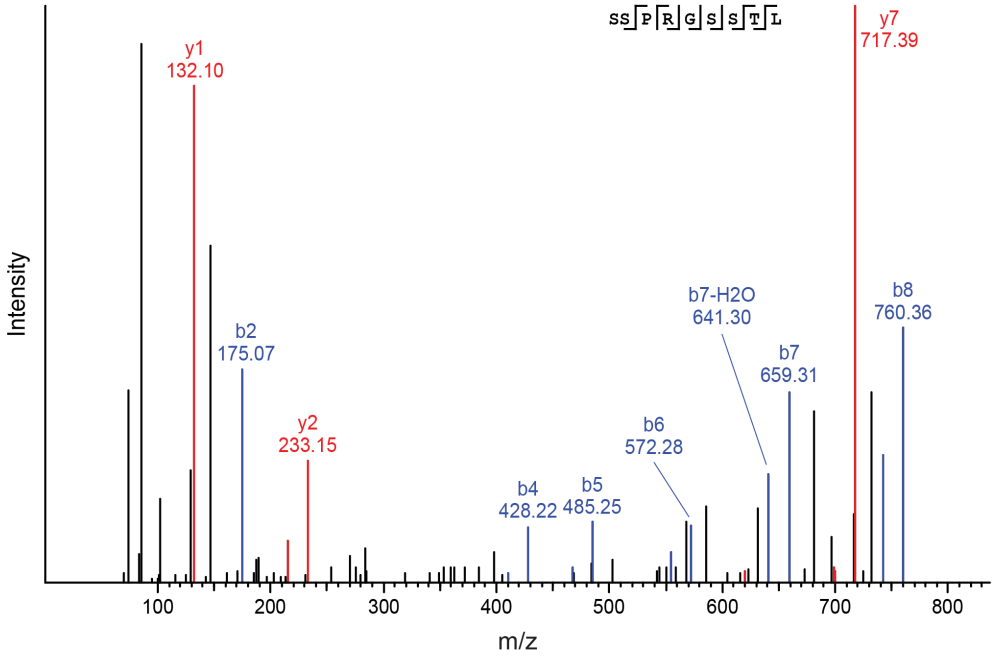
Endogenous peptide



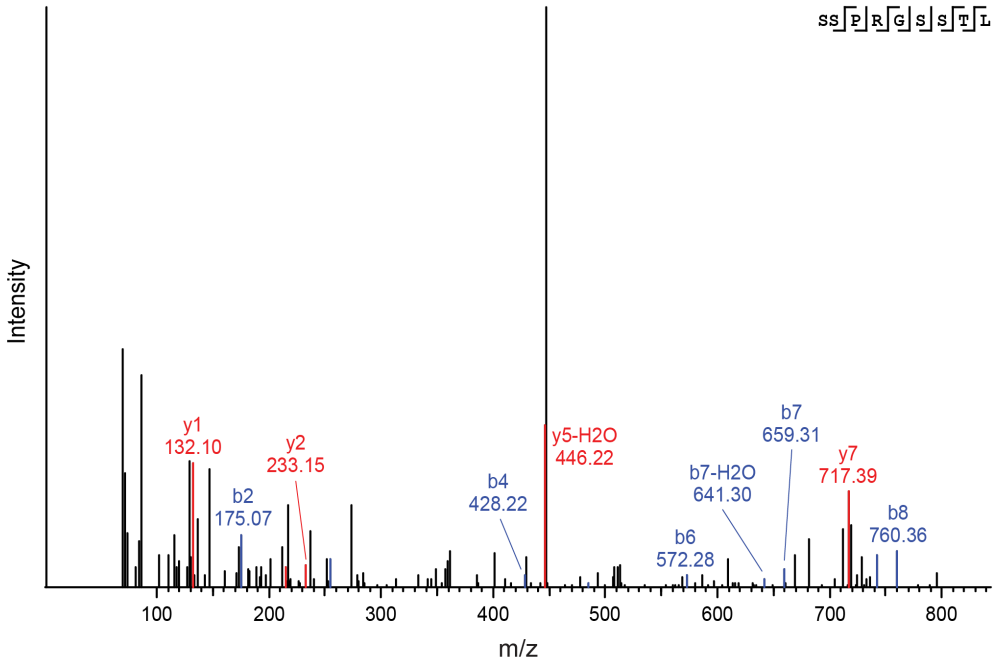
Q

SSPRGSSTL - ERE aeTSA

Synthetic peptide



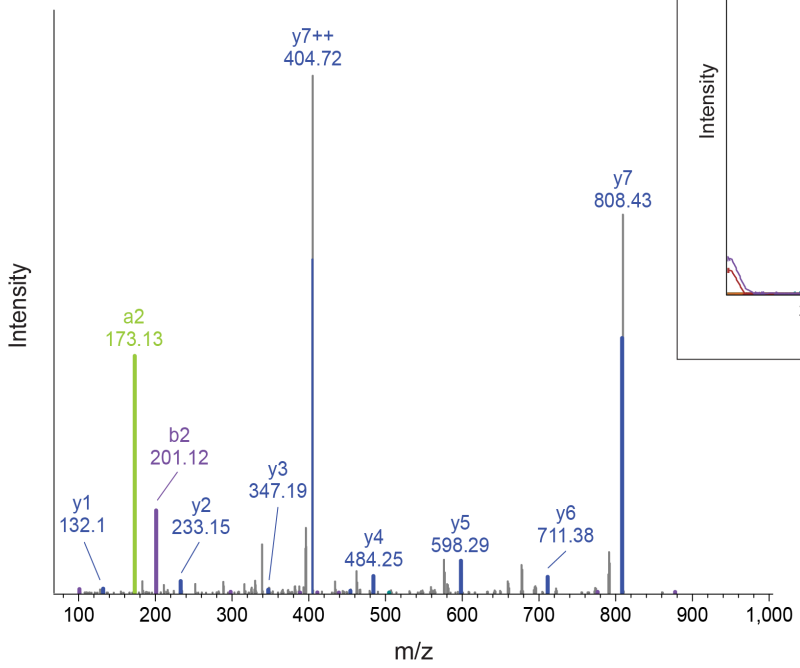
Endogenous peptide

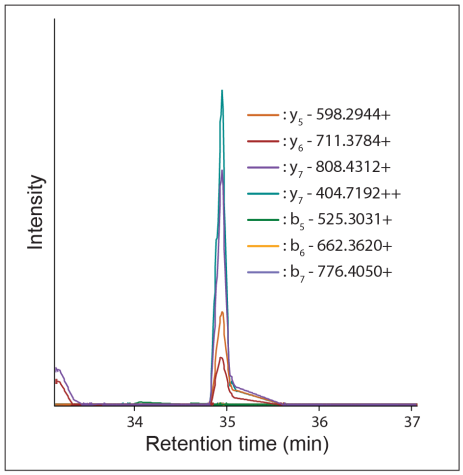


R

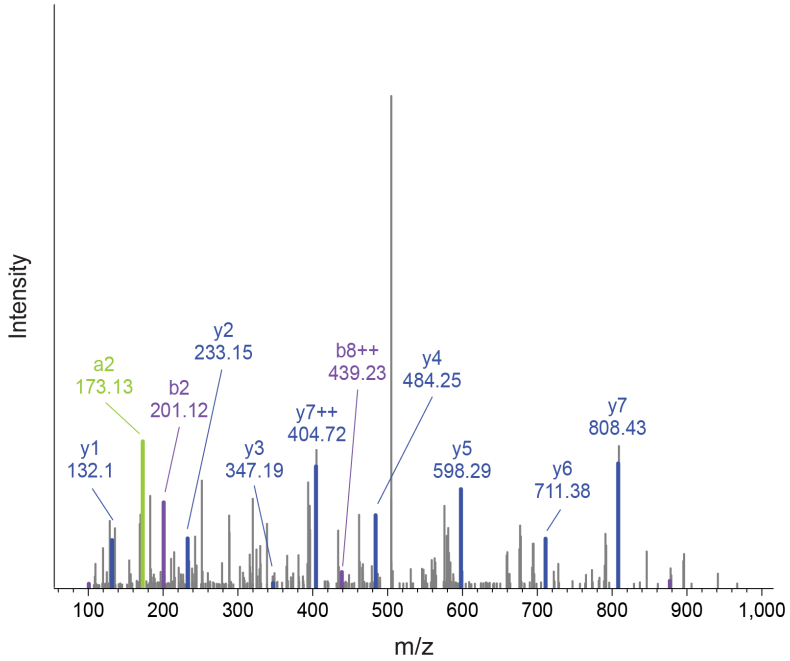
TVPLNHNTL - aeTSA

Synthetic peptide





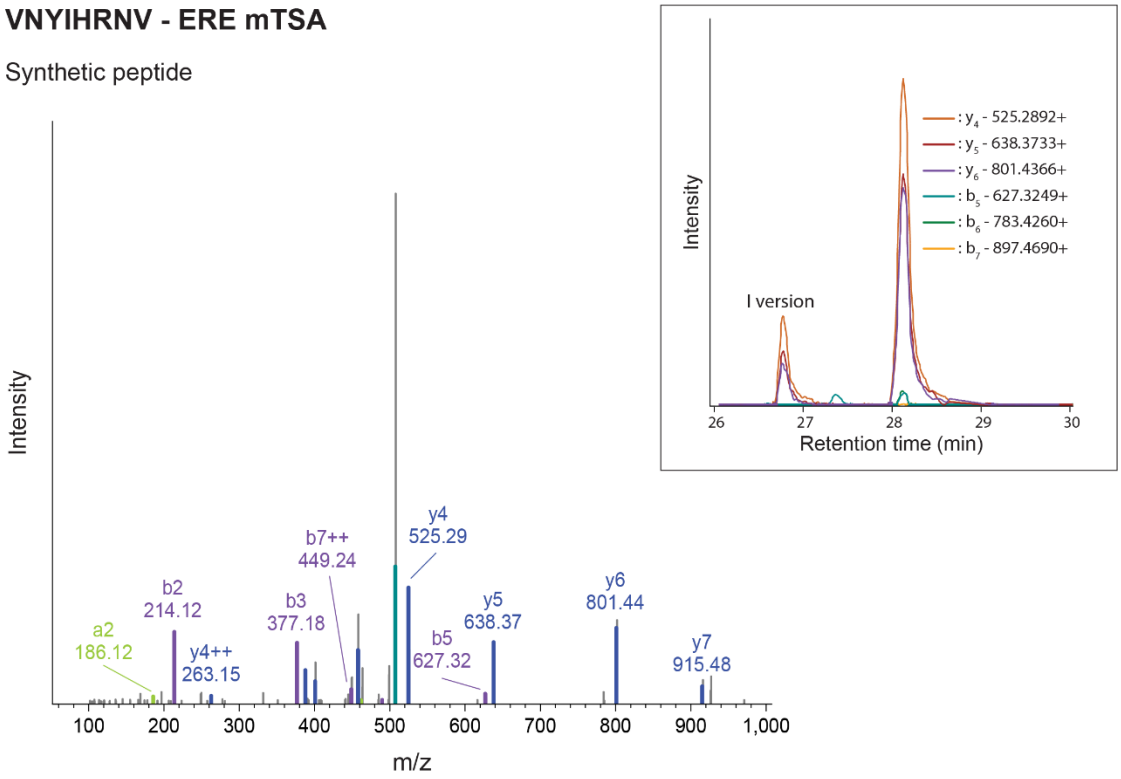
Endogenous peptide



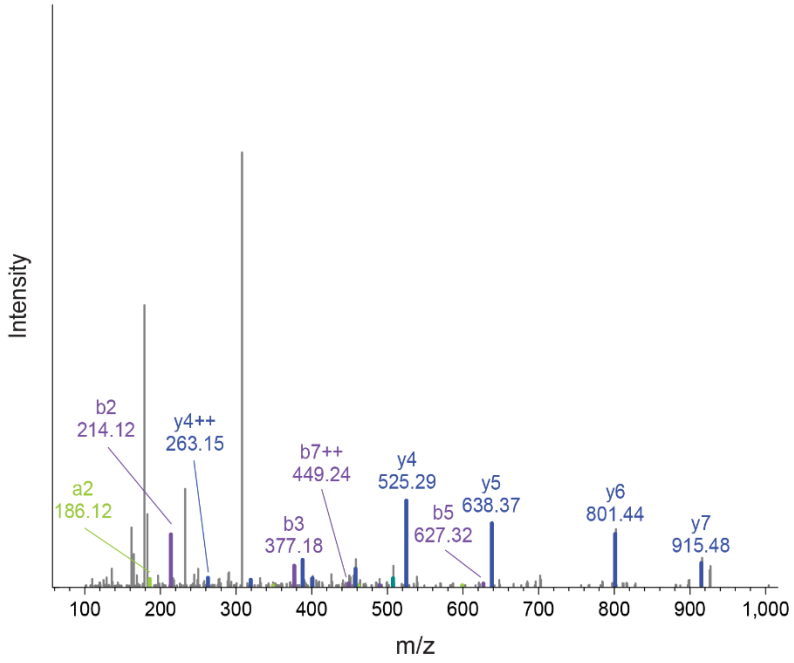
S

VNYIHRNV - ERE mTSA

Synthetic peptide



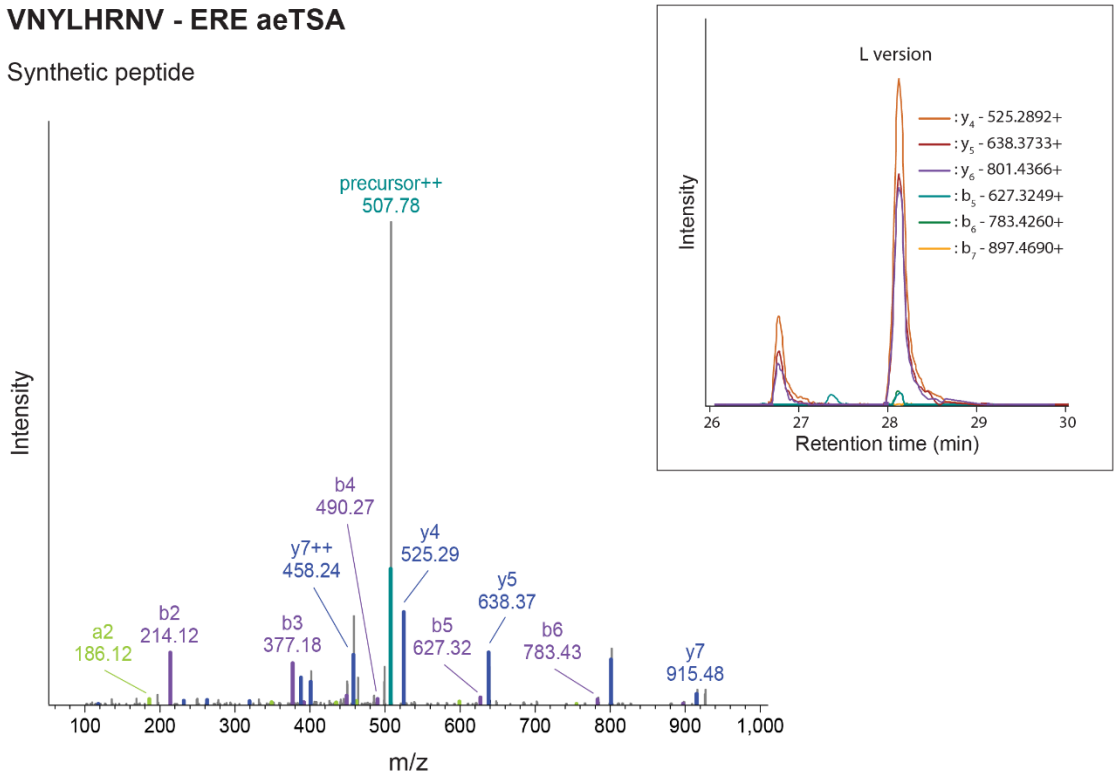
Endogenous peptide



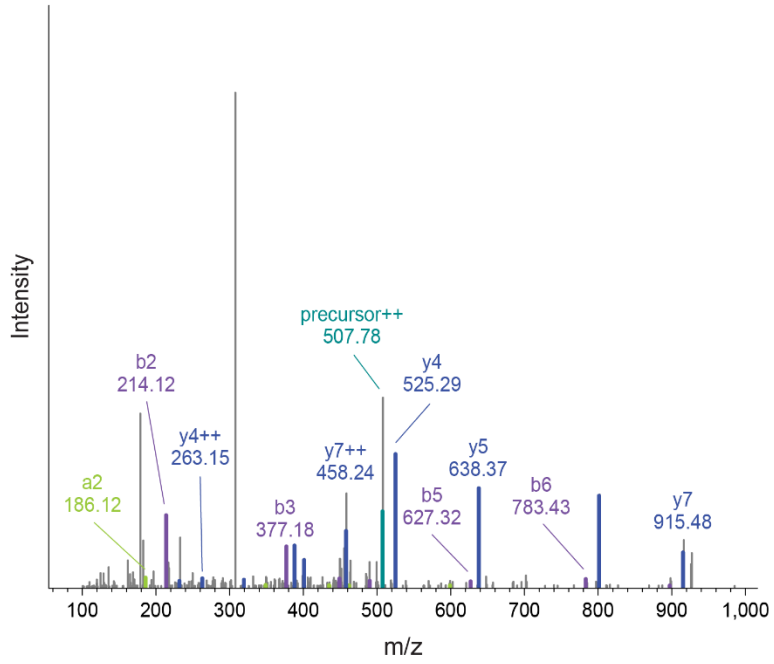
T

VNYLHRNV - ERE aeTSA

Synthetic peptide



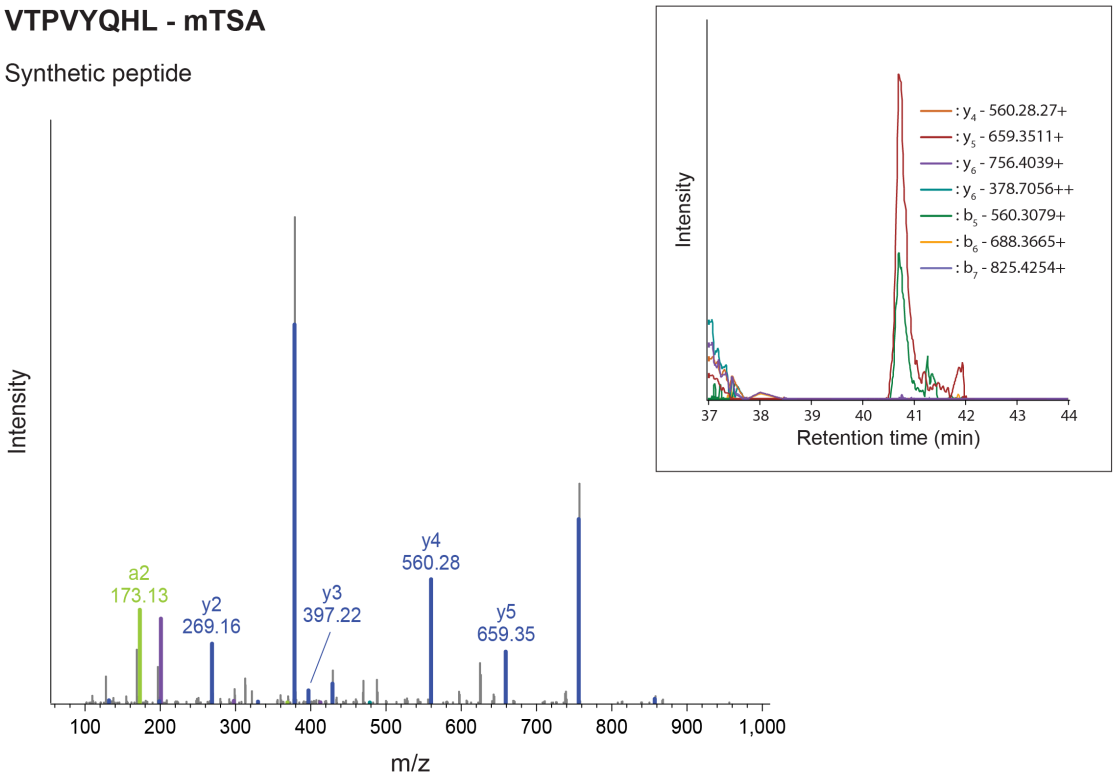
Endogenous peptide



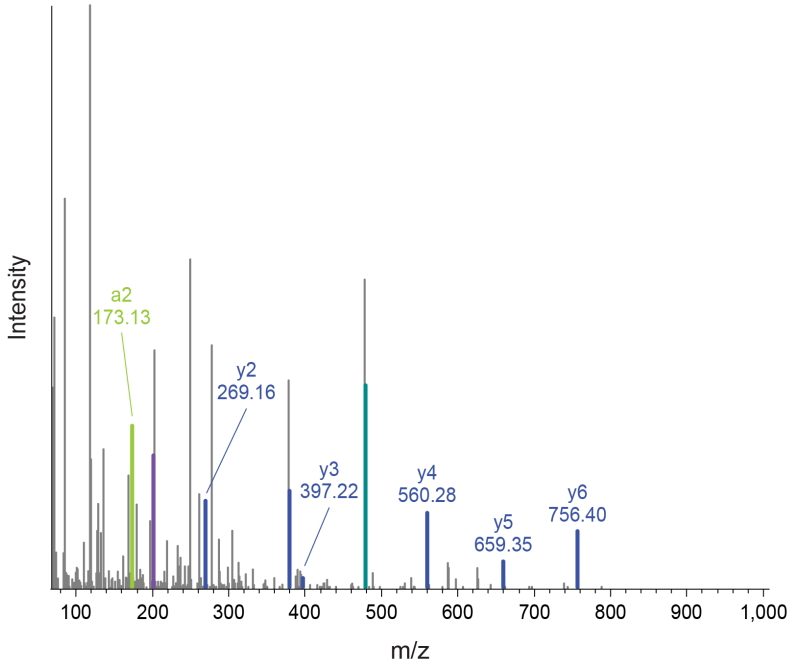
U

VTPVYQHL - mTSA

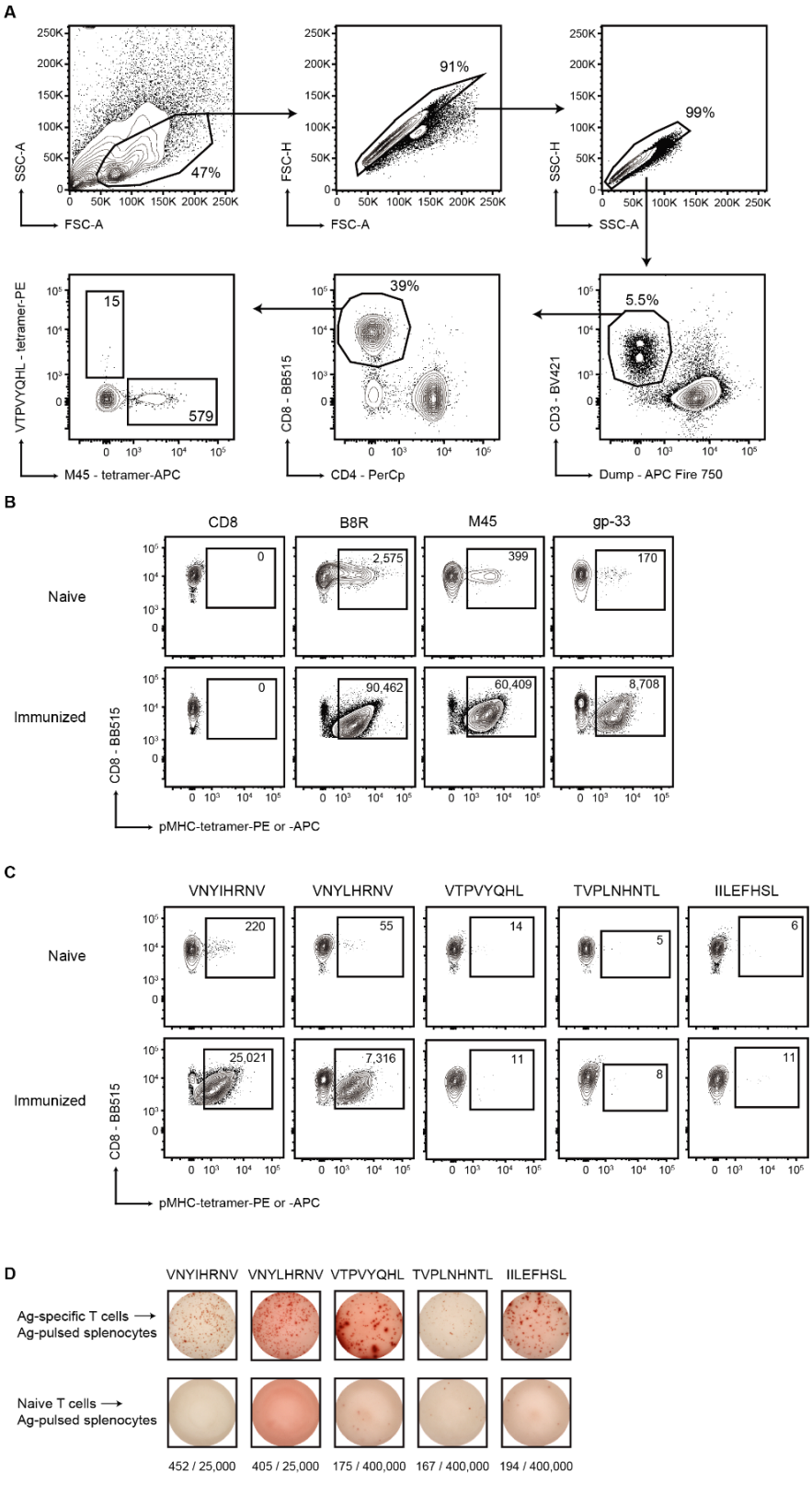
Synthetic peptide



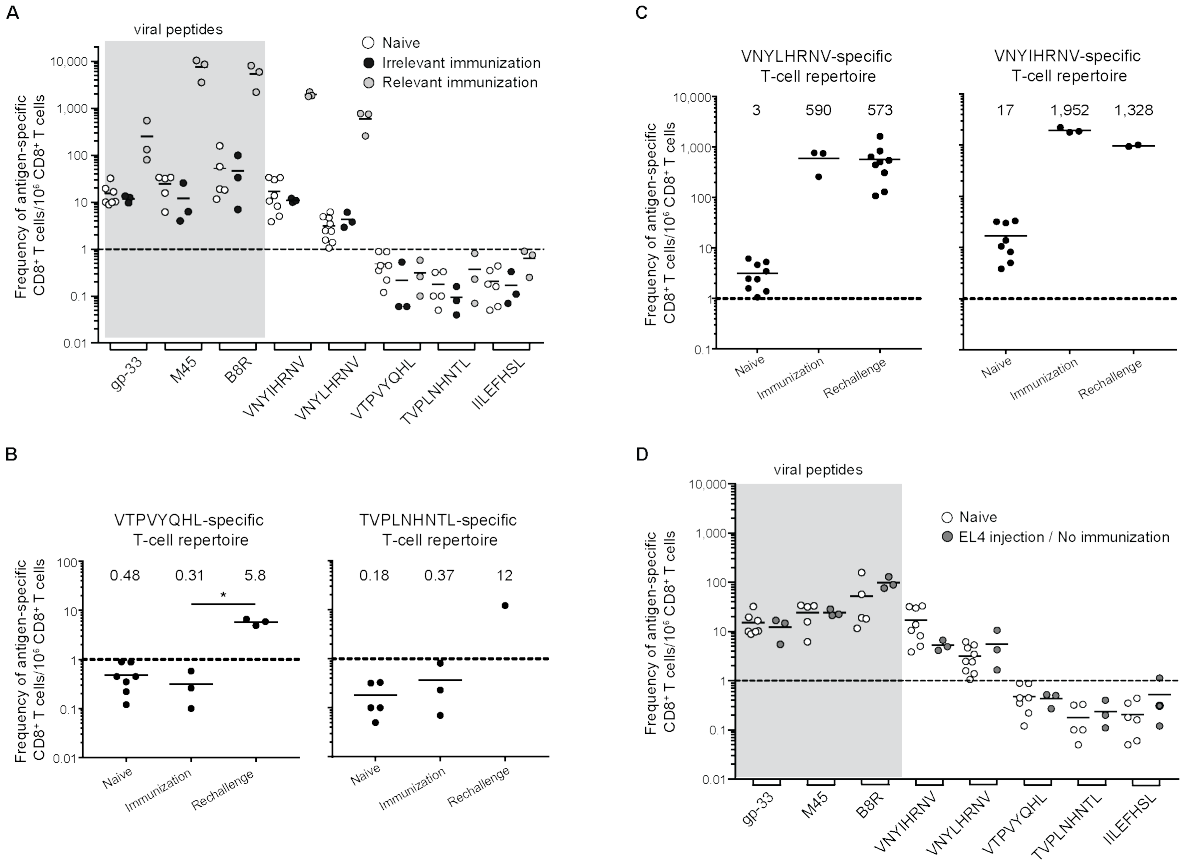
Endogenous peptide



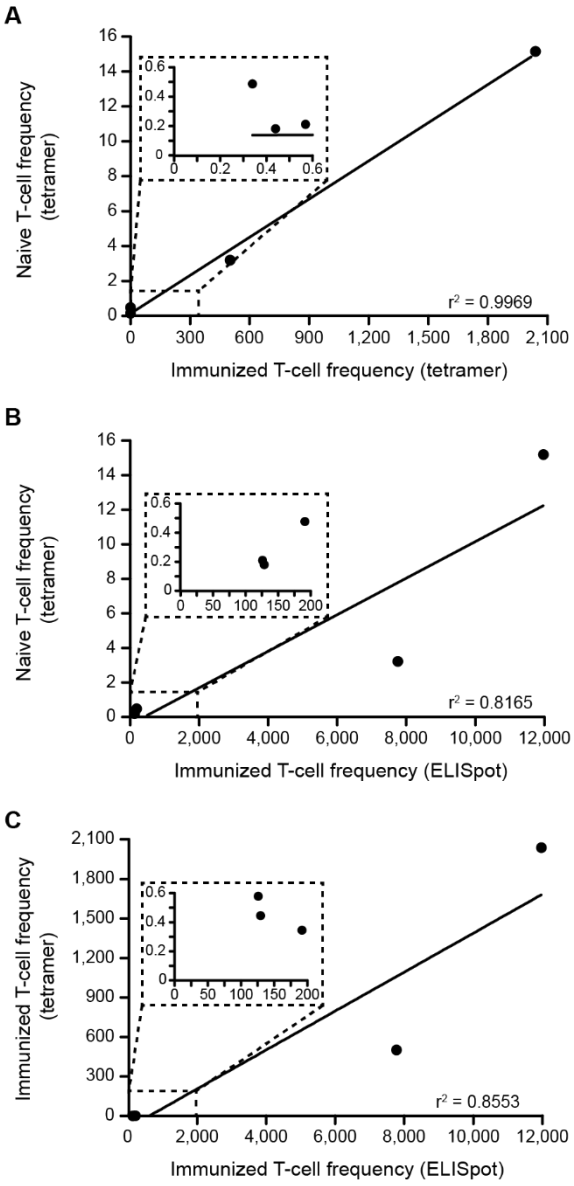
Supplementary Figure 4.4 | MS validation of CT26 and EL4 TSA candidates using synthetic analogs. (A-N) Synthetic and endogenous MS/MS spectra for CT26 TSA candidates. (**O-U**) Synthetic and endogenous MS/MS spectra for EL4 TSA candidates. Spectra presented in panels **P** and **R-U** come from additional EL4 replicates analyzed by PRM-MS. See **section 4.8.17** for details.


A
SSC-A
FSC-A
47%
FSC-H
91%
SSC-H
99%
VTPVYQHL - tetramer-PE
M45 - tetramer-APC
15
579
CD8 - BB515
CD4 - PerCp
39%
CD3 - BV421
Dump - APC Fire 750
5.5%
B
CD8
B8R
M45
gp-33
Naive
Immunized
0
2,575
399
170
90,462
60,409
8,708
CD8 - BB515
pMHC-tetramer-PE or -APC
C
VNYIHRNV
VNYLHRNV
VTPVYQHL
TVPLNHNTL
IILEFHSL
220
55
14
5
6
25,021
7,316
11
8
11
D
Ag-specific T cells
Ag-pulsed splenocytes
Naive T cells
452 / 25,000
405 / 25,000
175 / 400,000
167 / 400,000
194 / 400,000

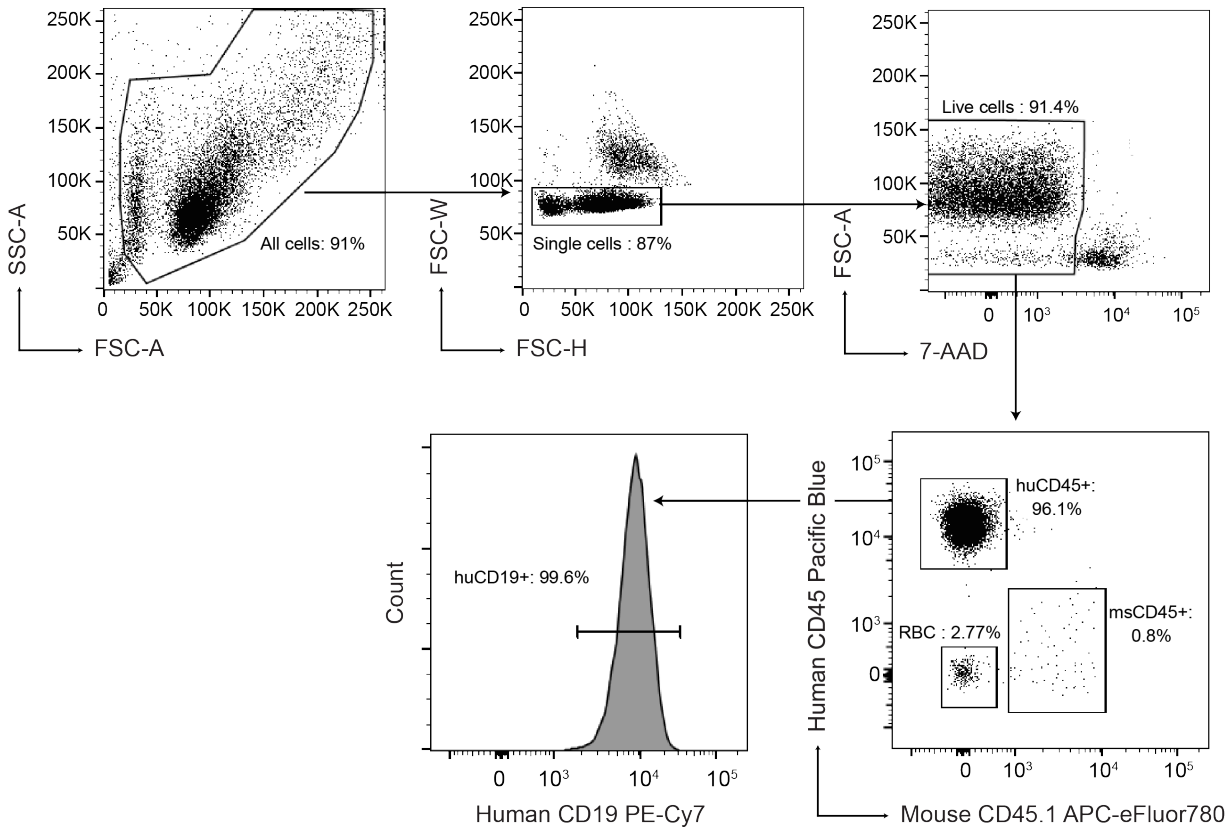
Supplementary Figure 4.5 | Detection of antigen-specific CD8$^+$ T cells in naive and immunized mice. (**A**) Gating strategy for the detection of tetramer$^+$ CD8$^+$ T cells *ex vivo*. Tetramer enrichment were performed on single-cell suspensions isolated from the spleen and lymph nodes of each mice. After doublets exclusion, Dump$^-$ CD3$^+$ cells were analyzed for CD8 and CD4 expression and tetramer$^+$ cells were analyzed in the CD8$^+$ compartment. A representative staining obtained following VTPVYQHL/H-2-K^b-PE and M45/H-2-D^b-APC tetramers enrichment in a naive mouse is shown. Absolute numbers of tetramer$^+$ CD8$^+$ T cells detected for each specificity are indicated. The Dump channel corresponds to pooled events positive for 7-AAD, CD45R and CD19, F4/80, CD11b and CD11c. (**B** and **C**) Representative analysis of tetramer$^+$ CD8$^+$ T cells in naive (upper row) and immunized (lower row) mice. CD8$^+$ cells before magnetic enrichment and after ex vivo enrichment for tetramer$^+$ viral specificities (**B**) as well as for TSA specificities (**C**) are shown. Percentages and numbers of tetramer$^+$ or tetramer$^-$ cells are indicated. (**D**) One representative experiment of the frequency of IFN-γ secreting CD8$^+$ T cells in immunized and naive mice. The number of spot forming cells relative to the number of plated CD8$^+$ T cells in each condition are indicated below each well.



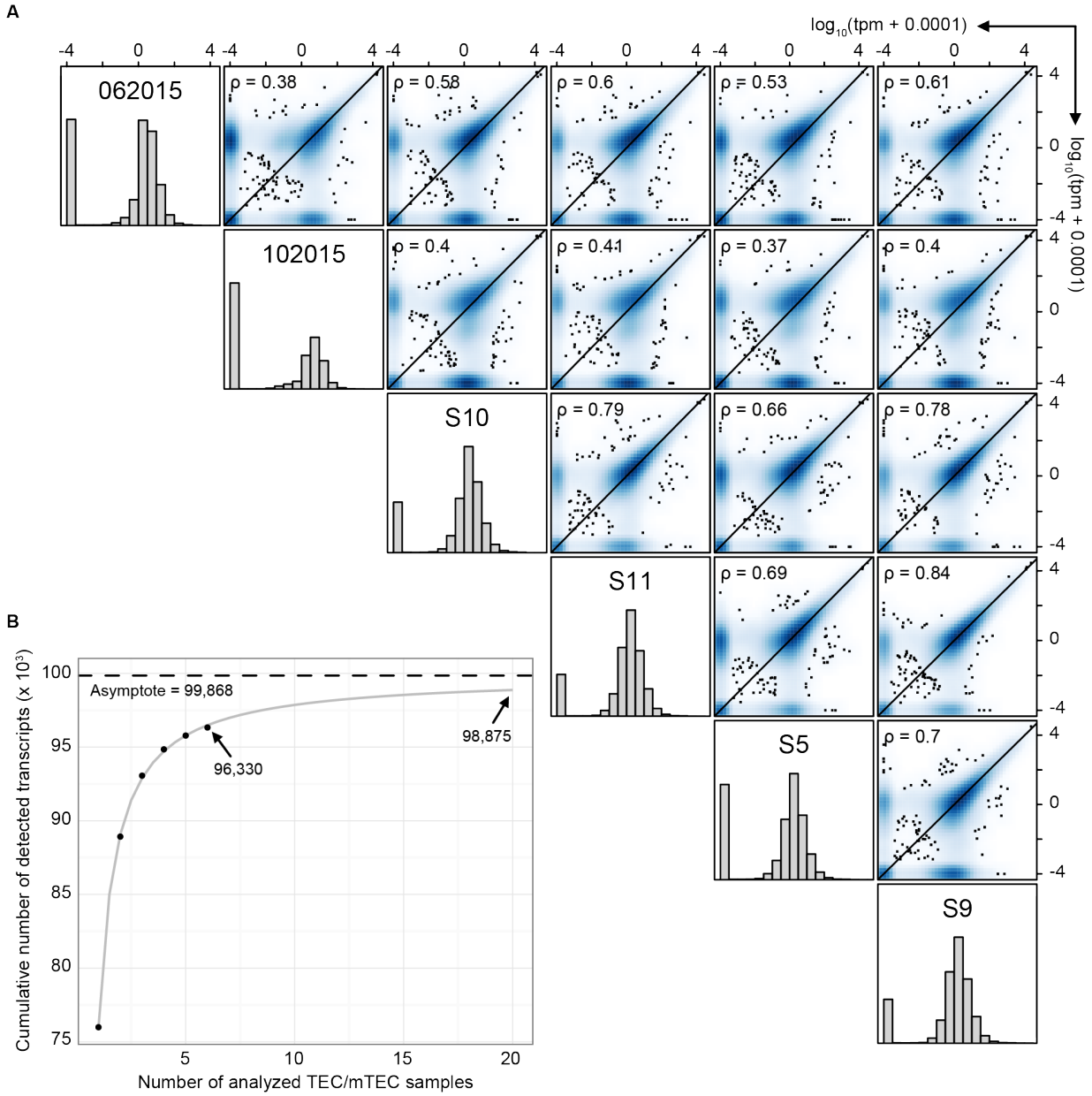
Supplementary Figure 4.6 | Frequencies of antigen-specific T cells. (**A**) Frequencies of antigen-specific T cells in naïve mice and mice immunized with relevant or irrelevant peptides. (**B** and **C**) Frequencies of antigen-specific $CD8^+$ T cells in mice immunized against VTPVYQHL or TVPLNHNTL (**B**) or against VNYLHRNV or VNYIHRNV (**C**) that were rechallenged with EL4 cells at day 150. For comparison purposes, frequencies of antigen-specific T cells in naive and immunized mice reported in panel (**A**) are reproduced. (**D**) Frequencies of antigen-specific T cells in non-immunized mice injected with EL4 cells. All calculated frequencies of $tetramer^+$ $CD8^+$ T cells are expressed as the number of antigen-specific $CD8^+$ T cells per 10^6 $CD8^+$ T cell. Each symbol represents one mouse (n = 1 to 9 mice). Dotted line represents a minimal detection level of one $tetramer^+$ T cell per 10^6 $CD8^+$ T cells. Viral peptides used as controls are highlighted in gray. p-values were calculated using one-sided Wilcoxon rank sum test (* $p \leq 0.05$).



Supplementary Figure 4.7 | Correlation between antigen-specific T cell frequencies in naive and immunized mice. (**A** and **B**) Correlation between the frequencies of antigen-specific $CD8^+$ T cells in naive and immunized mice as calculated by (**A**) tetramer staining and (**B**) IFN-γ ELISpot assays. (**C**) Correlation between the frequencies of antigen-specific $CD8^+$ T cells in immunized mice as calculated by tetramer staining and IFN-γ ELISpot assays. Average frequencies were used for plotting data. Fitness of curves was determined by the coefficient of determination (r^2).



Supplementary Figure 4.8 | Purity of the 10H080 B-ALL sample following expansion in NSG mice. After isolation of B-ALL from NSG mice, purity and viability were assessed using flow cytometry. 0.5 x 10^6 cells were stained with anti-human CD45, anti-human CD19, anti-mouse CD45.1 and 7-AAD. Dot plots showing the gating strategy leading to identification of B-ALL cells in one representative sample. B-ALL cells were defined as 7-AAD$^-$huCD45$^+$ cells that homogeneously expressed huCD19 and always represented about 96% of harvested cells. Remaining contaminants after Ficoll gradient were composed of red blood cells (RBC, 2-3%) that do not express MHC at their surface and murine CD45$^+$ cells (about 1%).



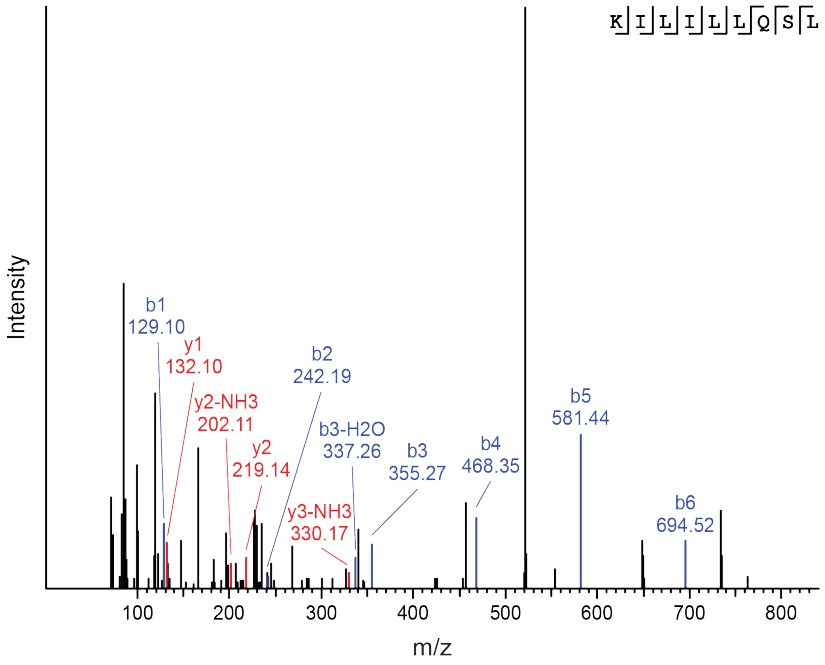
Supplementary Figure 4.9 | Overview of the human TEC and mTEC transcriptomic landscapes. (**A**) Human TEC (062015 and 102015) and mTEC (S5 to S11) isolated from unrelated donors display similar transcriptomic profiles. Following RNA-Seq, we selected transcripts expressed in at least one donor with a tpm > 1, as estimated by kallisto, to plot all one-to-one scatter plots. The Spearman's rank correlation coefficient (ρ) is indicated at the top left corner of each graph and the black line represents identical expression of transcripts. (**B**) RNA-Seq of additional human TEC/mTEC samples should result in a minimal gain of information. Using our set of expressed transcripts (tpm > 1 in at least one sample), we extrapolated the cumulative number of transcripts (*cT*) that should be detected by adding additional samples to our cohorts (*nS*, see **section 4.8.18** for details) using the following

function: $cT = \frac{a \times (nS-1)}{[b+(nS-1)]} + c$ with *a* = 23,892.73, *b* = 0.8243389 and *c* = 75,976.11 (grey line). On the graph, we indicated the cumulative number of transcripts detected by analyzing *nS* = 6 (our cohort, black dots) or *nS* = 20 samples, as well as the total number of transcripts that should be detected, which corresponds to $\lim_{nS\to\infty}\left(\frac{a \times (nS-1)}{[b+(nS-1)]} + c\right) = a + c = 99{,}868$ (asymptote value).

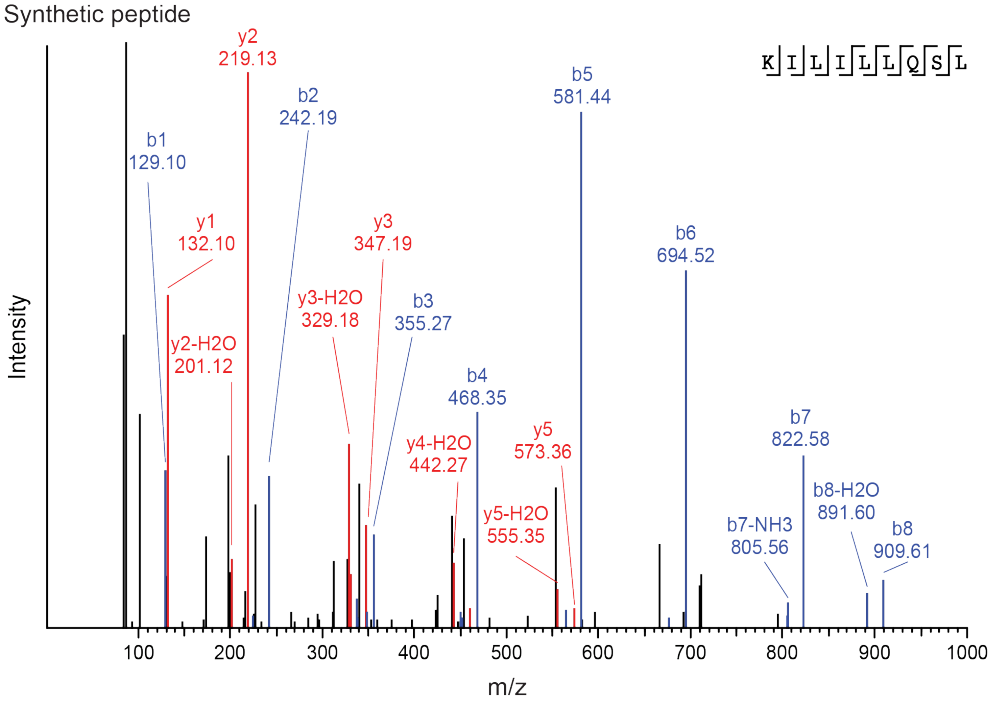
A

KILILLQSL - ERE aeTSA

Endogenous peptide



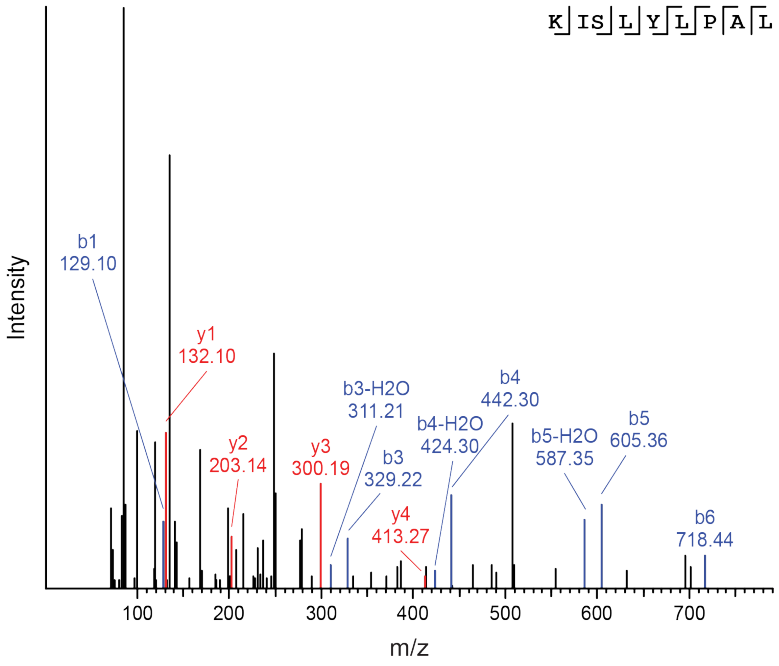
Synthetic peptide



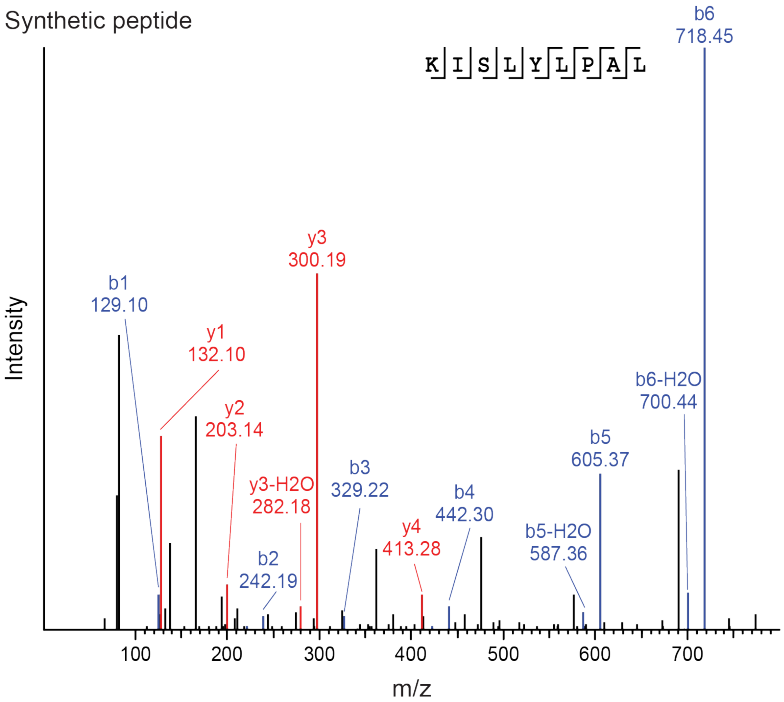
B

KISLYLPAL - ERE aeTSA

Endogenous peptide



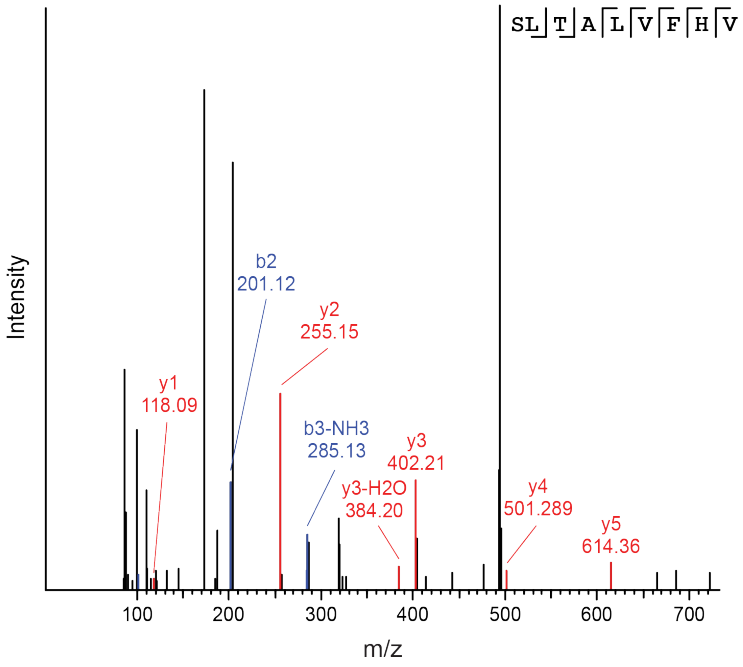
Synthetic peptide



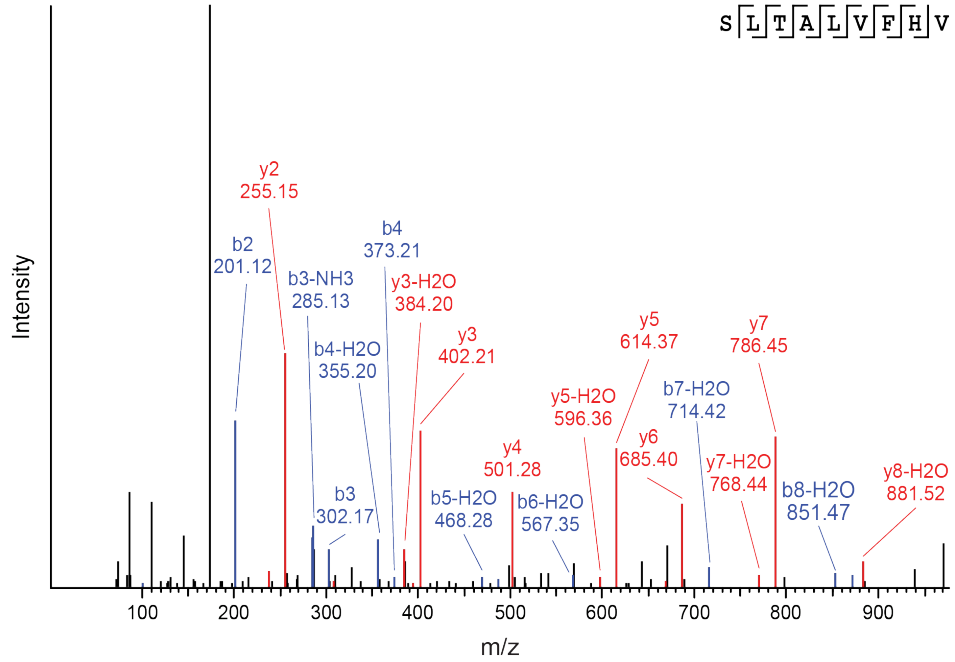
C

SLTALVFHV - aeTSA

Endogenous peptide



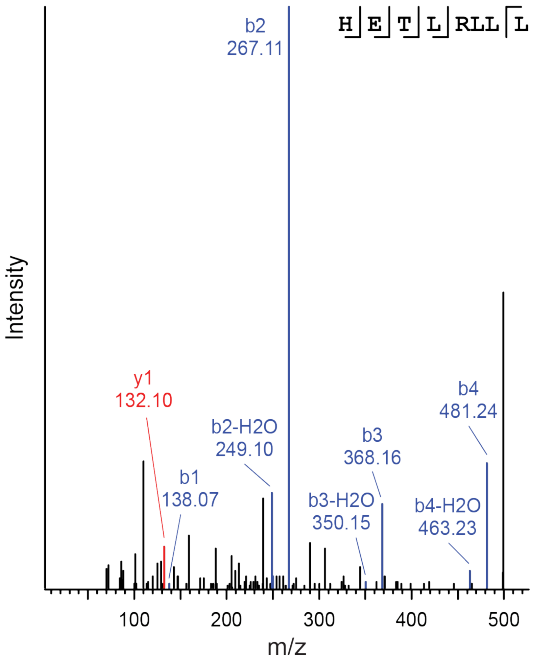
Synthetic peptide



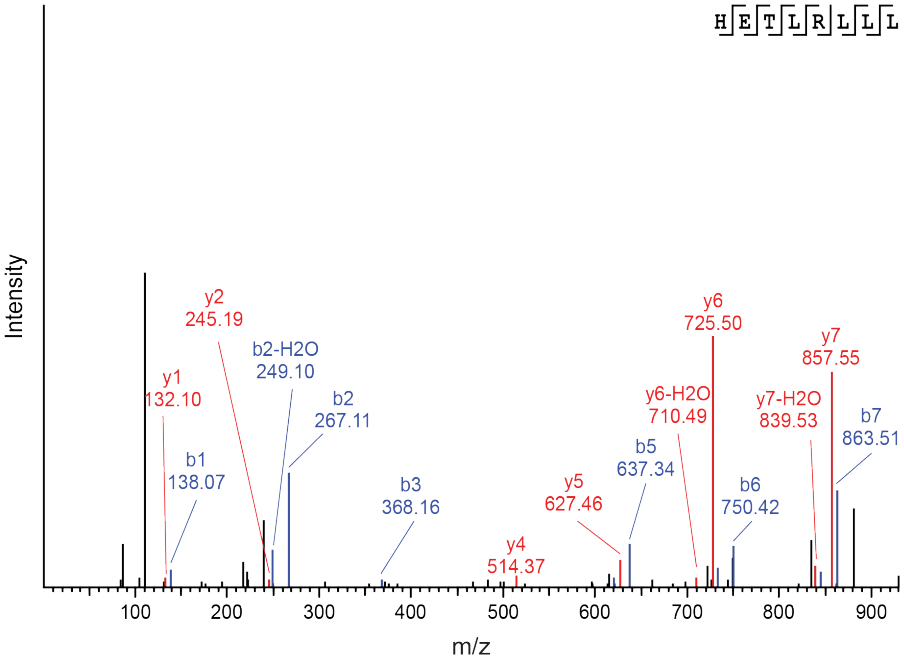
D

HETLRLLL - aeTSA

Endogenous peptide



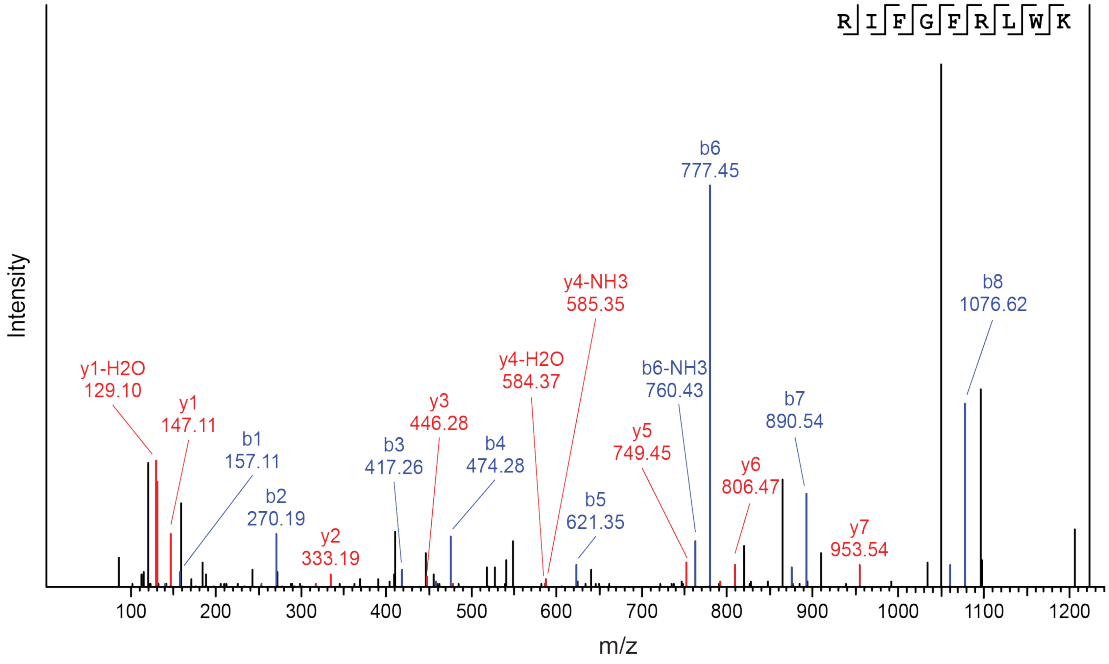
Synthetic peptide



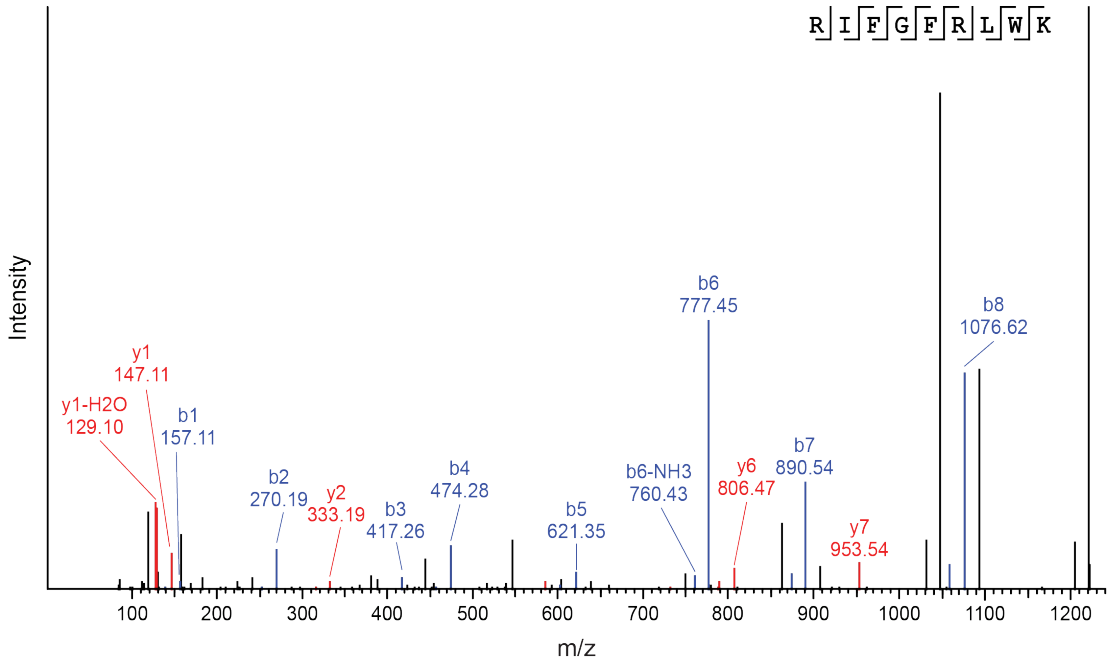
E

RIFGFRLWK - aeTSA

Endogenous peptide



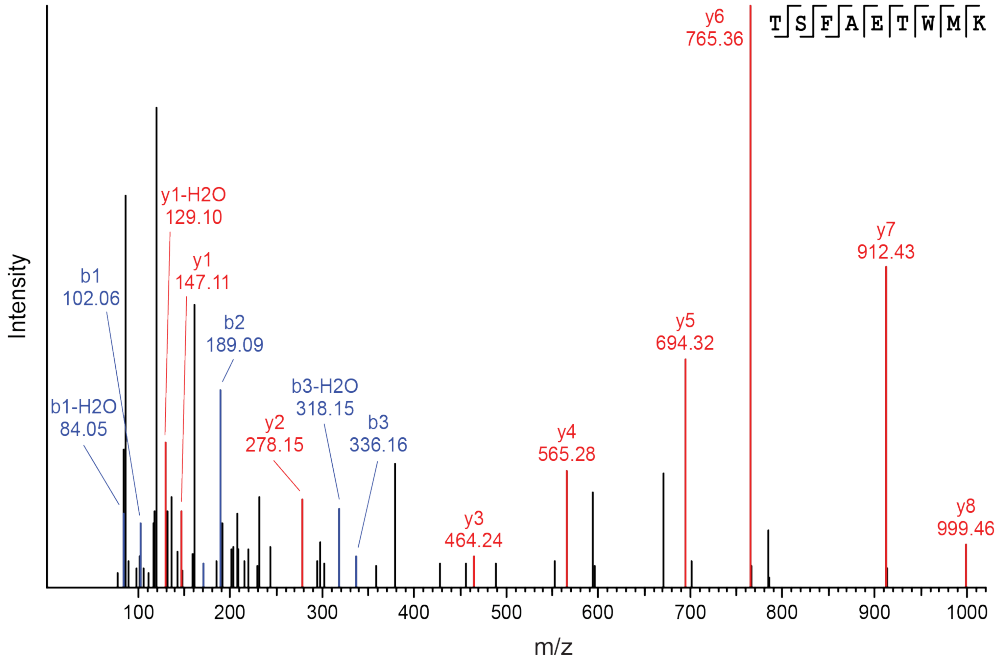
Synthetic peptide



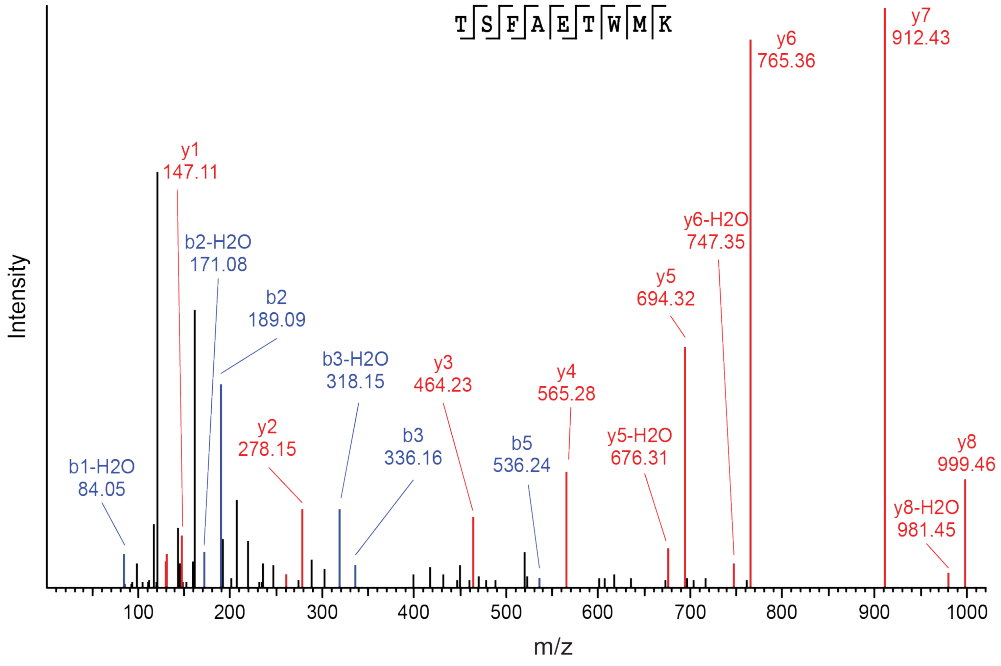
F

TSFAETWMK - ERE aeTSA

Endogenous peptide



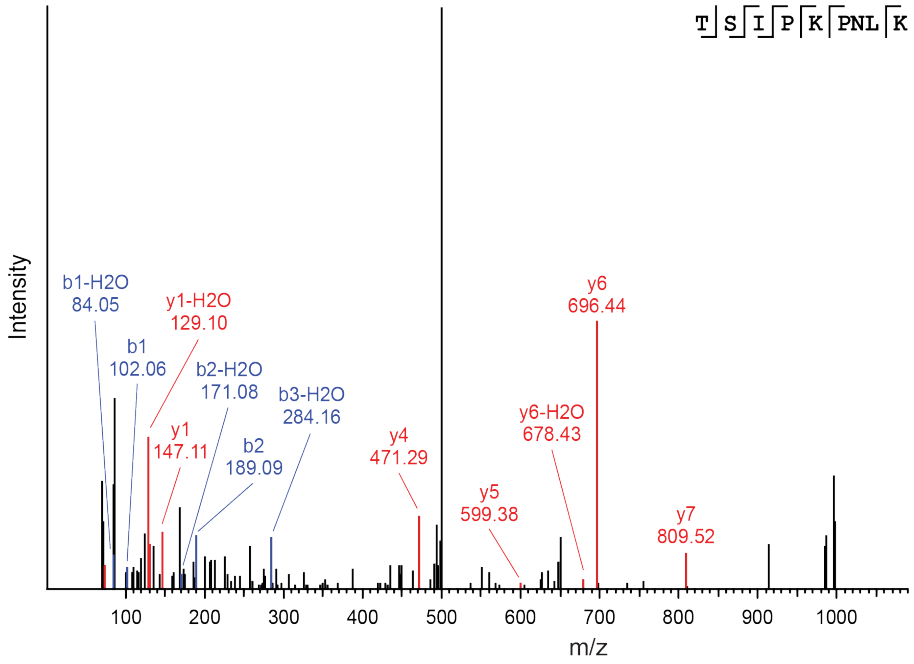
Synthetic peptide



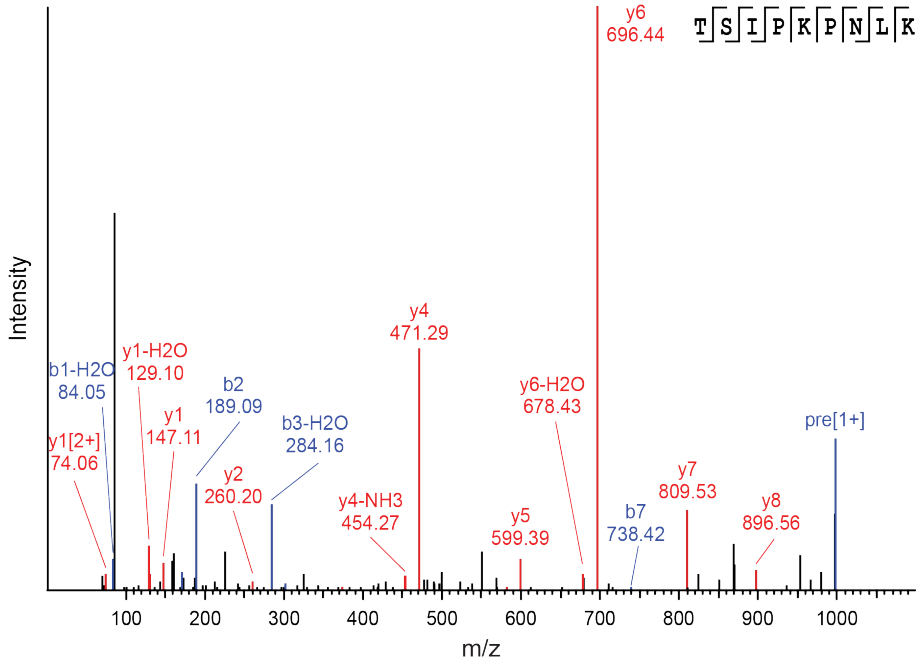
G

TSIPKPNLK - aeTSA

Endogenous peptide



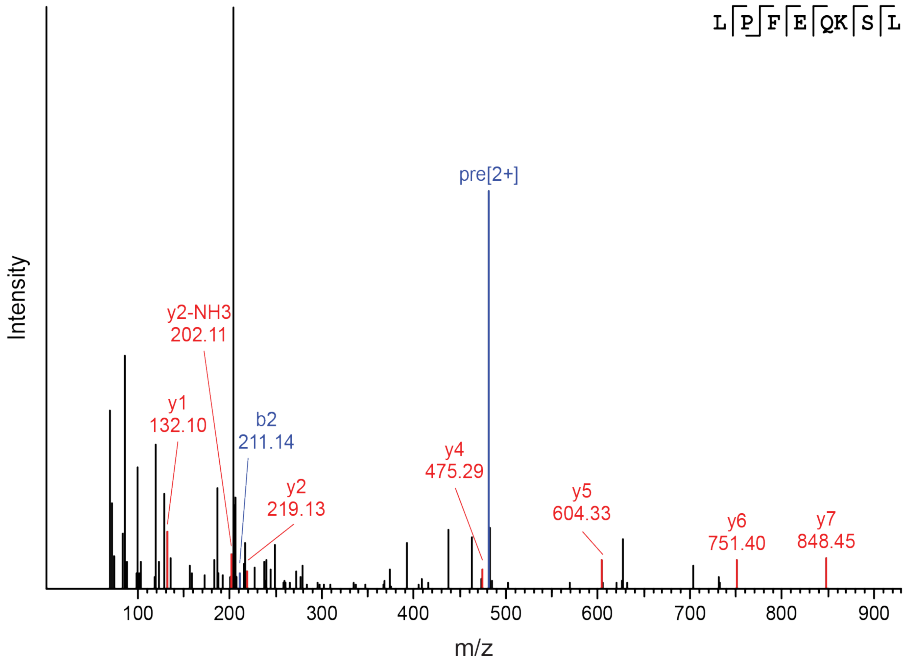
Synthetic peptide



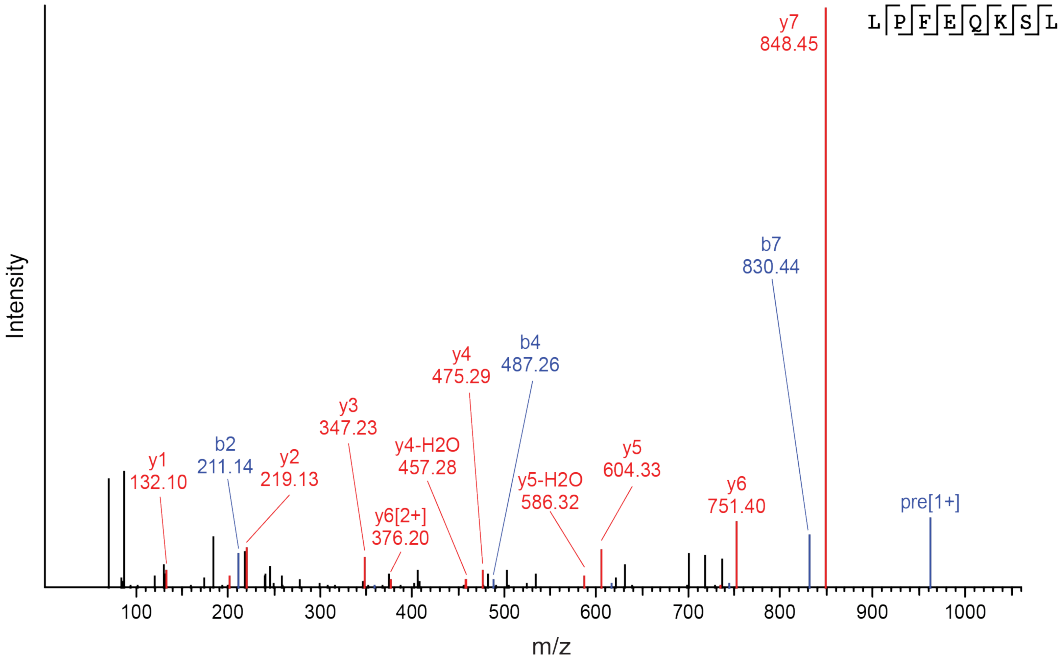
H

LPFEQKSL - aeTSA

Endogenous peptide



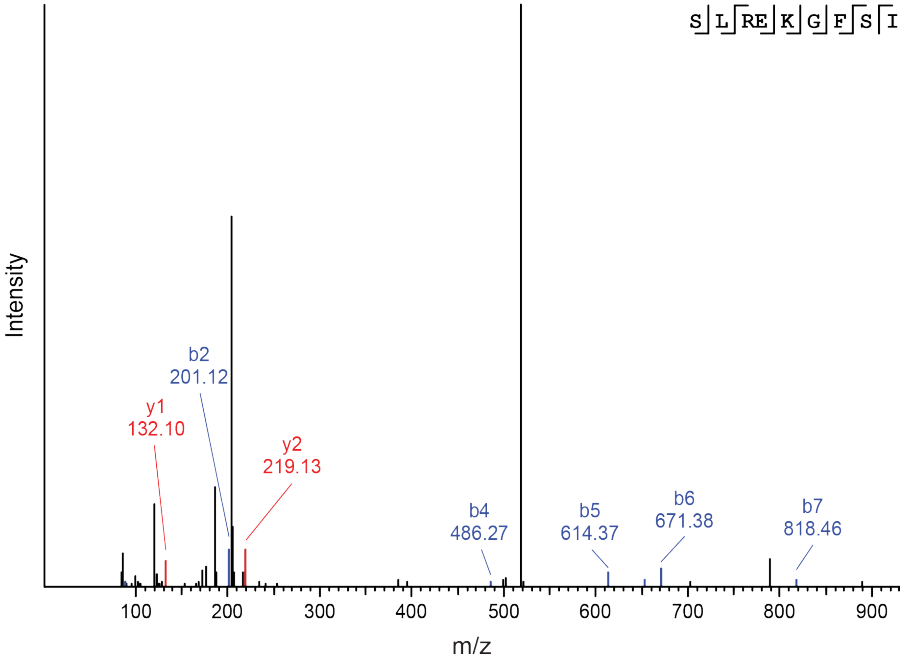
Synthetic peptide



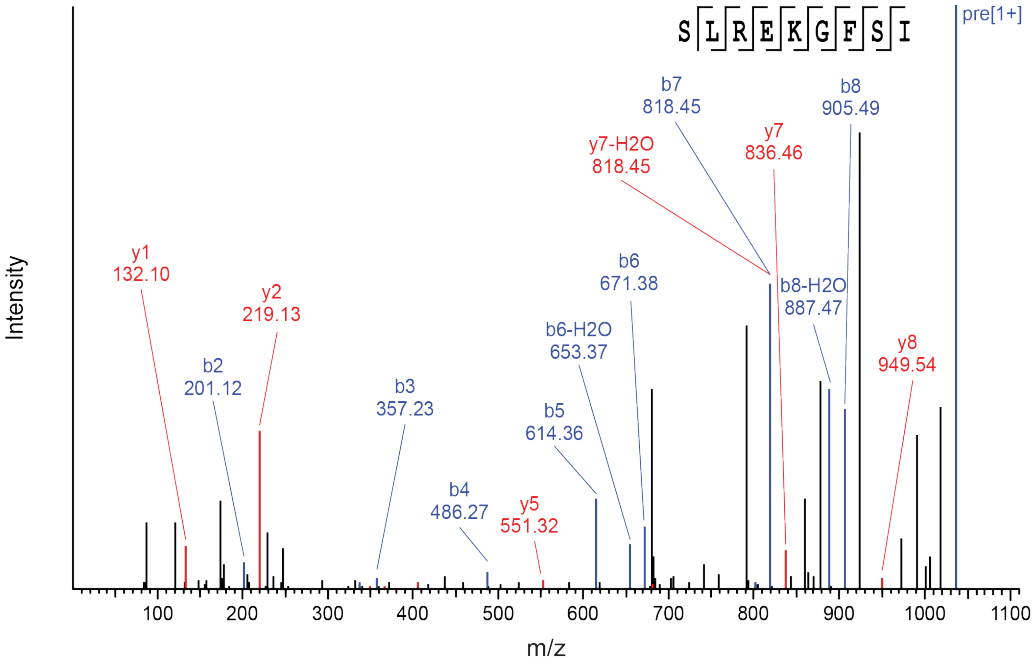
I

SLREKGFSI - aeTSA

Endogenous peptide



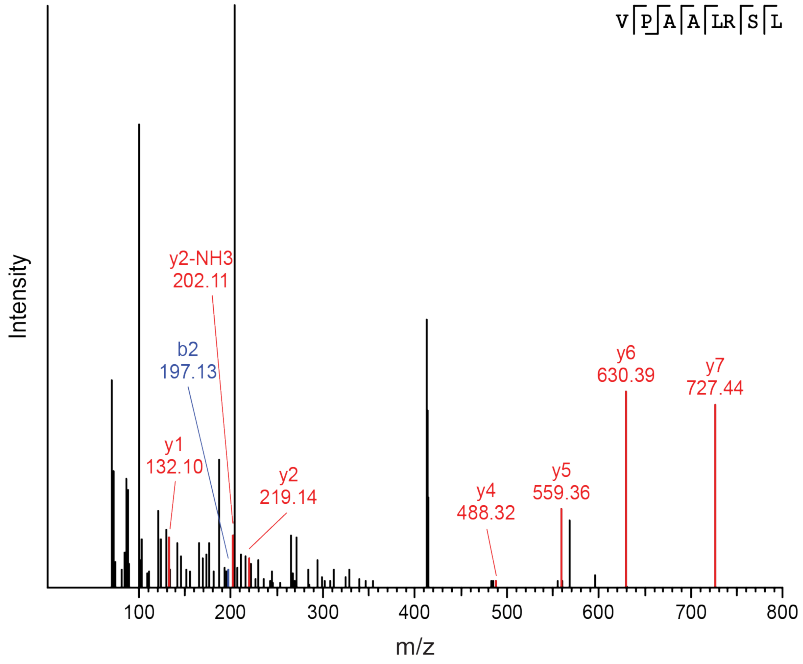
Synthetic peptide



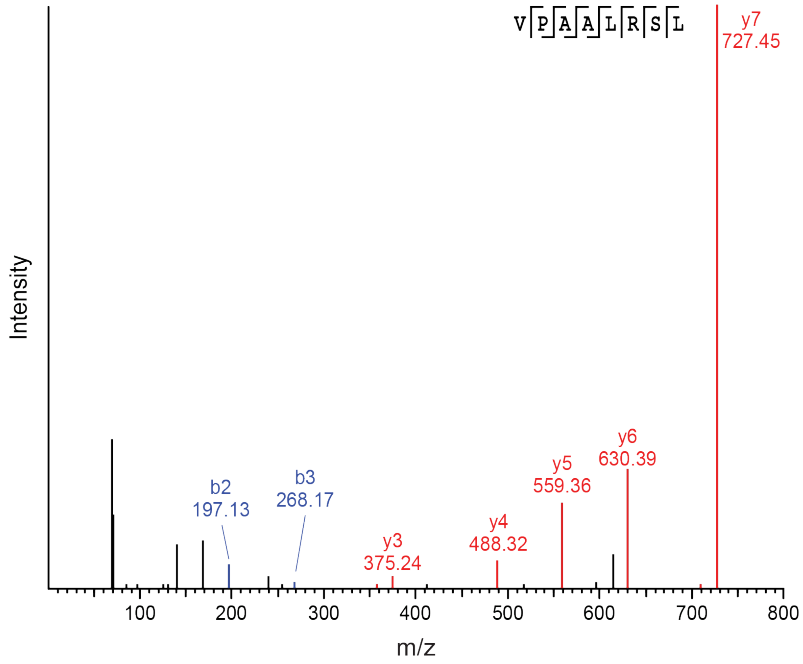
J

VPAALRSL - aeTSA

Endogenous peptide



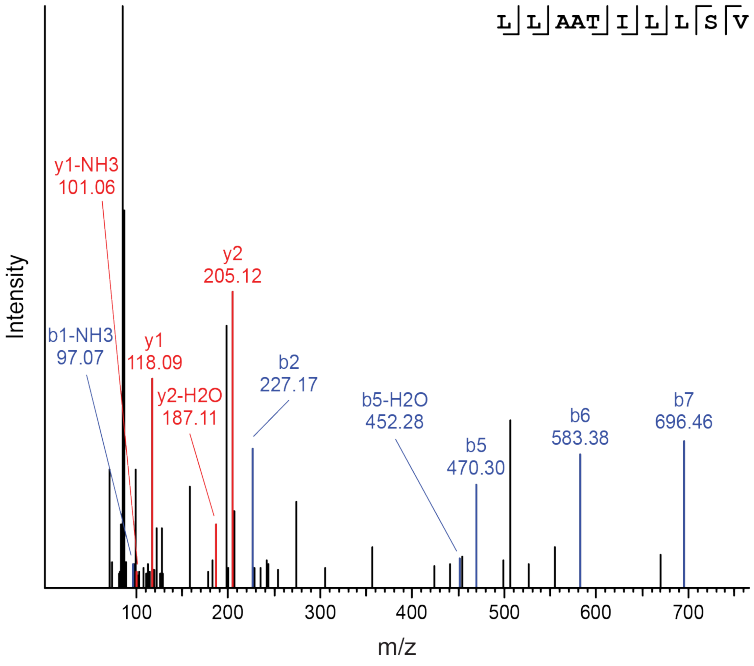
Synthetic peptide



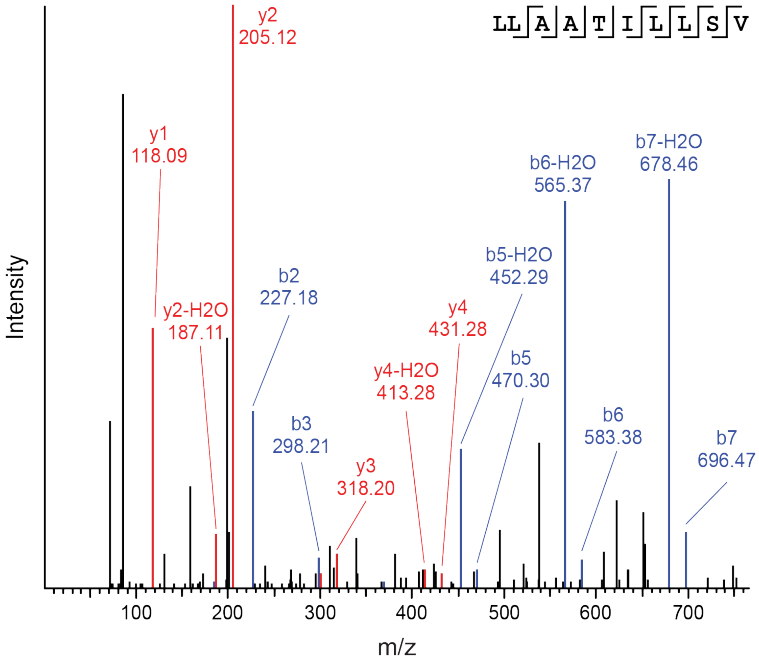
K

LLAATILLSV - aeTSA

Endogenous peptide



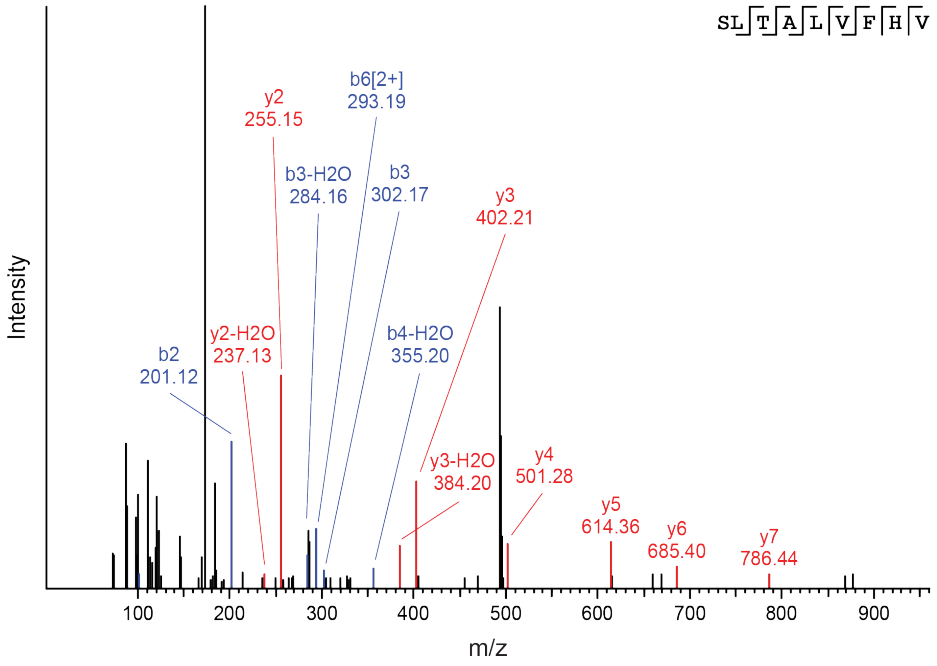
Synthetic peptide



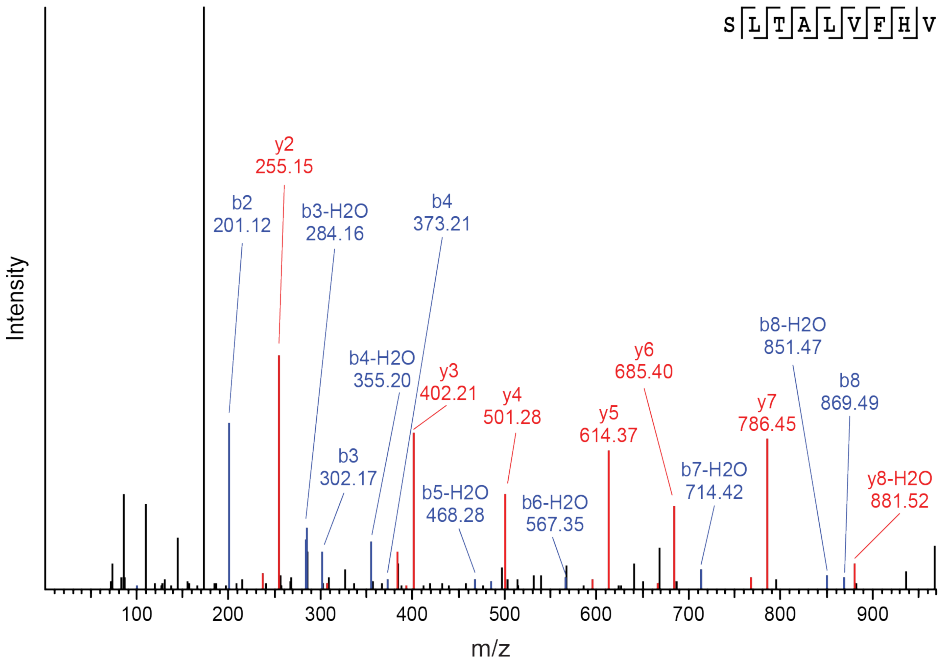
L

SLTALVFHV - aeTSA

Endogenous peptide



Synthetic peptide

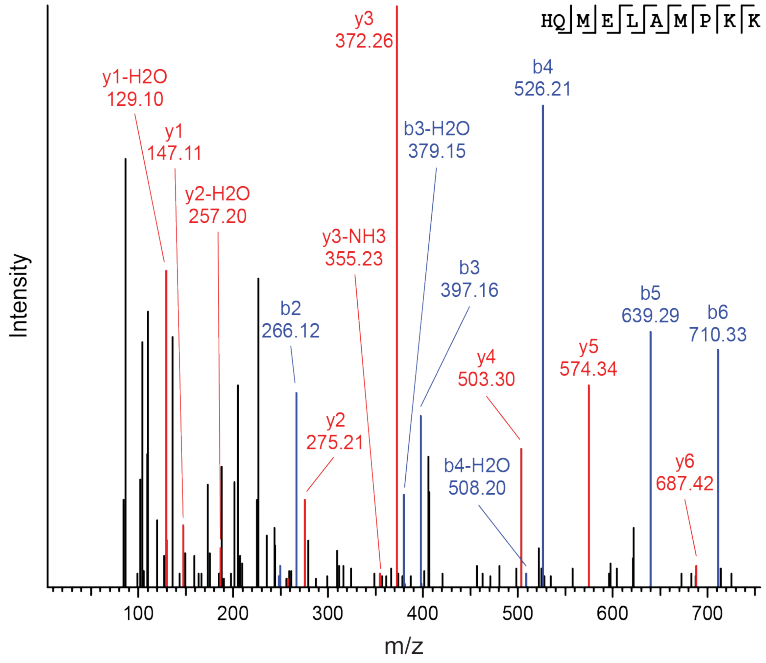


Supplementary Figure 4.10 | MS validation of B-ALL TSA candidates using synthetic analogs. Synthetic and endogenous MS/MS spectra for TSA candidates identified in each of our four B-ALL specimens: (**A-C**) 07H103, (**D-G**) 10H080, (**H-J**) 10H118 and (**K-L**) 12H018. See **section 4.8.17** for details.

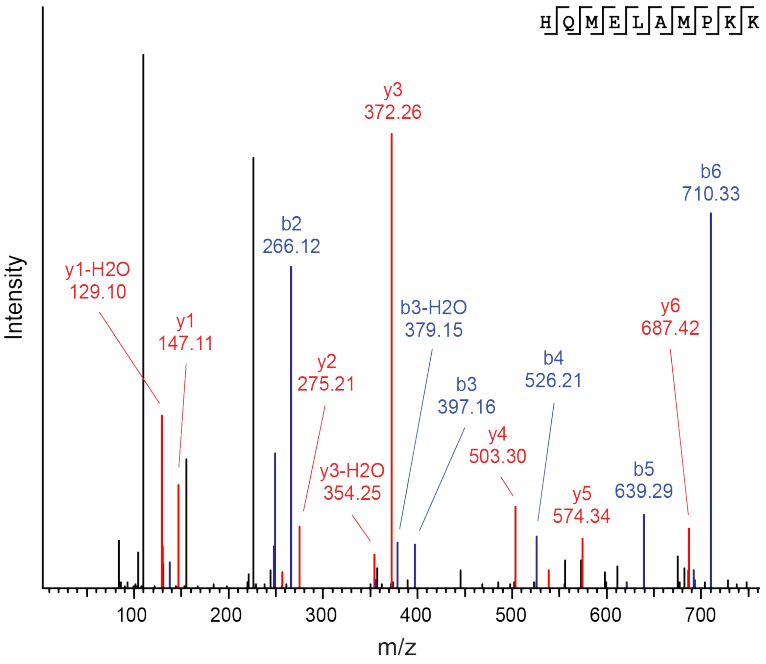
A

HQMELAMPKK - aeTSA

Endogenous peptide



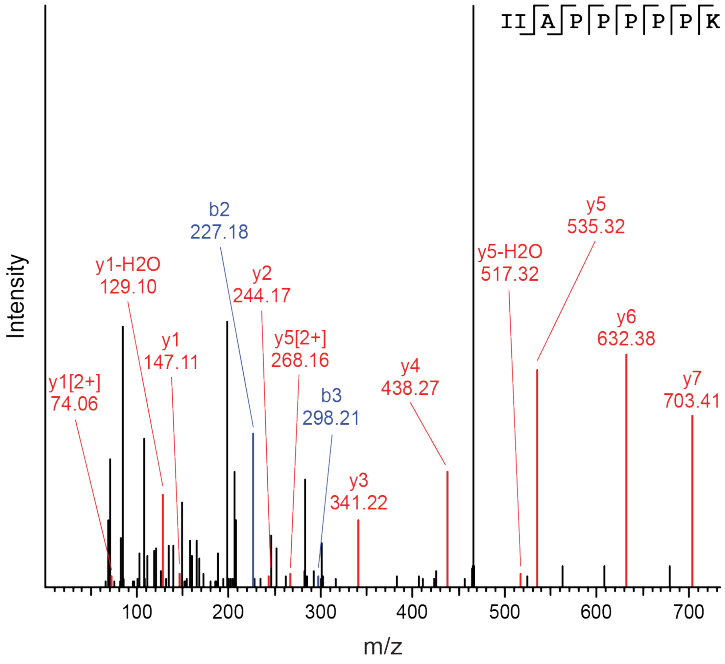
Synthetic peptide



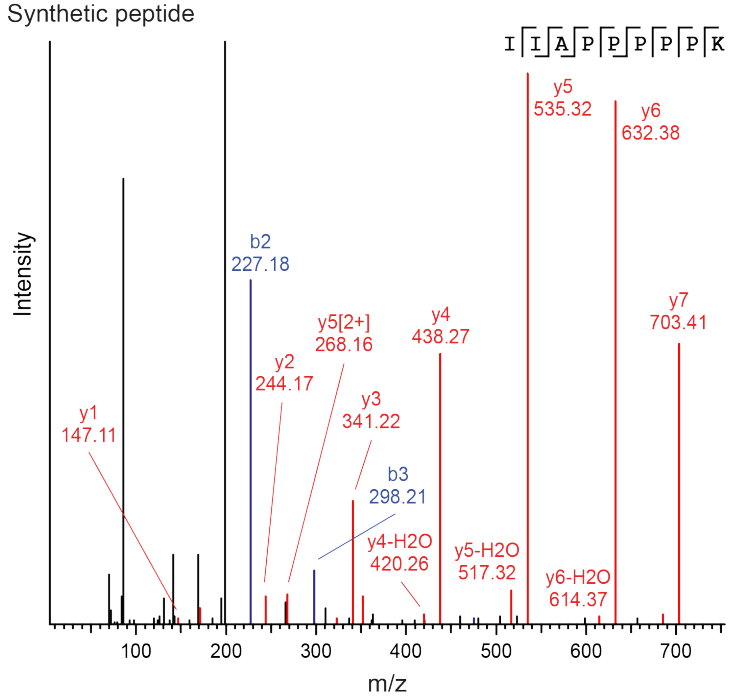
B

IIAPPPPPK - aeTSA

Endogenous peptide



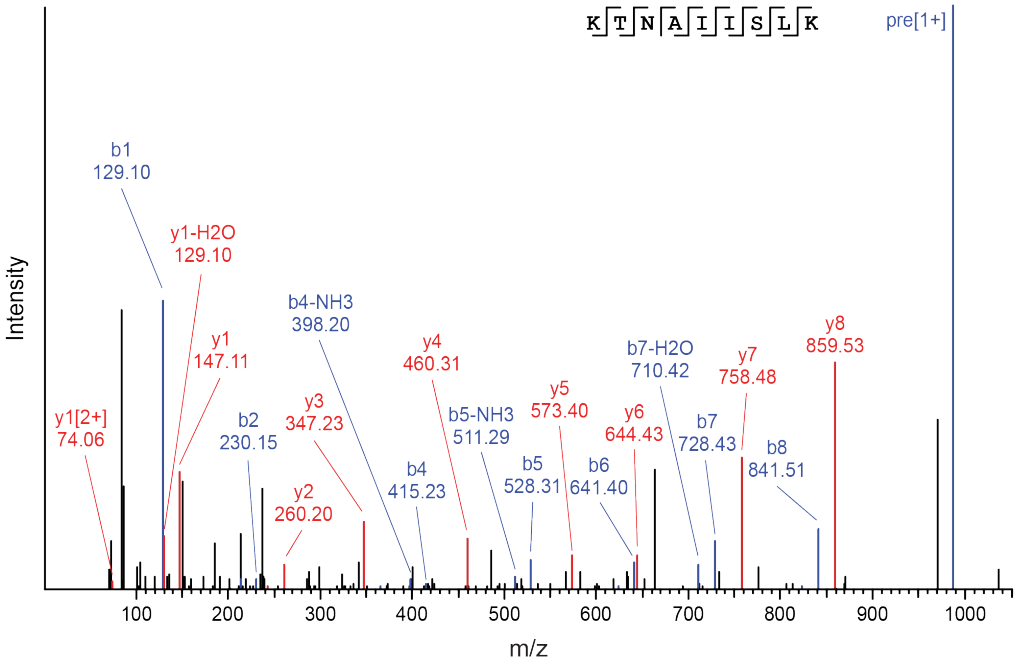
Synthetic peptide



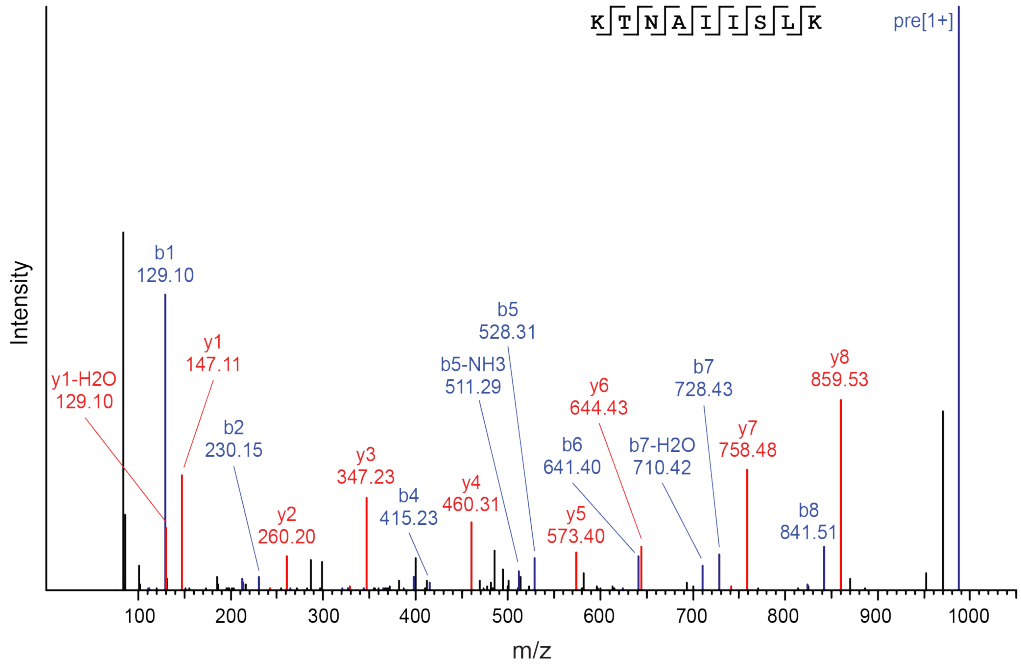
C

KTNAIISLK - aeTSA

Endogenous peptide



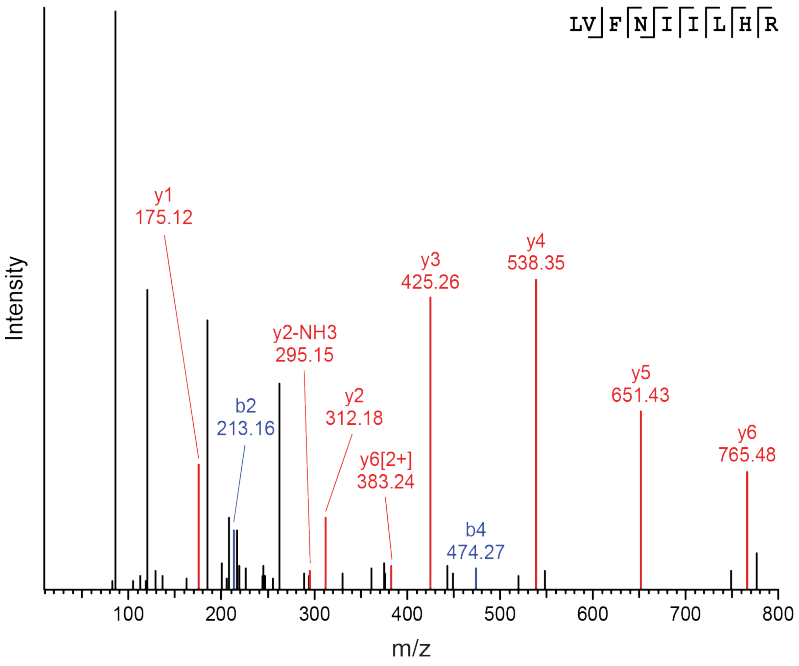
Synthetic peptide



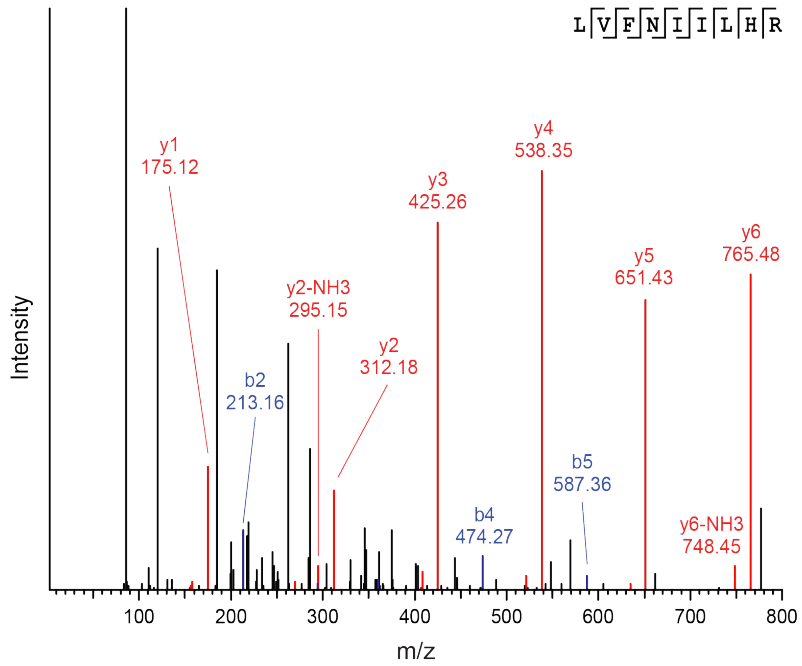
D

LVFNIILHR - aeTSA

Endogenous peptide



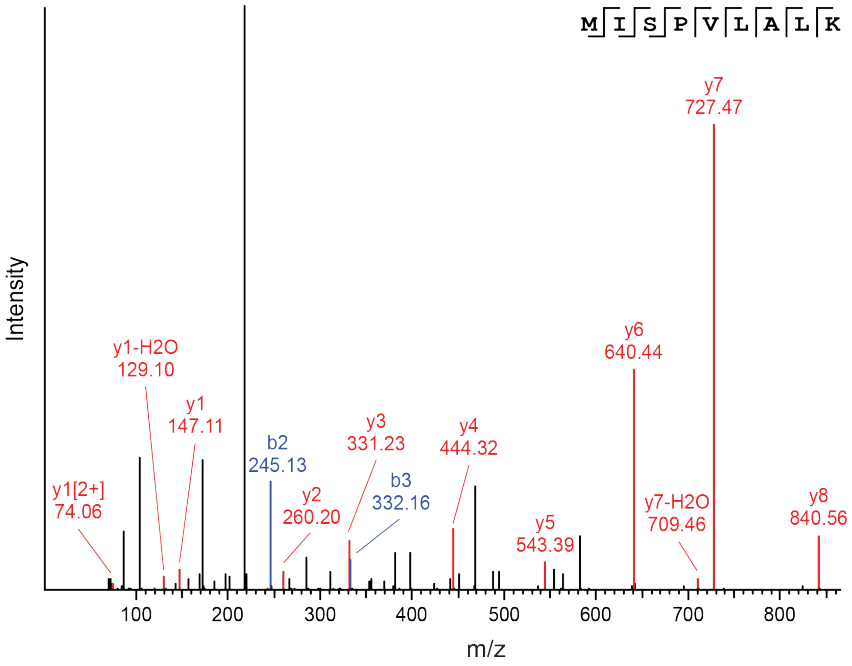
Synthetic peptide



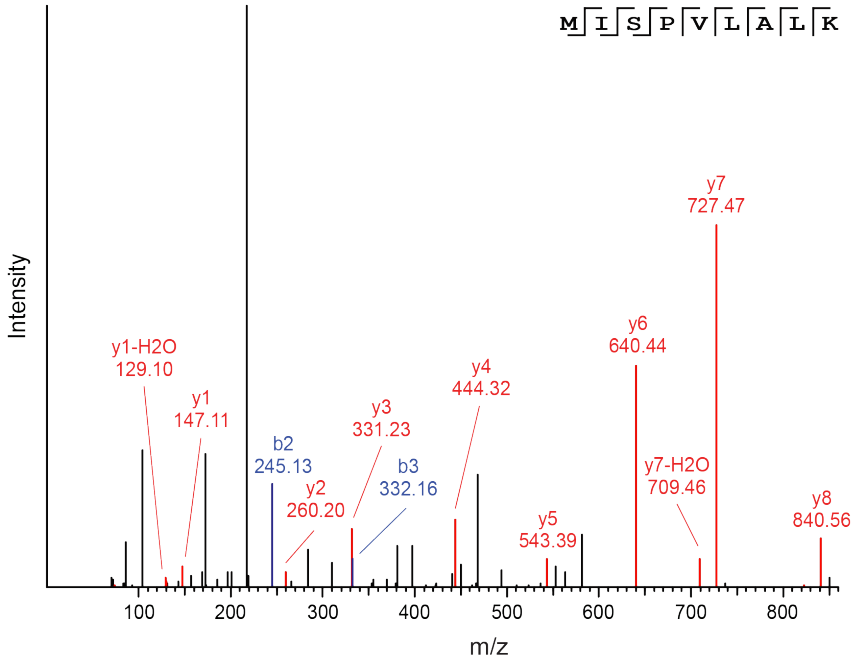
E

MISPVLALK - aeTSA

Endogenous peptide



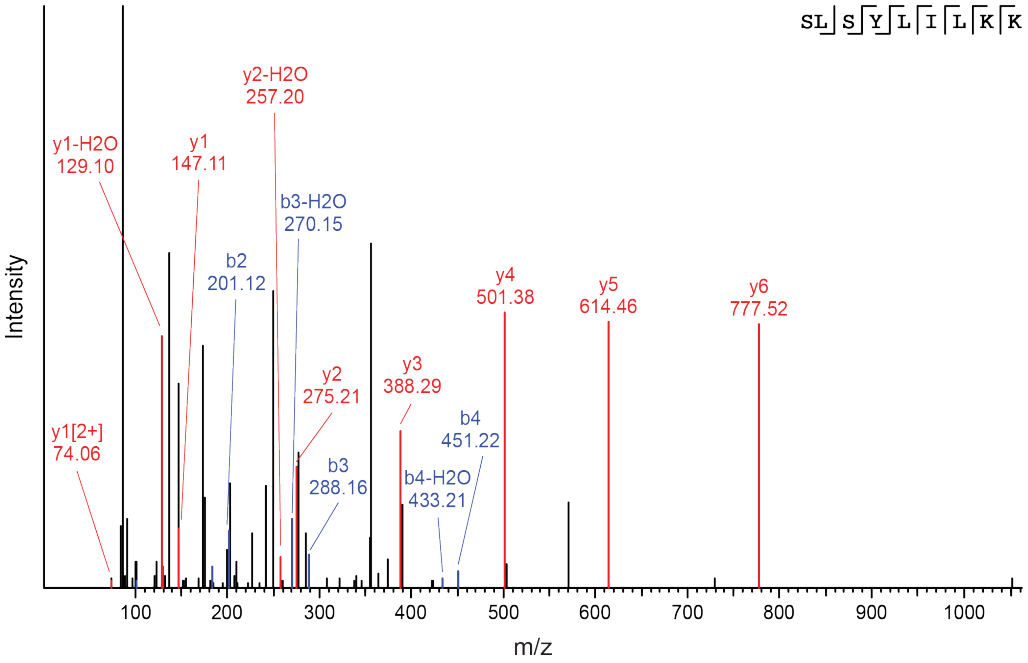
Synthetic peptide



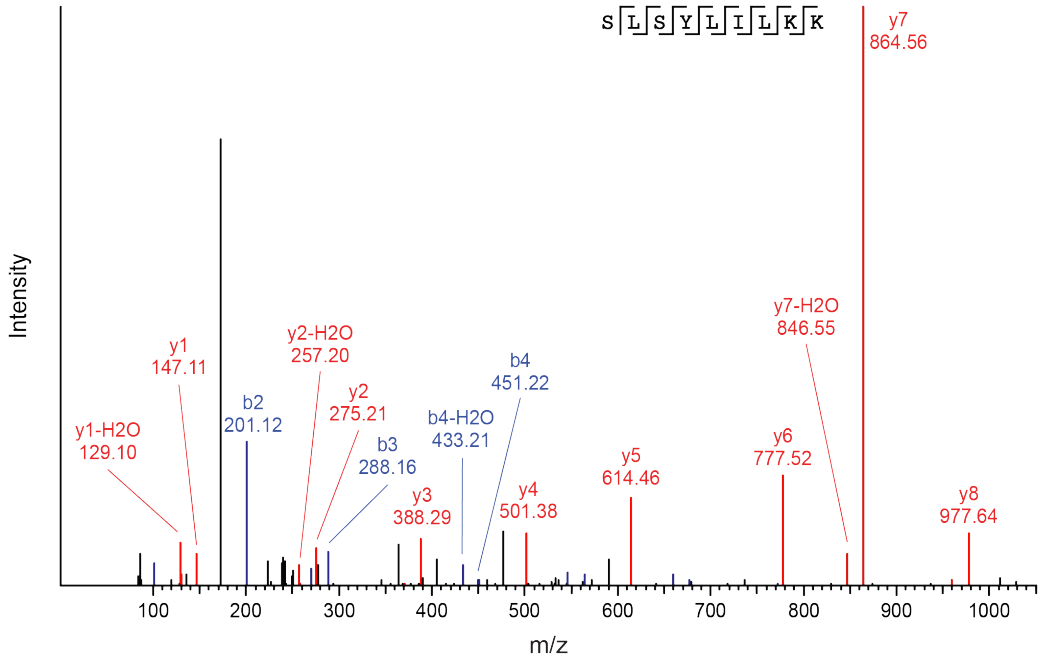
F

SLSYLILKK - aeTSA

Endogenous peptide



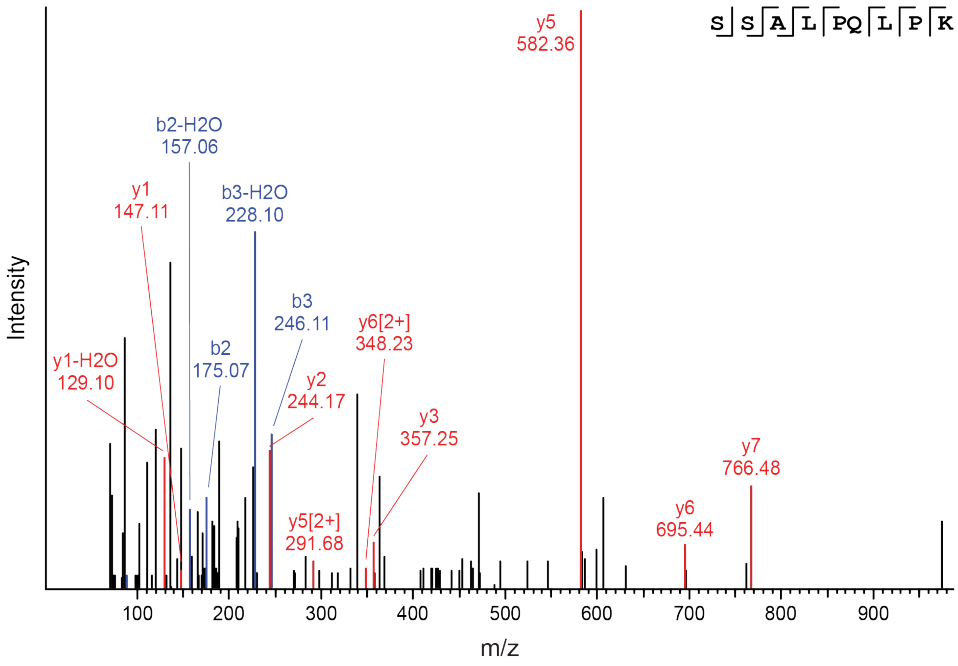
Synthetic peptide



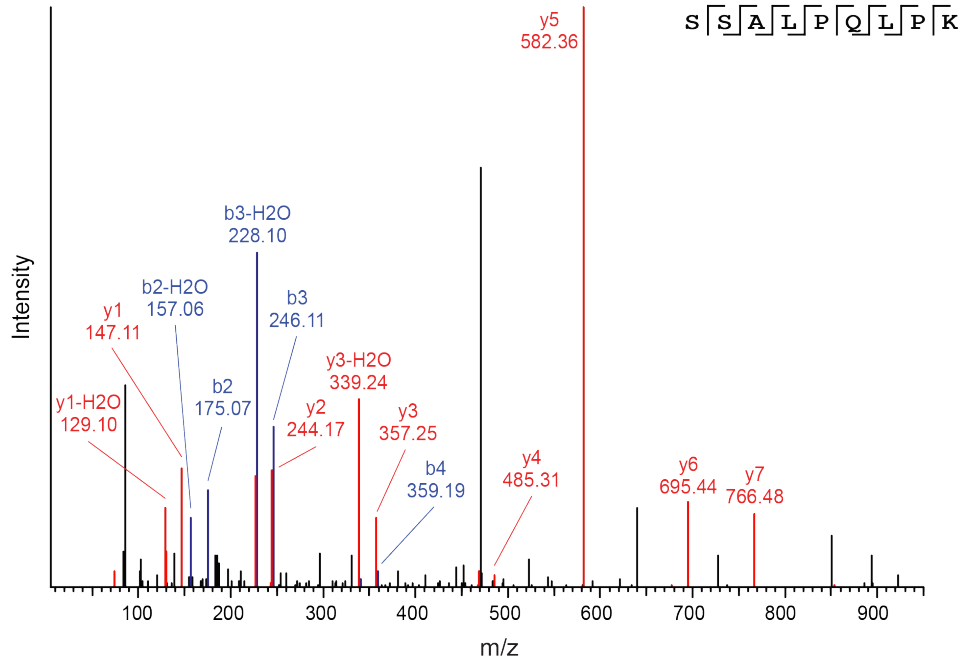
G

SSALPQLPK - aeTSA

Endogenous peptide



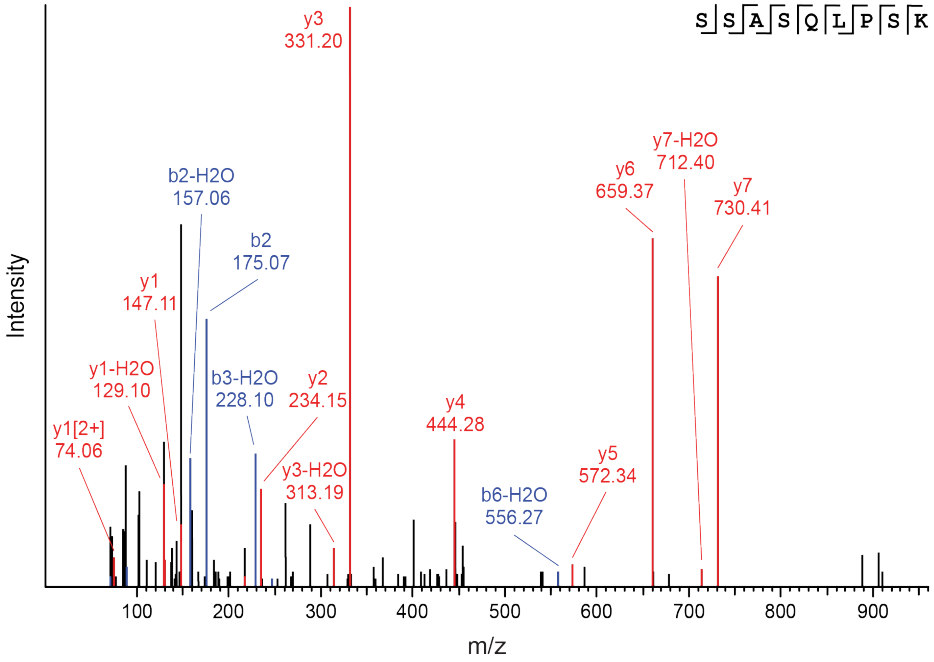
Synthetic peptide



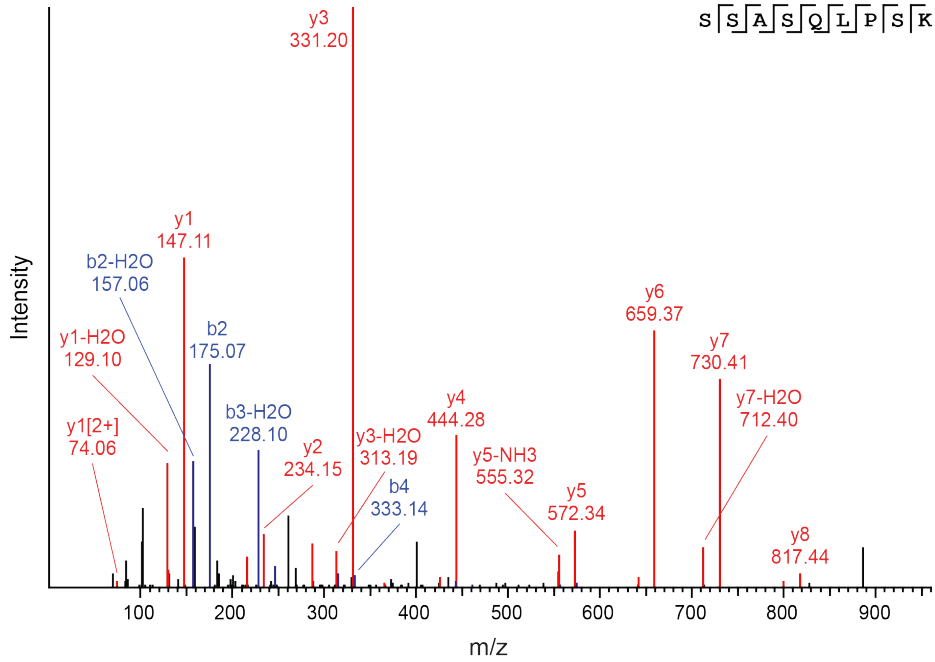
H

SSASQLPSK - ERE aeTSA

Endogenous peptide



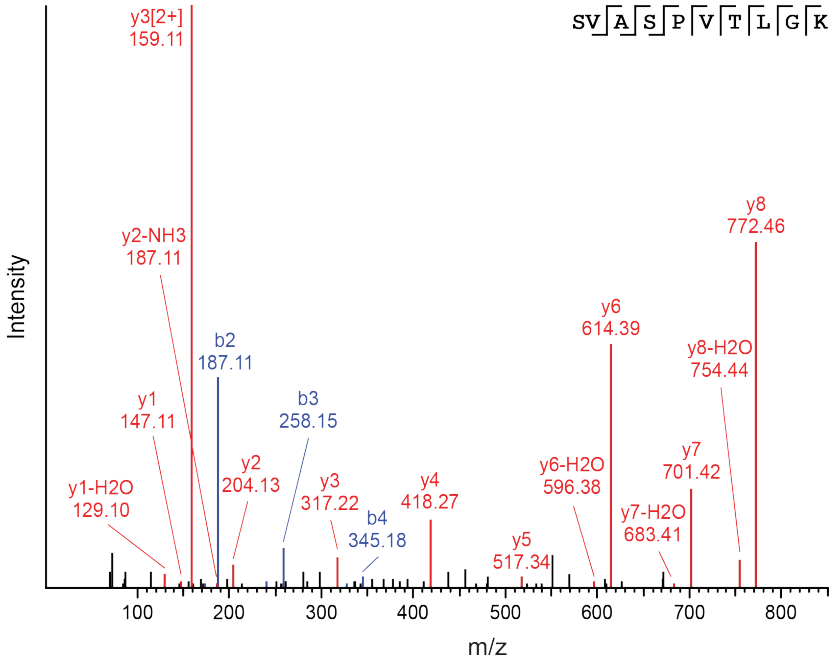
Synthetic peptide



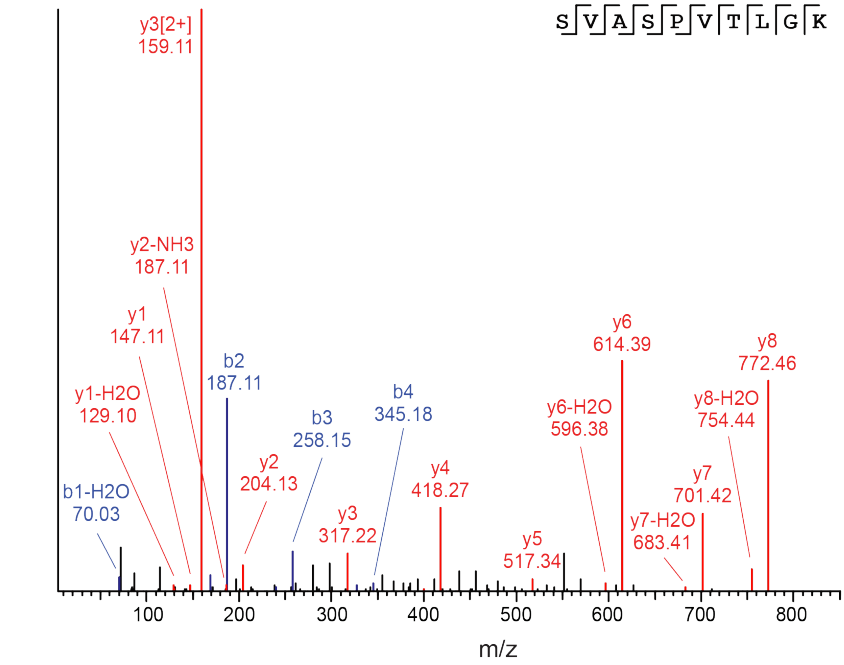
I

SVASPVTLGK - aeTSA

Endogenous peptide



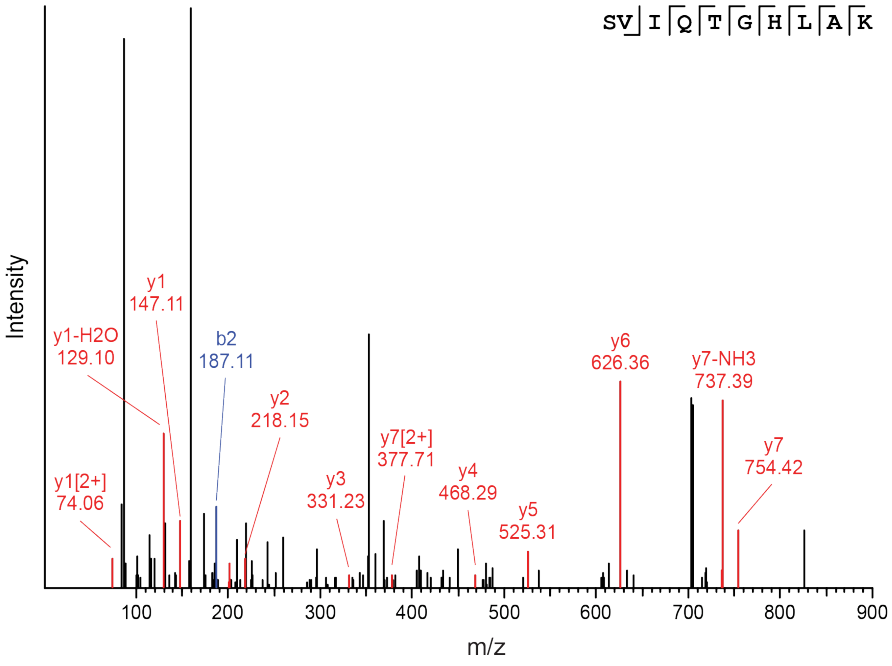
Synthetic peptide



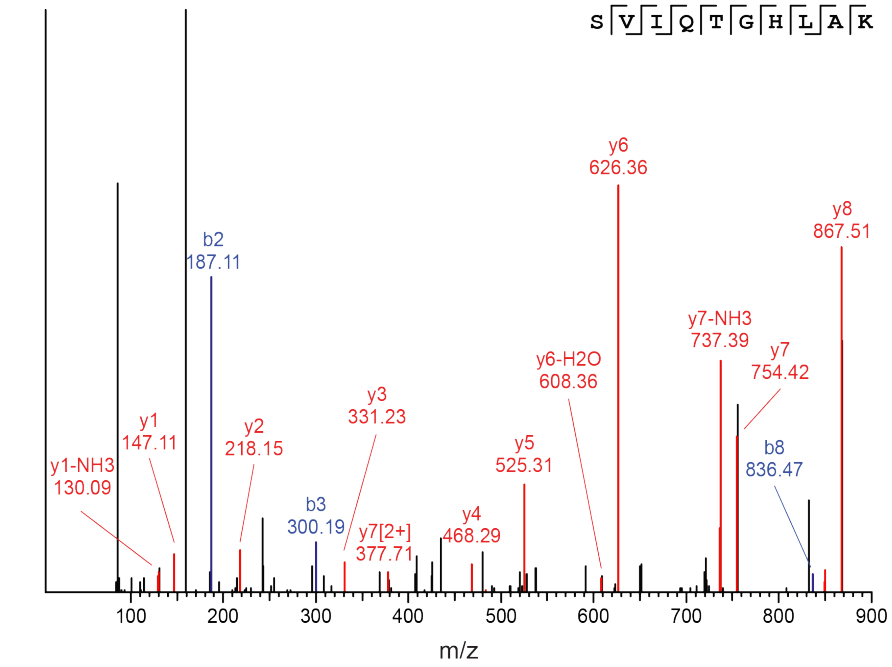
J

SVIQTGHLAK - aeTSA

Endogenous peptide



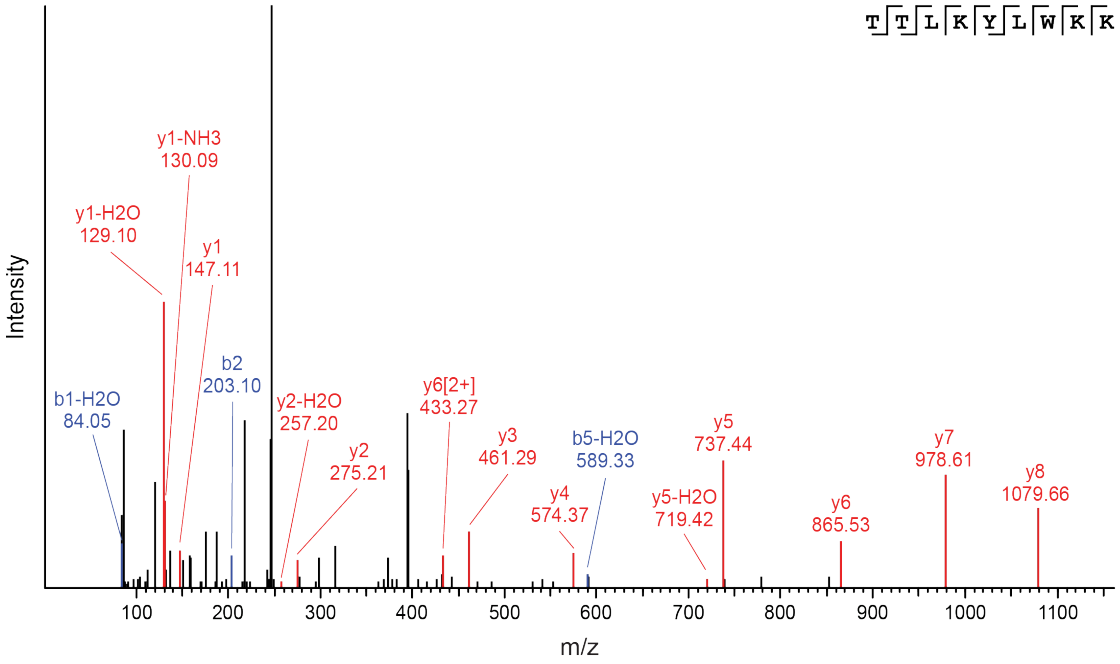
Synthetic peptide



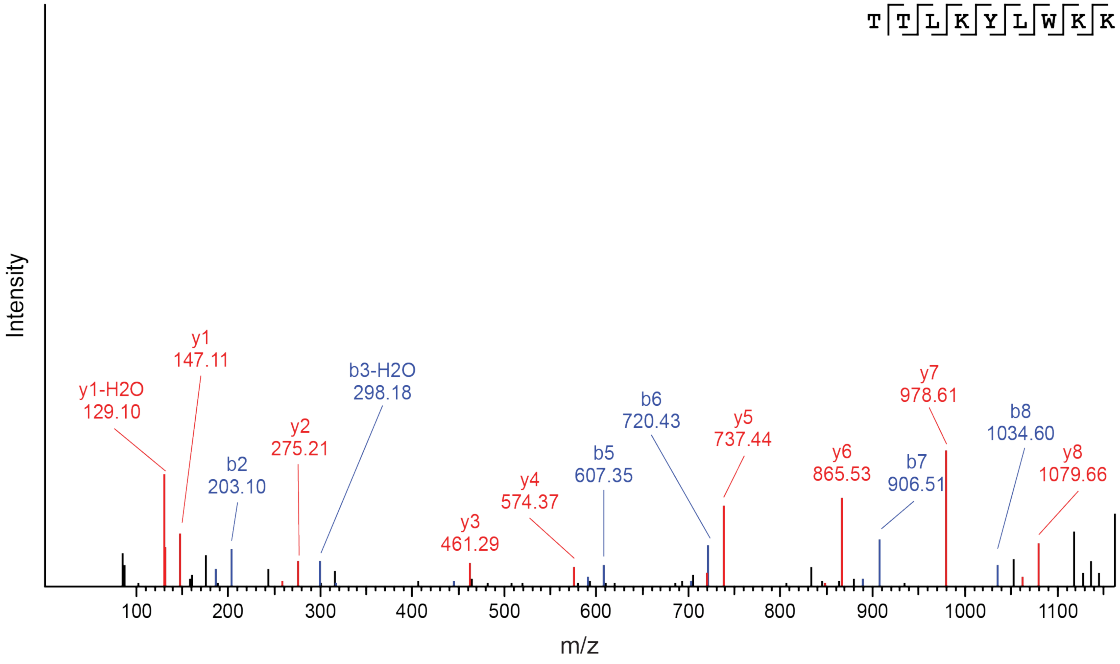
K

TTLKYLWKK - aeTSA

Endogenous peptide



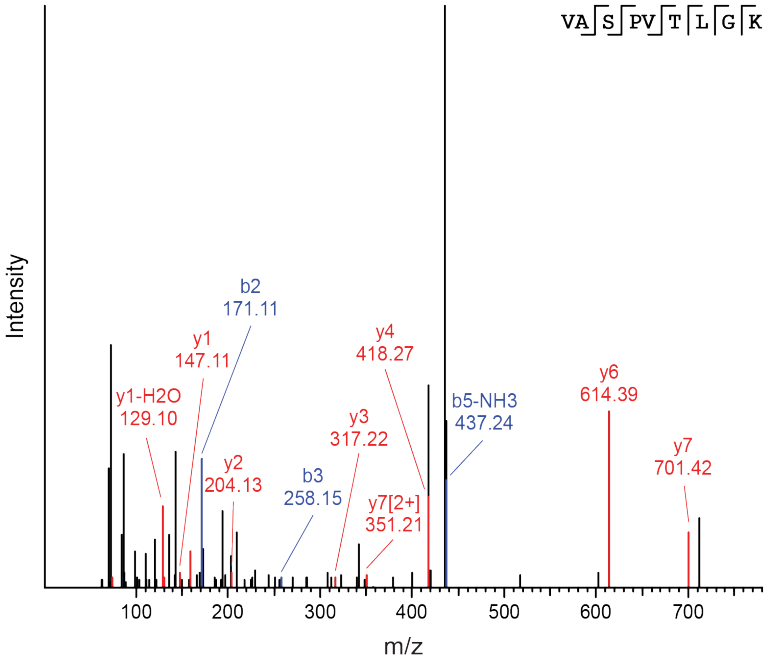
Synthetic peptide



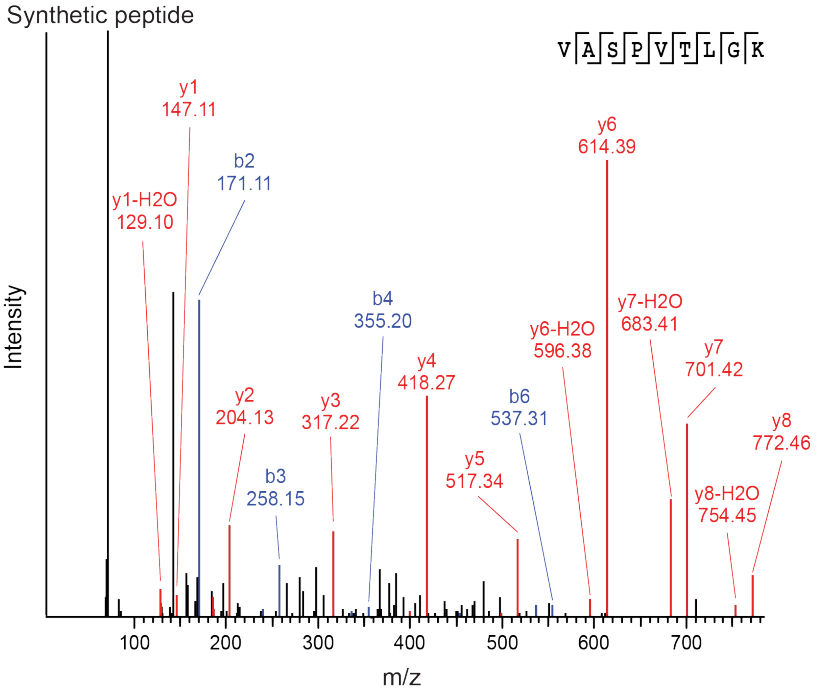
L

VASPVTLGK - aeTSA

Endogenous peptide



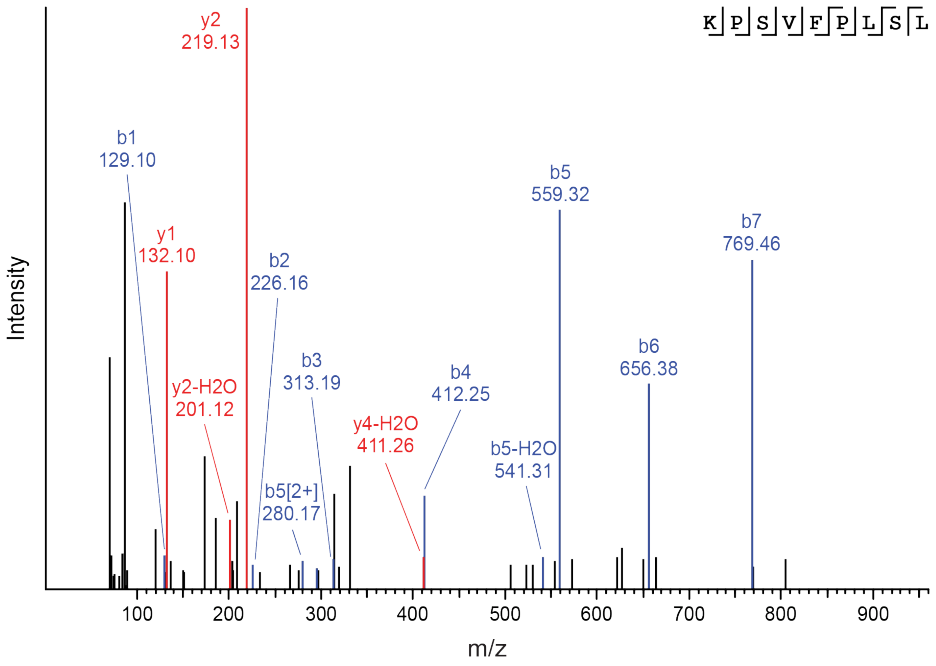
Synthetic peptide



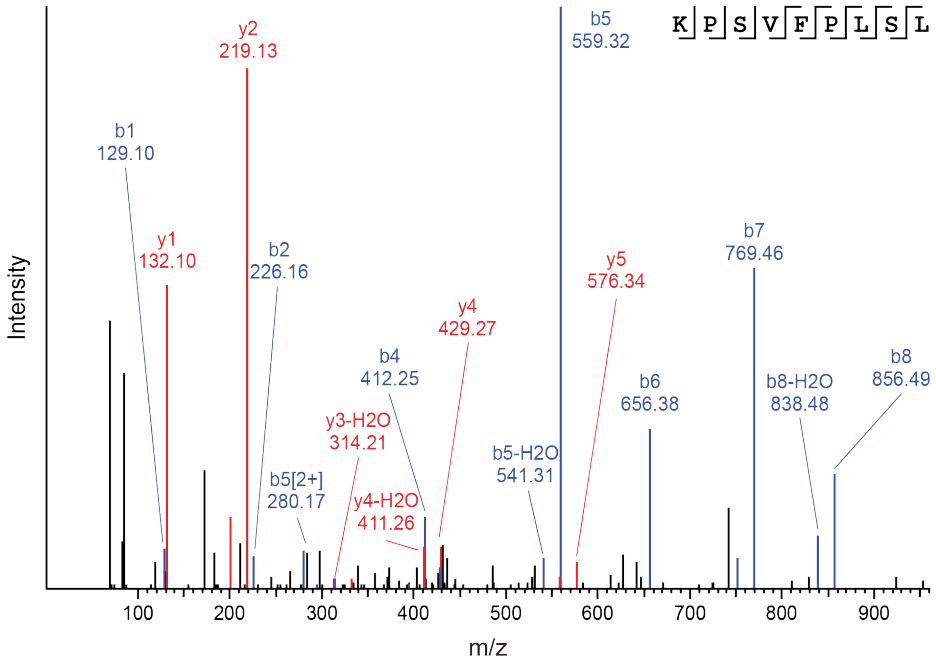
M

KPSVFPLSL - aeTSA

Endogenous peptide



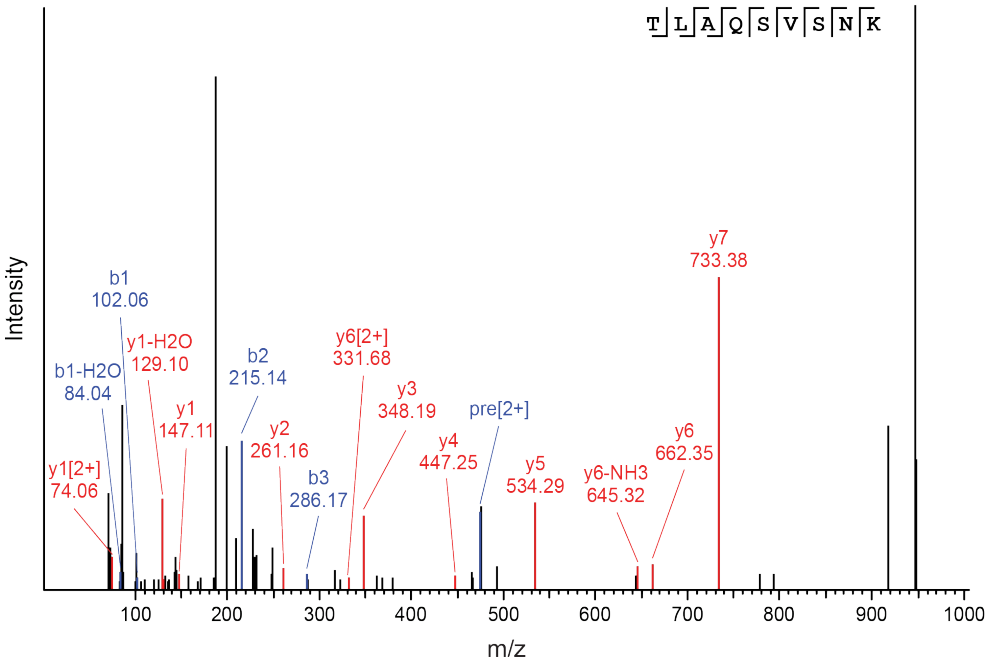
Synthetic peptide



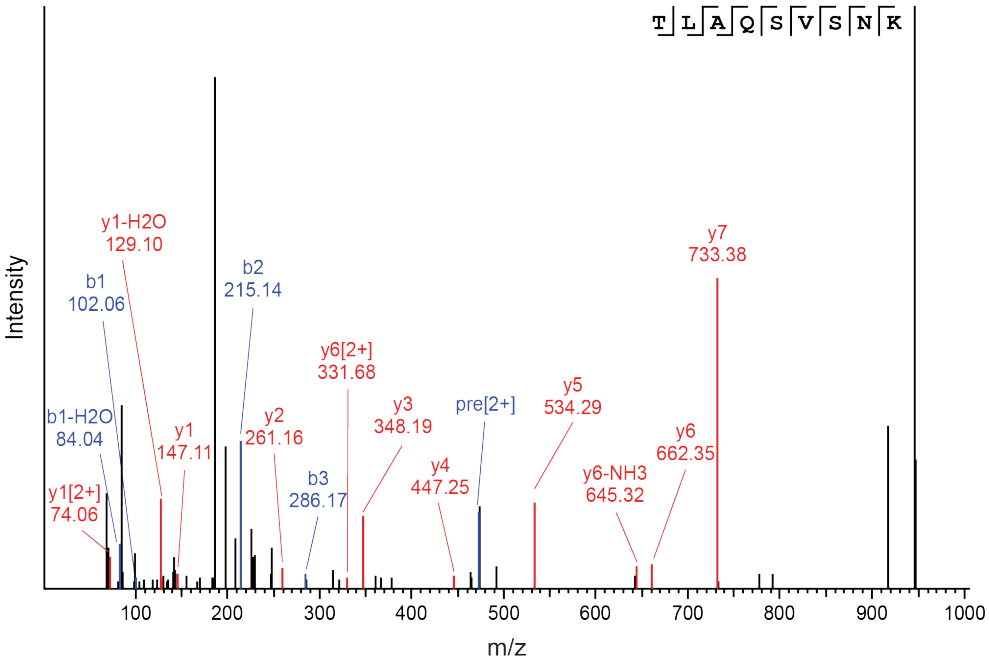
N

TLAQSVSNK - aeTSA

Endogenous peptide



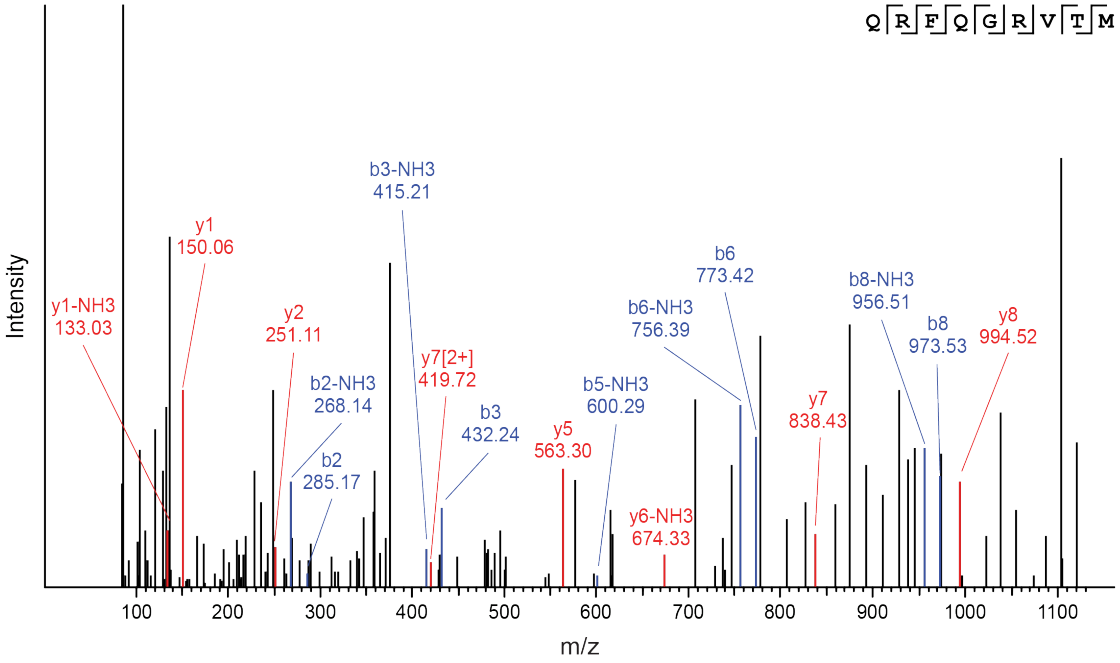
Synthetic peptide



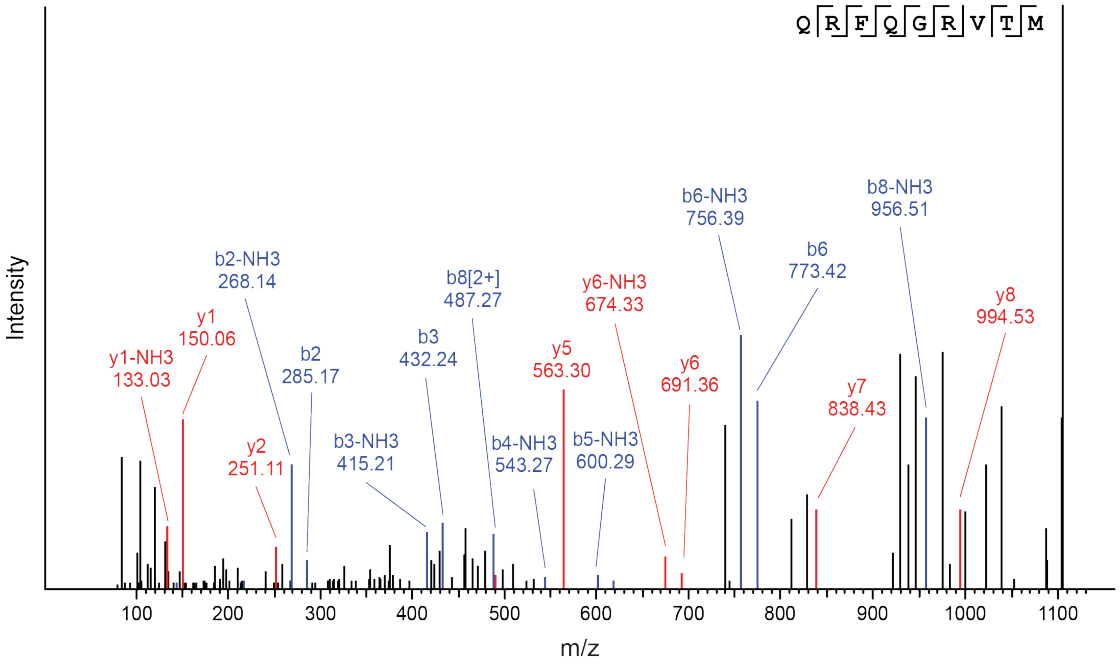
O

QRFQGRVTM - mTSA

Endogenous peptide



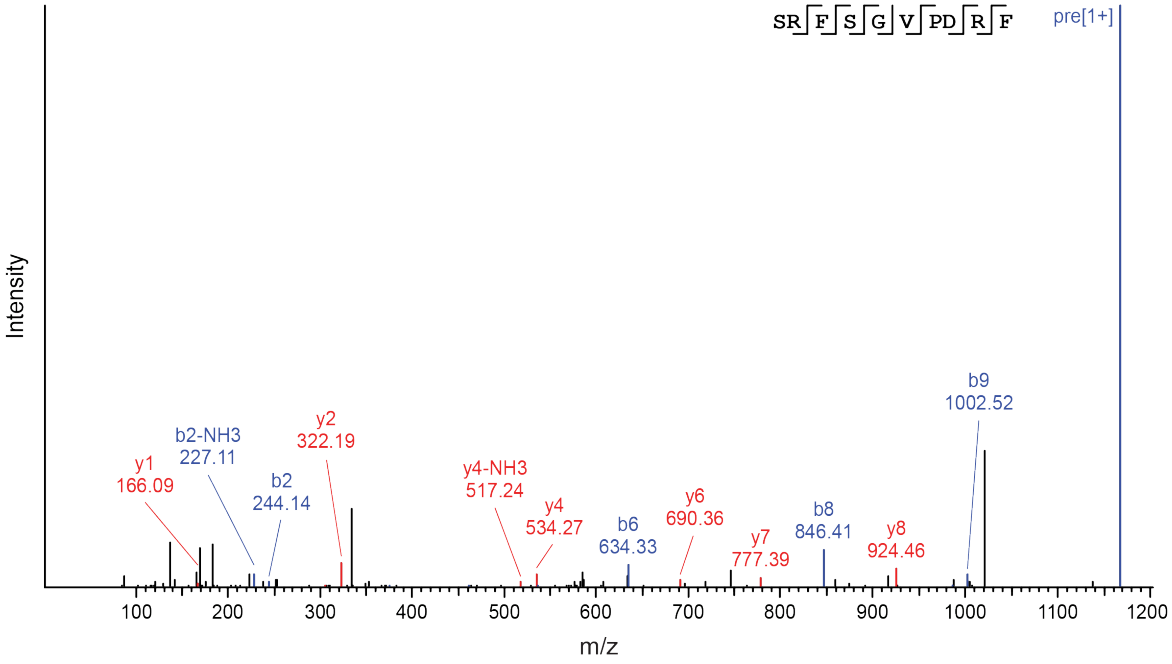
Synthetic peptide



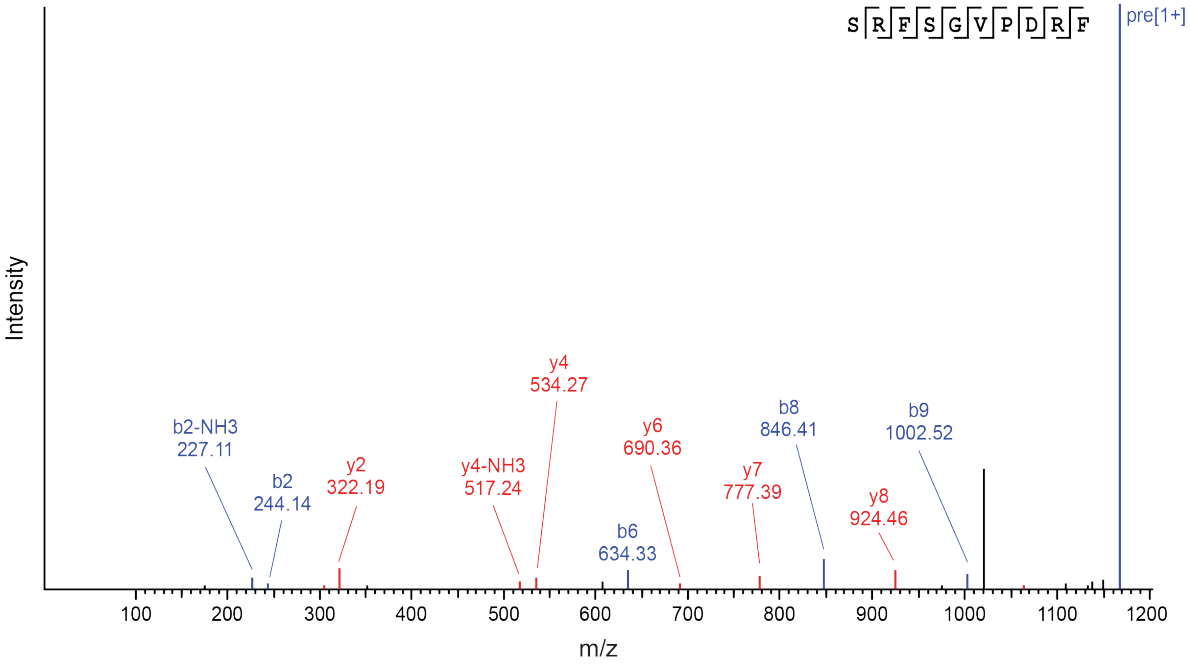
P

SRFSGVPDRF - aeTSA

Endogenous peptide



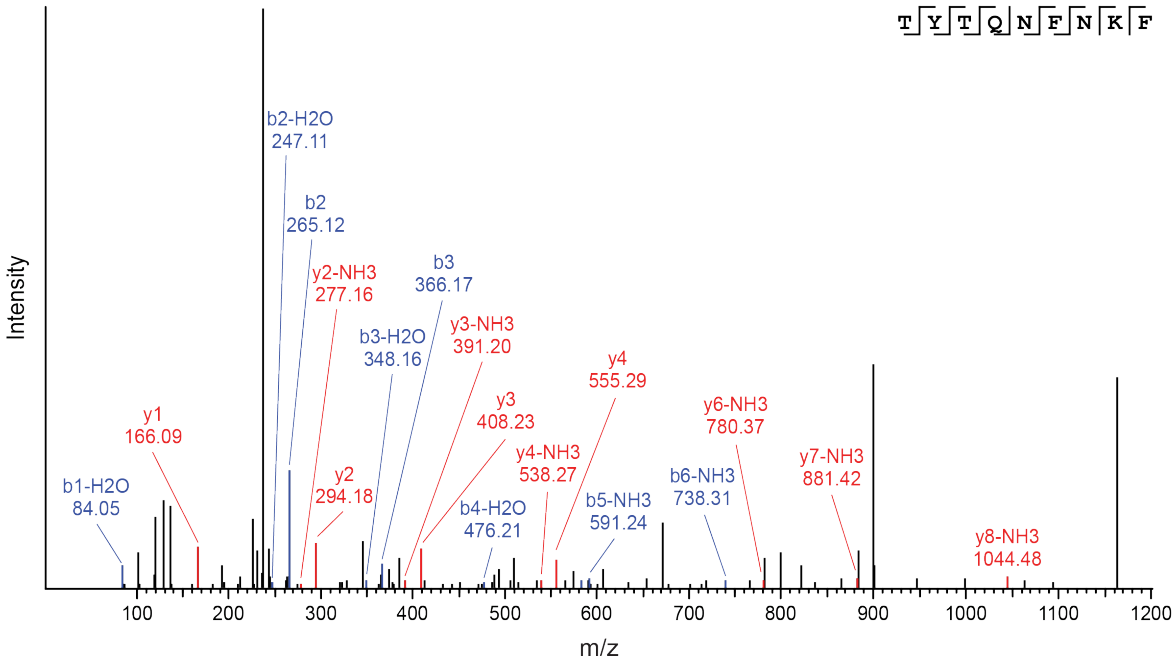
Synthetic peptide



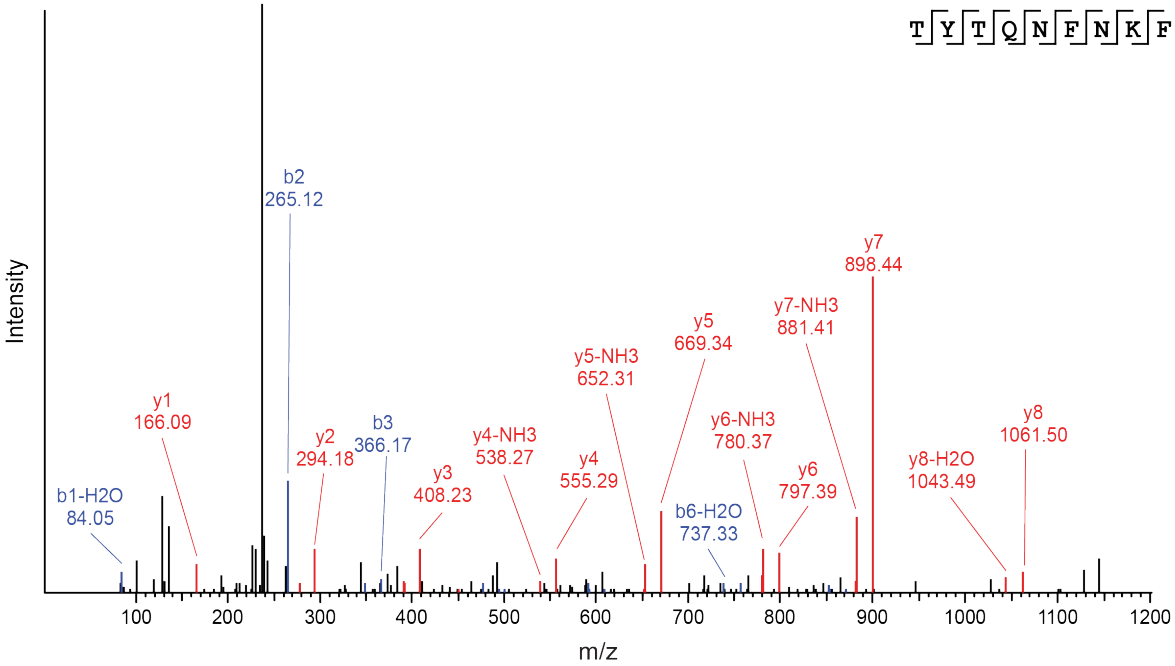
Q

TYTQNFNKF- mTSA

Endogenous peptide



Synthetic peptide



Supplementary Figure 4.11 | MS validation of lung cancer TSA candidates using synthetic analogs. Synthetic and endogenous MS/MS spectra for TSA candidates identified in each of our three lung cancers: (**A-L**) lc2, (**M-N**) lc4 and (**O-Q**) lc6. See **section 4.8.17** for details.

4.10.2 Supplementary Tables

All Supplementary Tables (i.e., **Table S1** to **S18**) are available online in '.xlsx' format (www.sciencetranslationalmedicine.org/cgi/content/full/10/470/eaau5516/DC1) and their full list is provided below:

- Supplementary Table 1. Statistics related to the generation of the global cancer databases.
- Supplementary Table 2. Information about samples used in this study.
- Supplementary Table 3. List of CT26 MHC class I–associated peptides.
- Supplementary Table 4. List of EL4 MHC class I–associated peptides.
- Supplementary Table 5. Accession numbers of the ENCODE datasets used in this study.
- Supplementary Table 6. Features of murine TSAs.
- Supplementary Table 7. Experimental values obtained in analyses of mouse TSA immunogenicity.
- Supplementary Table 8. List of 07H103 MHC class I–associated peptides.
- Supplementary Table 9. List of 10H080 MHC class I–associated peptides obtained by mild acid elution.
- Supplementary Table 10. List of 10H080 MHC class I–associated peptides obtained by immunoprecipitation.
- Supplementary Table 11. List of 10H118 MHC class I–associated peptides.
- Supplementary Table 12. List of 12H018 MHC class I–associated peptides.
- Supplementary Table 13. List of lc2 MHC class I–associated peptides.
- Supplementary Table 14. List of lc4 MHC class I–associated peptides.
- Supplementary Table 15. List of lc6 MHC class I–associated peptides.
- Supplementary Table 16. Accession numbers of the Genotype-Tissue Expression (GTEx) datasets used in this study.
- Supplementary Table 17. Features of human TSAs detected in B-ALL specimens.
- Supplementary Table 18. Features of human TSAs detected in lung tumor biopsies.

CHAPTER 5

5 Discussion

The overarching goal of this thesis was to provide the first-ever systems-level analysis of the cryptic MAP repertoire in normal and cancer cells. Because no MS-based approach were readily available to identify cryptic MAPs, i.e., MAPs derived from non-canonical translation events, we assembled our own proteogenomic toolbox by exploring novel ways of analyzing RNA-Seq data to build more informative databases for MS. Our first DB-building strategy, relying on the 6-frame translation of unmapped RNA-Seq reads, allowed us to characterize the overall MAP repertoire presented by human B-LCLs (**Chapter 2**). Throughout this study, we demonstrated that cryptic MAPs represent a sizeable portion of the MAP repertoire (168–342 out of 2,603 MAPs ~ 10%) and that they derive from (i) the out-of-frame translation of protein-coding transcripts as well as (ii) the translation of non-coding sequences (e.g., introns or UTRs) and transcripts (e.g., pseudogenes). Despite their unexpected origin, cryptic MAPs likely derive from genuine translational events rather than noise, as most of them can be reproducibly detected in subjects sharing the relevant MHC I molecules. Interestingly, their biogenesis appears to be quite peculiar, as it seems to involve the proteasome-independent processing of very short proteins produced by the translation of highly expressed yet unstable RNAs. Lastly, we observed that, in normal cells, cryptic MAPs originate from polymorphic genomic regions. This observation lead us to hypothesize that these regions might easily accumulate somatic mutations in cancer cells and therefore represent an unprecedented source of TSAs. To further investigate this idea, we developed a refined DB-building strategy and characterized the cryptic and conventional TSA landscape of nine tumor samples, including two murine tumor cell lines (EL4 and CT26) and seven human primary samples (four B-ALL specimens and three lung tumor biopsies, **Chapter 4**). With this analysis, we were the first to show that cryptic TSAs drastically outnumber conventional TSAs, as they represent 36 TSAs out of the 40 we identified. Unexpectedly, most of these TSAs derive from cancer-specific yet unmutated sequences (aeTSAs, 32 out of 40 TSAs), including EREs (13 out of 40 TSAs), thereby suggesting that we identified a shareable pool of highly immunogenic TSAs. To evaluate the *in vivo* protective effect of our TSAs, we selected five EL4 TSAs

that had no peripheral expression and could therefore be considered truly EL4 specific. Interestingly, vaccination against single TSA prior to EL4 challenge improved the survival of 10 to 100% of the mice, our best candidate being an ERE aeTSA, that is an unmutated antigen. By comparing properties of the EL4 TSAs used in this prophylactic vaccination study, we concluded that (at least) two crucial factors do influence the efficacy of TSA vaccination: TSA expression in cancer cells and the frequency of TSA-responsive T cells in the pre-immune repertoire. Knowing that estimation of such factors can be done in humans, these could be used to prioritize TSAs in clinical studies.

The work presented in this thesis contributed to deepen our knowledge of the MAP repertoire by showing that cryptic MAPs can significantly enlarge the scope of CD8+ T cell immunosurveillance. Besides providing the research community with two proteogenomic frameworks for the identification of such cryptic MAPs, we also demonstrated that cryptic TSAs represent promising targets for T-cell based cancer immunotherapy. Indeed, they are far more frequent than conventional TSAs and combine a strong immunogenic potential with the possibility of being shared between patients, most of them deriving from unmutated sequences. Nonetheless, far more work is needed to consolidate and expand our findings, from a technical, functional and even conceptual standpoint. Thus, in the following sections, we will highlight the strengths and weaknesses of our proteogenomic approaches compared to current strategies used in the field. Next, we will discuss improvements that could be made to our two DB-building strategies for the analyses of the MAP repertoire by proteogenomics. Then, we will move on to discuss more functional implications of our results by reviewing and expanding on what we have learned on (i) the origin of cryptic MAPs, (ii) their peculiar biogenesis and (iii) their representation in the MAP repertoire of tumor cells. Lastly, we will briefly discuss the revolution that cryptic TSAs could represent for T-cell based immunotherapy, as well as the challenges they bring along.

5.1 Us and them

Previous proteogenomic approaches developed by systems immunologists have used sample-specific databases resembling our *control database* presented in **Chapter 2** (**Figure 2.1a,** page 82) or our *canonical cancer proteome* presented in **Chapter 4** (**Figure 4.1a,** page 186) to identify MAPs[1-3]. Briefly, the mutational and transcript expression landscape of each sample are inferred from DNA (exome) and RNA-Seq data, respectively. These two pieces of information are then combined with the reference genome to build a sample-specific transcriptome that is further *in silico* translated to obtain a sample-specific proteome, which, once submitted to a DB search engine (such as Mascot or Peaks) along with the relevant collection of MS/MS spectra, allows for MAP identification. Though easy to implement, this DB-building strategy presents two main drawbacks: first, the identification of sample-specific mutations by any mutation callers implies mapping DNA or RNA-Seq reads to the reference genome. However, using such software tools to manipulate sequencing reads necessarily imply loosing data. Indeed, current mappers have a hard time mapping reads coming from repetitive sequences such as EREs[4], while current mutation callers only identify simple mutations from mapped reads, i.e., single-base mutations or small insertions and deletions. Moreover, the generation of sample-specific transcriptomes and proteomes relies on current genome annotations provided by consortium like ENCODE[5]. Consequently, such approaches only explored a narrow part of the MAP landscape, that is MAPs derived from known protein-coding transcripts bearing (or not) simple mutations.

To avoid identifying MAPs derived *only* from known transcripts, we developed two DB-building strategies that one could define as more flexible (**Figure 2.1a**, page 82 and **Figure 4.1a**, page 186). Although these two strategies differ through some of their properties, they both rely on the same core principle, that is, the use of unaligned RNA-Seq reads. In the first (and more comprehensive) strategy, we used the 6-frame translation of all RNA-Seq reads to build sample-specific DBs, while in the second (and more targeted) strategy, we only used the translation of cancer-specific RNA-Seq reads. Bypassing the mapping and mutation calling steps allowed our two alignment-

free strategies to leverage the full information contained in each RNA-Seq dataset. Consequently, these approaches were (and still are) the sole allowing for the identification of conventional and cryptic MAPs as well as mutated MAPs derived from either single-base mutations or more complex structural variants. With our targeted strategy we even went further by proposing a solution for the high-throughput identification of TSAs derived from normal yet cancer-specific sequences, that is, aeTSAs. Here, rather than combining fancy (and heuristic) approaches to identify the few cancer-specific sequences hidden in a sea of normal sequences, we used a naive (and precise) approach, termed k-mer profiling, to identify what was common between the transcriptomic landscape of normal and cancer cells. As our ultimate goal was to identify TSAs, i.e., peptides derived from the translation of non-tolerogenic and cancer-specific sequences, we decided to use the transcriptome of TECs and mTECs as normal control, these cells being essential for the establishment of central tolerance. Indeed, we reasoned that removing sequences detected in both TECs/mTECs and cancer cells from the pool of sequences to be translated would enrich our DB with mutated and unmutated sequences likely to be non-tolerogenic. Using this simple subtraction, we discovered that aeTSAs can be reliably identified by MS and that they outnumber mTSAs.

Despite all those advantages, it is important to understand that our comprehensive and targeted DB-building strategies are not some sort of super-combined approaches able to identify all MAPs. They should rather be seen as works in progress that could always be improved. First, *our two strategies likely underestimate the proportion of cryptic MAPs and TSAs in the MAP repertoire*. Indeed, as we decided to use poly(A) RNA-Seq reads, we are likely depleting our DBs from some non-coding RNAs that are not polyadenylated but could nonetheless generate cryptic MAPs. Thus, a better strategy to capture and therefore consider those non-polyadenylated transcripts would be to build our DBs for MS using an RNA extraction protocol involving rRNA depletion rather than poly(A) enrichment. Second, *our current workflows cannot derive reliable quantitative data on all identified MAPs*. Indeed, for each project, we decided to use a label-free shotgun MS approach, which does not require any *a priori*

knowledge of the analyzed MAPs and is therefore ideal to discover new MAPs. Despite its flexibility, this approach suffers from a poor reproducibility, implying that we are likely underestimating the extent of cryptic MAPs and TSAs sharing across samples. Consequently, it could be interesting to complement our shotgun MS approach with a more targeted MS approach, like PRM-MS, to derive reliable quantitative data on a pre-selected set of cryptic MAPs and TSAs. Third, *our strategies can only identify unmodified MAPs encoded by contiguous genomic sequences*. Indeed, as we did not perform enrichment for peptides bearing specific post-translational modifications such as phosphorylation, we are likely to miss most of them, though such epitopes have been shown to trigger anti-tumor responses[6,7]. Moreover, as our DBs were built from RNA-Seq reads, we can only identify MAPs derive from linear genomic sequences as well as sequences created by the splicing of pre-mRNAs. However, we do know that MAPs can be created by a process called proteasome splicing during which the proteasome fuses two distant protein segments (belonging, or not, to the same protein) thereby creating a so-called proteasome-spliced MAP. Since such splicing events occurs concomitantly to protein degradation, that is after transcription, these proteasome-spliced MAPs cannot be included in any of our RNA-Seq-based DBs and are therefore missed by our two approaches. Recently, these spliced MAPs have been shown to represent ~ 25 % of the MAP repertoire[8] though their importance is still questioned[9].

Altogether, it is important to understand that all approaches studying the MAP are equally valid but that they come with different blind spots. For instance, we decided to focus on the identifications of MAPs derived from non-canonical translation events as well as structural variants, while other research groups decided to ignore such MAPs and focus on the identification of other types of MAPs or more quantitative aspects. Thus, for the time being, getting a comprehensive picture of the MAP repertoire of any given sample does require to combine results obtained through those different approaches, including ours.

5.2 Thoughts to improve our DB-building strategies

Although our two DB-building strategies are optimal for the identification of cryptic MAPs, it is very important to understand that our *comprehensive strategy* presented in **Figure 2.1a** (page 82) and our *targeted strategy* presented in **Figure 4.1a** (page 186) were designed with two different objectives in mind and would not necessarily identify the same MAPs, even if applied to the same sample. Indeed, the first strategy tried to get a global picture of the overall MAP repertoire (conventional and cryptic), while the second one focused on identifying MAPs that could be of relevance to cancer immunotherapy, i.e., aeTSAs and mTSAs. As such, both strategies are worth improving for future studies of the MAP repertoire.

5.2.1 When 'less is more'!

In our comprehensive DB-building strategy, we simply replaced the reference proteome by the *in silico* 6-frame translation of all unmapped RNA-Seq reads. This is clearly the most comprehensive DB that one can build as it takes into account (i) the sample's mutational profile, encoded in RNA-Seq reads, as well as (i) any translational events occurring in the cell, thanks to the *in silico* 6-frame translation. As a corollary, the resulting DBs are likely to include a lot of irrelevant protein sequences (~ 80%) and therefore be extremely large. Accordingly, we observed that the final size of such all-frames databases could be 30 to a 100 times larger than the reference proteome and that this was directly influenced by the RNA-Seq depth (data not shown). As comprehensive as they might be, such large databases are poorly handled by current MS software tools because of the target-decoy strategy they use to estimate the FDR of their peptide-spectrum-matching procedure (**Figure 1.8**, page 40). Indeed, with this approach, the larger the database, the higher the chances of having high-scoring, yet incorrect, peptide-spectrum matches[10-12], thereby leading to a gross FDR over-estimation and consequently yielding less identifications than smaller, yet less complete, databases (at the same FDR).

Building a DB able to comprehensively characterize the MAP repertoire of cells, while limiting the impact of this 'DB size-induced identification drop', requires to trade-

off DB completeness vs. size. One (easy-to-implement) strategy to reduce the overall size of the DB, without losing meaningful information, would be to use RNA-Seq data obtained from strand-specific rather than standard libraries[13,14]. Indeed, with this information in hand, DBs could be generated by performing an *in silico* 3- rather 6-frame translation. As expected, this optimization step, which was actually implemented in our targeted DB-building strategy, drastically reduced the size of our DBs and has likely lowered the percentage of irrelevant protein sequences included down to ~ 60%. When large amount of live starting material are available, i.e., a minimum of 10 million cells[15], a more fancy optimization step would be to build the DB by integrating information obtained from stranded RNA-Seq and Ribosome profiling (Ribo-Seq) data. Although not trivial[16], the mapping and deconvolution of sequenced ribosome-protected mRNA fragments allows to determine the reading frame(s) in which mRNAs are actually translated[17]. Using such information should reduce the size of the DB to its lowest, as most translational events included are expected to be true[18]. As such, PROTEOFORMER, a software implementing such dual RNA/Ribo-Seq DB-building strategy, was shown to improve protein identification by MS[19], while an another software, called PRICE, improved MAP identification[20]. The sole caveats of this approach reside in the fact that Ribo-Seq experiments are likely to underestimate the complexity of the translatome, especially for low-complexity genomic regions (such as EREs)[18]. Thus, to avoid missing cryptic MAPs derived from EREs, it might be worth fishing out those low complexity RNA-Seq reads and include their *in silico* 3-frame translation to the RNA/Ribo-Seq-derived DB.

Altogether, implementing those solutions should (i) increase the scalability of our approach, by limiting the impact of sequencing depth on the final DB size, and (ii) allow for a better coverage of the overall MAP repertoire, including cryptic MAPs. Lastly, exploring alternative, and more accurate, ways to estimate the number of false identifications might be worth testing for proteogenomic searches[21]. The mixture model-based method, which uses an expectation maximization algorithm to discriminate between correct and incorrect identifications, might be a good option as it can bypass the need of a decoy database[22].

5.2.2 Targeted, not restrictive!

In our targeted DB-building strategy, we concatenated each canonical cancer proteome to its respective cancer-specific proteome, obtained by performing the *in silico* 3-frame translation of so-called non-tolerogenic sequences. These sequences were identified by comparing RNA-Seq data from normal (TEC/mTEC) and cancer samples with our novel alignment-free workflow called k-mer profiling. Briefly, normal k-mers were simply subtracted from cancer k-mers to identify cancer-specific k-mers that are likely to encode non-tolerogenic sequences (one might say TSA precursors), as they must overlap with either cancer-specific mutations or aberrantly expressed sequences. Up to now, this DB-building strategy has been successfully applied to the analysis of nine tumor samples and allowed for the identification of TSAs in each of them (**Figure 4.2**, page 190 and **Figure 4.6**, page 201)!

The use of k-mer profiling in our workflow present several advantages over more classical, mapping-based, DB-building strategies. First, it is an extremely *simple* and *fast* method to use, as it tries to extract and reconstruct the non-tolerogenic, rather than the entire, transcriptome[23,24]. Second, it is an *unbiased* method that can capture sequences likely to encode mTSAs and aeTSAs. Up to now, the identification of mTSAs by MS has been focused on single-base mutations, because complex rearrangements are often poorly captured by current mappers[25]. With our k-mer profiling strategy, the identification of cancer-specific mutations is freed from any mapping step and those generating at least two overlapping cancer-specific k-mers should be represented in the final DB. The identification of VTPVYQHL, an mTSA derived from a large deletion, in EL4 cells clearly illustrates this point (**Figure 4.2**, page 190) and suggests that our DB-building strategy has the potential to capture the overall landscape of mTSAs in tumors. Lastly, it is a *modular* and *highly flexible* method to generate small DBs for MS (~ 6 times bigger than the reference proteome). Indeed, the code was implemented so that changing or adding additional k-mer databases to filter the cancer k-mers against requires close to no change. Moreover, in this project, we only restricted our analysis to the landscape of TSAs presented by MHC I molecules. However, recent reports have shown that TSAs presented by MHC II molecules might be the main targets of anti-

tumor responses[26,27], so that their landscape might be worth investigating. Besides isolating MHC II- rather than MHC I-associated peptides when preparing tumor samples for MS, the sole change in our DB-building strategy would be to increase k-mer length from 33 to 48 nucleotides, MHC II-associated peptides being on average 16 amino acid-long[28,29].

Despite all those advantages, one significant improvement could be investigated in order to get a more comprehensive picture of the mTSA landscape. For now, we can identify mTSAs directly overlapping cancer-specific mutations. Indeed, our k-mer profiling strategy was designed to extract, reconstruct and translate mutation-centered sequences, provided that the underlying mutation is cancer-specific. However, when they break codon periodicity, frameshift mutations, including insertion/deletions, do generate cancer-specific proteins simply because translation of their direct downstream, yet unmutated, sequence occurs in a cancer-specific frame. Thus, for frameshift mutations, translation of mutation-centered contigs is not as important as identifying and translating downstream normal sequences in the created cancer-specific frame. To do so, one could simply extract mutation-centered contigs, expected to be 65 nts-long when k = 33 nts ($contig\ length = k \times 2 - 1$), and determine their genomic location using fm-indexes of the genome and the transcriptome. Creating the fm-index of a large input text file (like the human genome or transcriptome) is advantageous as it drastically compresses it, while allowing you to access extremely quickly the location of any substring of interest (contigs in our case) within this file[30]. Thus, to determine the genomic location of a contig, one could submit all of its possible substrings to pre-computed genomic and transcriptomic fm-indexes in order to find a concordant set of positions (chromosome, start, stop). Of course, longer substrings are expected to be more informative than shorter ones, as those are likely to have numerous matches in the genome. Besides, information gathered from the transcriptomic fm-index are here to help determining if contigs derived from non-contiguous genomic regions do overlap with known exon-exon junctions. Once the contig has been given a genomic location, it is now possible to compare its nucleotide sequence with the one of the reference genome and identify the nature of its underlying

mutation (single-base, insertion, deletion, etc.). Then, for each insertion/deletion-derived contig, one can (i) extract its downstream sequence (according to the strand it derives from) and (ii) *in silico* 3-frame translate this extended contig down to the first stop codon. The implementation and automation of the first steps of this workflow, i.e., assign a genomic location to contigs and identify the underlying mutation, are currently undertaken by Eric Audemard at the bioinformatic platform. Thus, in the near future, we should be able to identify TSAs derived from frameshift mutations, antigens which are likely to be more immunogenic than those derived from single-mutations[31,32].

5.3 On the origin of cryptic MAPs

Despite the methodological concerns discussed in the **sections 5.1** and **5.2**, our comprehensive proteogenomic analysis provides the first, and sole, systems-level analysis of the cryptic MAP repertoire isolated from normal B-LCLs. Because this dataset is the biggest dataset of cryptic MAPs ever collected on one cell type, this allowed us to determine what is captured by the cryptic MAP repertoire.

5.3.1 Giving a voice to non-coding transcripts and another perspective on protein-coding ones

On normal cells, we showed that cryptic MAPs represent ~ 10% of the MAP repertoire and that they derive from (i) the out-of-frame translation of protein-coding transcripts, (ii) the translation of non-coding regions within protein-coding transcripts (5'/3' UTRs, introns) and (iii) the translation of unknown transcripts (intergenic) as well as transcripts assumed to be non-coding (pseudogenes, antisense, **Figure 2.3**, page 86). In terms of proportions, it is interesting to note that those cryptic MAPs derive mostly from atypical translation of protein-coding transcripts rather than from translation of truly non-coding transcripts. In line with this, ~ 13% of human short ORFs annotated up to now derives from non-coding transcripts, the rest being attributed to translation of 5'UTRs, 3'UTRs and coding sequence (out-of-frame)[33]. Moreover, this observation is coherent with the overall translational activity of cells obtained from Ribo-Seq studies, which have shown that (i) only a small fraction of non-coding RNAs are subjected to translation[34,35] and that (ii) cryptic proteins preferentially arise from atypical translation of protein-coding transcripts[36]. Lastly, even when translated, non-coding RNAs are likely to represent a poor source of MAPs, as they present with features likely to dampen their MAP-generating potential. First, non-coding transcripts tend to be express in a few copies per cell[37], while high transcript abundance is one of the key factors increasing the likelihood of transcripts to generate MAPs[38-40]. Second, non-coding transcripts tend to be shorter and contain fewer exons than protein-coding transcripts[41], two features known to positively influence MAP generation[38]. Lastly, they

tend to generate short protein products[33,42], which are expected, just by virtue of their size, to generate fewer MAPs than long proteins[38,43].

Within protein-coding transcripts, cryptic MAPs tend to derive from: out-of-frame translation of exons > 5'UTR > 3'UTR > intron (**Figure 2.3c**, page 86). This gradation is not unexpected as it correlates with the representation of those regions within Ribo-Seq reads where ~ 85% of reads map to annotated coding sequences, ~ 5-10% to 5'UTRs and the rest to ~ 3'UTRs, intronic regions and so on[16]. Nonetheless, we know that our workflow likely underestimates the proportion and the diversity of cryptic MAPs derived from intronic regions. Indeed, Apcher *et al.* demonstrated that nuclear translation of pre-spliced mRNA was able to generate intronic MAPs[44]. Although this discovery is (for now) only supported by data involving transfection of cells with model peptide-containing constructs, it strongly suggests that nuclear translation is instrumental for the generation of cryptic MAPs. However, because we performed a poly(A) enrichment prior to RNA-Seq, those pre-spliced transcripts are likely to be underrepresented in our RNA-Seq data and in the resulting DB.

5.3.2 More mutated than conventional MAPs?

One startling observation we made when looking at the genomic origin of cryptic MAPs was that their peptide-coding sequences tend to overlap with more non-synonymous polymorphisms than the ones encoding conventional MAPs. Although surprising at first, two explanations can be envisioned. First, it is a fact that non-coding regions do accumulate more mutations than coding ones[45], as the probability of selecting against a deleterious non-synonymous mutations is higher in coding than in non-coding regions[46]. Consequently, cryptic MAPs coming from non-coding regions have a higher chance of being mutated than conventional ones. Second, the evolutionary constraints exerted on the protein-coding genome pushes the synonymous-to-non-synonymous mutational ratio towards more synonymous mutations, in order to decrease the likelihood of functional consequences for the organism[47]. Due to the redundancy of the genetic code, those synonymous mutations are expected to mainly affect the third nucleotide of each codon, as changing this

nucleotide often yields the same amino acid during translation. Because we observed that a large proportion of cryptic MAPs derive from the out-of-frame translation of protein-coding transcripts, such mutations considered synonymous when translated in their original reading frame are likely to become non-synonymous in any of the two alternative reading frames. Consequently, cryptic MAPs derived from the out-of-frame translation of protein-coding transcripts are likely to turn synonymous mutations into non-synonymous ones. Knowing that such synonymous mutations can be deleterious for translation by altering its speed or accuracy[48], one could hypothesize that such mutated cryptic MAPs flag soon-to-be neoplastic cells, thereby triggering their rapid elimination by the immune system.

5.3.3 More than just translational noise?

Because cryptic MAPs derive from non-coding regions within protein-coding transcripts, and to a lesser extent, from non-coding transcripts, one might assume that they represent 'translational noise' with limited to no physiological relevance. Several points argue against this hypothesis. First, most cryptic MAPs are likely to derive from reproducible non-canonical translation events, as we showed that cryptic MAPs can be reproducibly detected in the MAP repertoire of unrelated subjects in an MHC I-dependent fashion (**Figure 2.2**, page 84). Second, cryptic proteins, which generate cryptic MAPs, show evolutionary conservation. Indeed, homologs of those proteins can be found across several species, including *S. Cerevisiae*, suggesting that those proteins are functional[33,49]. In line with this, several studies have reported the implication of short proteins in various cellular processes[50,51]. Third, a Ribo-Seq study performed by the group of Jonathan Weissman revealed that, for a subset of cryptic proteins, it is not about sequence conservation but rather about conservation of translational activity[36]. This observation suggests that, if the resulting protein is not necessarily functional and likely to be quickly degraded, the act of translation might have an important regulatory role. In line with this, we observed that cryptic MAPs (i) derived from regions that were more polymorphic than conventional MAPs at the population level, hence less conserved (**Figure 2.6c**, page 94) and that (ii) their cryptic

source transcripts were likely targeted by the nonsense-mediated decay pathway, a translation-coupled mRNA degradation mechanism (**Figure 2.4**, page 89).

Altogether, it is clear that both the overall translational activity and the mutational landscape of cells are well captured by the MAP repertoire, as MAPs derived from both coding and allegedly non-coding regions are presented. Moreover, cryptic MAPs are unlikely to derive from translational noise as they originate from either conserved sequence or conserved translational activity. Consequently, presentation of cryptic MAPs significantly expands the scope of CD8+ T-cell immunosurveillance.

5.4 On the peculiar biogenesis of cryptic MAPs

Besides deriving from genomic regions initially thought to be non-coding, several line of evidence suggest that the biogenesis of cryptic MAPs significantly differs from the one of conventional MAPs.

5.4.1 Translation initiation

It is well-established that translation initiation requires an AUG within a strong Kozak context[52]. However, for ~ 40% of our cryptic MAPs, translation initiation was predicted to occur at near-cognate or non-cognate start codons rather than an AUGs (**Figure 2.5a,b**, page 92). This observation suggests that translation of cryptic proteins, hence cryptic MAPs, requires an initiation factor different than the one involved in translation of conventional proteins. eIF2A would be an interesting initiation factor to look at, as it has been shown to initiate translation at near-cognate AUGs and that its activity can be selectively inhibited in cells[53]. Besides, eIF2A knock out mice do exist and could be leveraged for this analysis[54].

The fact that cells might produce cryptic and conventional MAPs using different translation initiation mechanisms is likely to ensure the presentation of a wider repertoire of MAPs. In line with this, we observed that most transcripts generating cryptic MAPs do not generate conventional MAPs (**Figure 2.4a**, page 89). Moreover, it might allow cells to present MAPs under a wider range of environmental conditions. In line with this, the integrated stress response can shut down eIF2-dependent translation but not eIF2A-dependent translation, which maintains translational activity on stress-related transcripts[55] as well as viral transcripts[56].

5.4.2 Proteasome-independence of cryptic MAPs

It is well-established that proteasomal cleavage determines the C-terminal end of most conventional MAPs. However, when comparing the C-terminal end of conventional vs. cryptic MAPs, we observed that they differ in their amino acid usage, thereby suggesting that cryptic MAPs were processed in a proteasome-independent fashion (**Figure 2.5d**, page 92). In line with this, we observed that a significant

proportion of cryptic MAPs were located at the C-terminal of their predicted source proteins (**Chapter 3**, page 156).

To test if cryptic MAPs are indeed processed in a proteasome-independent fashion, one could predict and compare the presence of proteasomal degradation motifs[57-59] and the extent of disorder[60-63], which is known to favor proteasomal degradation, between conventional and cryptic proteins. Besides, a chemoproteomic screen could be carried out as follow: a panel of inhibitors for proteases, that have been previously involved in the MHC I antigen presentation, could be used to probe the proteolytic landscape of cells and determine those that are instrumental for cryptic MAP generation[64]. Briefly, B-LCLs could be treated with increasing doses of each inhibitor and MAP would be extracted by immunoaffinity purification for each condition. Then, using synthetic ^{13}C-analogs of a pre-defined set of conventional and cryptic MAPs, those samples could be analyzed by PRM-MS in order to derive MAP copy number per cell for each peptide in each condition. From there, one could compute the percentage of presentation inhibition for each peptide in each condition as follow: $(N_t/N_u) \times 100$, with N_t and N_u the copy number obtained for the treated and untreated condition, respectively. If our hypothesis is true, proteasome inhibition should reduce the copy number per cell of most conventional MAPs in a dose-dependent fashion, while leaving cryptic MAPs unaffected.

Altogether, our observations suggest that the differences between the biogenesis of cryptic and conventional MAPs is instrumental to the survey of different portion of the transcriptome. Although interesting, this hypothesis will require further validation implying the characterization of the MAP repertoire from additional samples across several tissue types to see if this is indeed the case and if those transcripts are enriched in specific functional categories.

5.5 Cryptic MAPs in cancer…

Having established that cryptic MAPs represent a significant portion of the MAP repertoire and that they are not mere translational noise, we wanted to use proteogenomics in order to investigate the contribution of cryptic MAPs, with a focus on TSAs, to the MAP repertoire of murine and human tumors.

5.5.1 Sensor for neoplastic transformation?

Cryptic MAPs represent ~ 10% of all MAPs on normal cells (**Figure 2.1b**, page 82), while cryptic TSAs represent ~ 90% of all identified TSAs on tumor cells (**Figure 4.2d,** page 190 and **Figure 4.6c,** page 201). Assuming that there is no difference of processing by the MHC I antigen presentation pathway between proteins generating TSAs and those that do not, one could extrapolate that the overall MAP repertoire of tumor cells contains a larger proportion of cryptic MAPs than the one of normal cells. Before getting excited, it is important to be aware that at least three factors can bias this comparison:

1. The normal and tumor samples were not analyzed using the same proteogenomic approach. Indeed, our comprehensive approach, used to analyze the normal samples, generates bigger than databases than our targeted approach and is consequently more likely to underestimate the contribution of cryptic MAPs to the MAP repertoire.
2. The contribution of cryptic TSAs to the overall TSA repertoire might not accurately estimate the contribution of cryptic MAPs to the overall repertoire, as they represent less than 1% of it.
3. The tissue of origin as well as the MHC I haplotype of each sample is likely to influence presentation of cryptic MAPs[65,66].

Nonetheless, the idea that neoplastic transformation somehow favors the generation of cryptic translation products, thereby increasing the presentation of cryptic MAPs, is an interesting idea that could mechanistically make sense for several reasons.

Compared to normal cells:

1. The transcriptome of tumor cells is likely to express a wider pool of transcripts derived from non-coding regions. Indeed, the overall genome of cancer cells tends to be hypomethylated, while DNA methylation is an epigenetic mark usually associated to transcriptional repression[67-69]. Thus, hypomethylated DNA loci, which mainly include low-complexity DNA such as EREs, should be highly transcribed in cancer cells. Besides that, cancer cells have been shown to express a wide repertoire of abnormally-spliced transcripts, i.e., transcripts retaining introns[70]. In line with this, we observed that about half of cryptic TSAs derived from EREs or introns (**Supplementary Tables 4.6, 17 and 18**, page 288).
2. The proteome of tumor cells might contain, or at least produce, more cryptic proteins. It is well-established that cancer cells are *de facto* stressed cells[71], probably because most of them have to deal with aneuploidy[72] besides being exposed to difficult environmental conditions[73]. Consequently, several cancer types show a constitutive activation of the unfolded protein response[74], which decreases cap-dependent translation by triggering the phosphorylation of eIF2 on its α subunit[75]. However, studies using reporter constructs have shown that non-canonical translation initiation at near-cognate AUGs relies on eIF2A rather than eIF2[53,76]. Under such cellular stress response, eIF2A-mediated translation appears instrumental to the maintenance of translational activity on stress-related transcripts[55] and can also turn on IRES-dependent translation[56,77,78].
3. The immunogenicity of cancer cells *in vivo* positively correlates with their level of ER stress[79]. By challenging mice with parental or hyperploid tumor cell lines (obtained by inducing tetraploidization with nocodazole), our collaborators showed that, compared to their parental counterpart, hyperploid clones exhibit constitutive ER stress and are more efficiently eliminated by the immune system. This elimination, which is dependent on CD8+ T cells, does not involve the

recognition of conventional hyperploidy-associated antigens (**Figure AII.6**, page xlv), suggesting a role for cryptic MAPs in tumor cell recognition.

To formally test this hypothesis, one could use our (improved) comprehensive DB-building strategy (RNA/Ribo-Seq) to compare the landscape of cryptic MAPs in the parental vs. hyperploid tumor cell lines generated by our collaborators[79]. An increased representation of cryptic MAPs in the repertoire of hyperploid vs. parental clones would suggest that this increase in ER stress is sufficient to favor the generation of cryptic MAPs. Additional evidence could be gathered by evaluating to which extent the induction of ER stress in parental clones modulates the repertoire of cryptic MAPs. Provided that this artificial model validates our hypothesis, one could then move on to a more physiological model, i.e., human primary tumor samples and their normal counterpart.

5.5.2 Source of shared TSAs?

Looking more precisely at the origin of those cryptic TSAs, we made the startling observation that aeTSAs were way more frequent than mTSAs, even for our two highly mutated murine tumor cell lines. The fact that aeTSAs represent ~ 85 % of all identified TSAs could be ascribed to the fact that we might underestimate the proportion of mTSAs. Indeed, we know that our method is not optimal for the detection of all mTSAs (**section 5.2**). Moreover, tumor populations are often composed of multiple subclones, each expressing a different set of mutations[80-82]. As such, mTSAs might be less abundant than aeTSAs, which could prevent their detection by MS. In line with this, the aeTSA VNYLHRNV, which derive from a highly expressed transcript, is present at an extremely high copy number, while the mTSA VNYIHRNV, which derived from a lowly expressed transcript (suggesting its subclonality), could not be re-detected, even when using PRM-MS (**Figure 4.5b**, page 197). Besides, as discussed in **section 5.5.1**, this identification of aeTSAs is coherent with the increase in transcriptome diversity that neoplastic transformation induces.

The fact that aeTSAs are the main type of TSAs presented by tumors is in itself highly interesting. Indeed, because those sequences are cancer-specific, yet

unmutated, it opens up the idea that even lowly mutated tumor types can generate such TSAs. In line with this, we observed that B-ALLs, which are known to have a very low mutational load[83,84], did present aeTSAs (**Supplementary Table 4.17**, page 288). Moreover, these aeTSAs have the potential to be shared (i) by several patients within the same tumor type, as exemplified by the identification of SLTALVFHV, shared by two of our B-ALLs samples (**Figure 4.6b**, page 201), and (ii) by various tumor types. To estimate the extent of TSA sharing, it might be worth looking at the expression of regions coding for those TSAs across publicly available cohorts of cancer patients, such as those available from The Cancer Genome Atlas. Lastly, aeTSAs could be induced artificially using DNA demethylating agents. Indeed, such agents have been shown to increase the presentation of cancer-testis antigens, a particular type of aeTSAs[85,86] as well as expression of genes involved in antigen processing[87].

5.5.3 Source of immunogenic TSAs?

To evaluate the immunogenic potential of aeTSAs and mTSAs, we decided to use five EL4 TSAs that were truly tumor-specific, as they did not show any expression in normal peripheral tissues. By vaccinating mice with TSA-pulsed DCs, we observed that those TSAs were not equally efficient to prevent or delay leukemia onset (**Figure 4.3**, page 192). This anti-leukemic effect was dictated by two main factors, namely the frequency of TSA-specific T cell in the naive repertoire and the expression of the antigen by cancer cells, likely reflecting the proportion of tumor cell expressing it (**Figure 4.4a**, page 194 and **Figure 4.5a,b**, page 197). This last observation is consistent with the fact that, in humans, neoantigens predicted to be more clonal were associated with an increased immune cell infiltration in tumors and a better response to checkpoint blockade[88].

Because our workflow was supposed to enrich for non-tolerogenic sequences (i.e., absent from TEC/mTEC), the fact that two our TSAs poorly protected mice against the EL4 challenge was quite surprising (**Figure 4.3a**, page 192). Knowing that mTECs express TRAs in a mosaic fashion[89], those TSAs might derive from lowly expressed sequences that were missed by our RNA-Seq data despite its depth (**Supplementary**

Table 4.2, page 288). If this is the case, this would argue that those antigens were actually subjected to central tolerance and therefore explain the rare frequency of their T-cell repertoire and their absence of expansion following immunization (**Figure 4.4a,b**, page 194). The same phenomenon might have occurred when we profiled the expression of those antigens in peripheral tissues (**Figure 4.2c**, page 190) and would argue that those antigens are also subjected to peripheral tolerance mechanisms[90].

VNYLHRNV, which was able to protect 100% of the mice (**Figure 4.3b**, page 192), is an aeTSAs derived from a murine ERE (**Supplementary Table 4.6b**, page 288). Its mutated counterpart, VNYIHRNV, could only protect 20% of the mice, emphasizing the importance of high expression/clonality (**Figure 4.3b**, page 192 and **Figure 4.5a,b**, page 197). In humans, we expect ERE TSAs, which are remnants of viral infection, to be highly immunogenic. Indeed, mTSAs that resemble viral antigens were shown to be enriched in long-term survivors of pancreatic cancer and correlate with immune infiltration[84]. Moreover, recent studies have shown that de-repression of such ERE sequences in human cancer correlates with T-cell infiltration, increased antigenicity[91] and improved response to checkpoint inhibitors[92,93].

Our last antigen, VTPVYQHL, which was able to protect 30% of the mice (**Figure 4.3c**, page 192), is an interesting TSAs as it derives from a large intergenic deletion. Although we have no proof that this mutation is a driver in EL4 cells, it is known that several cancer types are characterized by fusions. In prostate cancer, the presence of predicted MAPs overlapping gene fusion correlated with tumor infiltration[94]. In leukemia, several peptides overlapping oncogene fusion proteins were explored for binding to MHC I[95]. Because some of those fusion genes are recurrent drivers mutations across patients and that such TSAs are expected to be more immunogenic than TSAs derived from single-base mutations, those antigens might be worth targeting in humans.

Altogether, those observations demonstrate that cryptic TSAs are a valuable pool of TSAs for the development of immunotherapies as they are more numerous than conventional TSAs, probably because their production is favored by the neoplastic transformation process. Moreover, these antigens have the potential to be shared

between tumors, while being highly immunogenic. A note of caution though… the peripheral expression profile of human aeTSAs seems to be less clear cut than the one of murine TSAs, as there is often a few samples of various tissues slightly expressing the antigen (**Figure 4.2c**, page 190 vs. **Figure 4.6c**, page 201). Knowing that high transcript expression is often a prerequisite for MAP generation, there is no certainty with regard to their actual processing in normal tissues. Nonetheless, those antigens, as opposed to TAAs, derive from sequences absent from the transcriptome of TEC/mTEC and are therefore less likely to have been centrally tolerized. Thus, it might be worth testing a few murine antigens displaying such low-grade peripheral expression to see if these can trigger autoimmune responses. In the eventuality of such autoimmune reactions, a more suitable way to remove k-mers encoding potentially harmful TSAs might be to subtract all normal k-mers from cancer k-mers, that is, to remove k-mers detected in TEC/mTEC as well as those detected in MHC I-expressing peripheral tissues.

5.6 Cryptic TSAs, next-generation targets for T-cell based cancer immunotherapy?

The observation that infiltration of tumors by some types of immune cells, including CD8+ T cells, positively correlates with patient survival lead to the development of so-called immunotherapies. These therapies, which try to harness the power of the immune system to fight cancer, can range from complex adoptive cell transfers to simple vaccines. In *adoptive cell transfer,* the goal is to replace the patient's immune system with *ex vivo* expanded immune cells that are better suited to fight the tumor[96]. Transferred cells can either be (i) autologous TILs[97] and (ii) autologous or donor-derived T cells engineered to express an antigen-binding receptor able to specifically recognize tumor cells. These engineered T cells can either express *engineered TCRs*, that is TCRs designed to better recognize a given peptide/MHC complex[98], or *chimeric antigen receptors* (CARs), which combine the antigen-recognition domain of a BCR with the signaling domain of a TCR to recognize a given cell surface protein. In both cases, these engineered T cells can eliminate any cell presenting with the cell surface marker they are design to target. Aside from the use of anti-CD19 CARs in the treatment of B-ALLs, no other adoptive cell transfer protocols have been approved by the Food and Drug Administration probably because these therapies can lead to severe (and even lethal) adverse effects[99] while being extremely costly[100]. An alternative, and less invasive, way of treating cancer patients is to leverage our knowledge of the immune system to modulate pre-existing anti-tumor responses, thereby favoring tumor clearance. In the 90s, discovery of PD-1 and CTLA-4, negative regulators of activated T cells, catalyzed the development of so-called *immune checkpoint inhibitors*, which can increase T-cell reactivity against tumors by masking those negative regulators on the surface of T cells. Despite some long-lasting remissions observed in melanoma patients, injections of immune checkpoint inhibitors only show a 25-40% response rate and can lead to severe adverse effects for patients[101,102], breaks being removed on both tumor-specific and bystander T cells. Assuming that this low response rate is due to the poor cross-presentation of relevant TSAs by DCs (**Figure 4.5d**, page 197), increasing antigen display with *anti-tumor*

vaccines appears as a relevant and cost-effective strategy to amplify immune responses relevant to tumor rejection. Despite the existence of promising vaccination platforms such as the one developed by Ugur Sahin's group[103,104], vaccination does require an *a priori* knowledge of the antigens to be targeted. Up to now, only vaccines with conventional mTSAs, identified through prediction-based or MS-based workflows, have been tested because these antigens are the sole for which reliable identification workflows are available[3]. However, as demonstrated by our study (**Chapter 4**), conventional mTSAs only represent a small fraction of the antigenic landscape of tumors, cryptic aeTSAs being the main class of anti-tumor antigens. Having developed a TSA identification workflow able to identify both conventional and cryptic aeTSAs, we do believe that we can now push forward the development of therapeutic anti-cancer vaccines for both highly and lowly mutated cancer types. Nonetheless, several questions need to be addressed before proceeding to any clinical implementation of such anti-cancer vaccines.

First, it will be essential to thoroughly evaluate how common are those aeTSAs in tumors. Indeed, in our study, we only analyzed two cancer types located at opposite ends of the mutational spectrum, namely some highly mutated lung tumors (n = 3) and some lowly mutated B-ALL specimens (n = 4). Although we did identify TSAs in both tumor types, a demonstration that such aeTSAs can be found in other, if not all, cancer types is definitely required. Thus, similarly to what was done by The Cancer Genome Atlas consortium, one could apply our TSA identification workflow, which uses shotgun MS, to small cohorts of each tumor type (n = 20) and get a sneak peek to their respective antigenic landscape. Then, additional tumor samples should be analyzed to precisely quantify the extent of TSA sharing among patients both at the RNA level, using tumor RNA-Seq data, and at the peptide level, by PRM-MS, a technique far more sensitive than shotgun MS. Indeed, PRM-MS requiring a consequent amount of tumor material, it might be worth comparing results obtained at the RNA and at the peptide level to evaluate if RNA-Seq data alone could be used to predict TSA presentation. Lastly, because we observed that the frequency of TSA-specific T cells in the pre-immune repertoire of mice positively influenced tumor rejection (**Figure 4.4**, page 194),

estimating the mean frequency of TSA-specific T cells in cohorts of healthy donors for each aeTSA seems to be a must. Performed on enough donors, such T-cell reactivity screening could give us a pretty good idea on the recognition potential of each aeTSA at the population level. From there, one could create a database aggregating the sequences of all aeTSAs with information on their shared and recognition potential in order to rank them according to their relevance for anti-cancer vaccination, going from *universal antigens* –shared by all patients of all tumor types– to *private antigens* –shared by few patients of few tumor types– that must be recognized by frequent T-cell repertoires in healthy subjects.

Second, although adding aeTSAs to the pool of actionable TSAs surely increases the likelihood of finding universal anti-tumor antigens, it is important to keep in mind that tumor heterogeneity is not only intertumoral (i.e., between different tumor types) but also intratumoral (i.e., between tumor cells of a given patient) ! Thus, we are in need of studies evaluating the impact of intratumoral heterogeneity on the TSA landscape of tumors. To do so, one could apply our TSA identification workflow on cohorts of patients for which multiple tumor sites can be sampled (e.g., different regions of a given primary tumor, primary vs. metastatic tumor lesions). However, with the low amount of starting material available, it might be more realistic to infer the TSA landscape of each tumor site by cross-referencing multi-region RNA-Seq data with our database of aeTSAs, that is, to consider a given tumor region positive for a given aeTSA only if this region expresses the coding sequence of this TSA. Then, on a per patient basis, those data could be used to derive a 3D map of their TSA landscape in order to classify aeTSAs as *clonal antigens*, when they are shared by all tumor sites of a given patient, or as *subclonal antigens*, when they are present in one to a few tumor sites. As several studies, including ours (**Figure 4.5**, page 197), suggested that clonal TSAs lead to better anti-tumor responses than subclonal ones[88], an ideal anti-tumor vaccine should target either (i) clonal TSAs or (ii) enough subclonal TSAs to cover the overall clonal architecture of the tumor. Although a 3D map of the TSA landscape could help for such patient-specific TSA prioritization task, it is important to remember that the antigenic landscape of each tumor is likely to evolve over time and treatments, with

some TSAs disappearing and others appearing. Thus, this analysis should be regularly repeated to ensure that the composition of the vaccine stays optimal for the patient. Moreover, collecting such longitudinal data on the evolution of so many TSA landscapes could help identifying other factors that do influence anti-tumor responses. For instance, it might be interesting to compare the overall survival and the relapse rate of patients following vaccination with cryptic vs. conventional aeTSAs. Indeed, one could hypothesize that tumor genomes are more permissive to the loss of cryptic TSA-coding sequences, as those might encode non-functional proteins.

Third and last, it will be important to explore ways of turning non-immunogenic tumors into immunogenic ones to ensure that all patients could receive such vaccine. As a first step, it might be interesting to conduct some sort of exploratory 'chemo-proteogenomic' screen on a collection of well-established human tumor cell lines to evaluate the impact of clinically used anticancer compounds on their respective TSA landscape. Among others, it might worth including drugs that could favor the generation of cryptic TSAs such as aminoglycosides, known to promote stop codon bypass and consequently increase presentation of MAPs derived from 3'UTRs[105], demethylating agents, known to de-repress EREs which are a rich source of immunogenic TSAs[106-108] (**Figure 4.3**, page 192), and proteasome inhibitors, known to increase non-canonical translation in response to proteotoxic stress[55]. Then, by comparing the induced and baseline TSA landscapes, one could derive three important metrics for each tested compound:

1. a *diversity index* to estimate changes in the width of the TSA landscape, that is the balance between *de novo* expressed and repressed TSA-coding sequences.
2. an *intensity index* to estimate changes in the amplitude of the TSA landscape, that is the proportion of highly vs. lowly expressed TSA-coding sequences.
3. a *similarity index* to estimate the extent of inter-cell line variability, that is the proportion of TSA-coding sequences induced in all vs. only a few cell lines.

Compounds triggering the *de novo* expression of numerous TSAs (high diversity index) at high levels (high intensity index) across all cell lines (high similarity index) are likely

to be the most interesting ones to treat patients prior to vaccination. Indeed, by increasing the diversity and the expression level of TSAs, these compounds increase tumor immunogenicity and therefore widen the pool of TSAs that might be targeted by anti-cancer vaccines. Moreover, by forcing the expression of similar TSAs across tumors of different origin, these compounds are likely to decrease the need for personalized vaccines and therefore decrease the cost of the overall therapy.

5.7 Conclusion

Throughout this thesis, we developed two proteogenomics approaches, readily available, that helped us unravel the contribution of non-canonical translation events to the MAP repertoire of normal and cancer cells.

In normal cells, we found that cryptic MAPs represent ~ 10% of the MAP repertoire. These MAPs derive from (i) the out-of-frame translation of protein-coding transcripts, (ii) the translation of non-coding regions within protein-coding transcripts (UTRs, introns) and (iii) the translation of unannotated transcripts as well as transcripts assumed to be non-coding. Interestingly, we report that their biogenesis differ from the of conventional MAPs, as they derived from the translation of unstable transcripts which produce very short proteins that might by degraded in a proteasome-independent fashion. Lastly, we showed that regions coding for cryptic MAPs tend to accumulate more polymorphisms than conventional ones, and some polymorphic cryptic MAPs can be immunogenic.

Across a panel of two murine cancer cell lines and seven human primary tumors (four B-ALLs and three lung tumor biopsies), we were able to show that ~ 90 % of TSAs are cryptic TSA. Moreover, we show that most of these TSAs are aeTSAs, which include EREs. Because of their normal, yet cancer-specific nature, these antigens are likely (i) to be shared by multiple patients and (ii) to be detected on additional lowly mutated tumors types. Lastly, we also demonstrate that, in mice, the anti-tumor potential of TSA vaccination is influenced by (i) the frequency of TSA-specific T cells in the naïve compartment and (ii) the expression of the antigen in tumors cells, i.e., clonality.

Altogether, our findings demonstrate that cryptic MAPs significantly expand the immune self and, consequently, the scope of CD8+ T cell immunosurveillance. Provided that aeTSAs are indeed shared in humans, they could prompt the development of off-the-shelf cancer immunotherapies targeting those antigens.

5.8 References

1. Granados, D.P., et al., *Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides.* Nat Commun, 2014. **5**: p. 3600.
2. Granados, D.P., et al., *Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers.* Leukemia, 2016.
3. Bassani-Sternberg, M., et al., *Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.* Nat Commun, 2016. **7**: p. 13404.
4. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2012. **13**(1): p. 36-46.
5. Gardai, S.J., et al., *Cell-surface calreticulin initiates clearance of viable or apoptotic cells through trans-activation of LRP on the phagocyte.* Cell, 2005. **123**(2): p. 321-34.
6. Cobbold, M., et al., *MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia.* Sci Transl Med, 2013. **5**(203): p. 203ra125.
7. Malaker, S.A., et al., *Identification of Glycopeptides as Posttranslationally Modified Neoantigens in Leukemia.* Cancer Immunol Res, 2017. **5**(5): p. 376-384.
8. Liepe, J., et al., *A large fraction of HLA class I ligands are proteasome-generated spliced peptides.* Science, 2016. **354**(6310): p. 354-358.
9. Mylonas, R., et al., *Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome.* Mol Cell Proteomics, 2018.
10. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.
11. Noble, W.S., *Mass spectrometrists should search only for peptides they care about.* Nat Methods, 2015. **12**(7): p. 605-8.

12. Murphy, J.P., et al., *MHC-I Ligand Discovery Using Targeted Database Searches of Mass Spectrometry Data: Implications for T-Cell Immunotherapies.* J Proteome Res, 2017. **16**(4): p. 1806-1816.
13. Levin, J.Z., et al., *Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nat Methods, 2010. **7**(9): p. 709-15.
14. Zhao, S., et al., *Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap.* BMC Genomics, 2015. **16**: p. 675.
15. King, H.A. and A.P. Gerber, *Translatome profiling: methods for genome-scale analysis of mRNA translation.* Brief Funct Genomics, 2016. **15**(1): p. 22-31.
16. Calviello, L. and U. Ohler, *Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome.* Trends Genet, 2017.
17. Michel, A.M., et al., *Observation of dually decoded regions of the human genome using ribosome profiling data.* Genome Res, 2012. **22**(11): p. 2219-29.
18. Brar, G.A. and J.S. Weissman, *Ribosome profiling reveals the what, when, where and how of protein synthesis.* Nat Rev Mol Cell Biol, 2015. **16**(11): p. 651-64.
19. Crappe, J., et al., *PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration.* Nucleic Acids Res, 2015. **43**(5): p. e29.
20. Erhard, F., et al., *Improved Ribo-seq enables identification of cryptic translation events.* Nat Methods, 2018.
21. Li, H., et al., *Systematic Comparison of False-Discovery-Rate-Controlling Strategies for Proteogenomic Search Using Spike-in Experiments.* J Proteome Res, 2017. **16**(6): p. 2231-2239.
22. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.
23. Marcais, G. and C. Kingsford, *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.* Bioinformatics, 2011. **27**(6): p. 764-70.
24. Ballouz, S., et al., *The fractured landscape of RNA-seq alignment: the default in our STARs.* Nucleic Acids Res, 2018. **46**(10): p. 5125-5138.

25. Kumar, S., et al., *Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data.* Sci Rep, 2016. **6**: p. 21597.
26. Kreiter, S., et al., *Mutant MHC class II epitopes drive therapeutic immune responses to cancer.* Nature, 2015. **520**(7549): p. 692-6.
27. Tran, E., et al., *Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer.* Science, 2014. **344**(6184): p. 641-5.
28. Chicz, R.M., et al., *Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size.* Nature, 1992. **358**(6389): p. 764-8.
29. Clement, C.C., et al., *The Dendritic Cell Major Histocompatibility Complex II (MHC II) Peptidome Derives from a Variety of Processing Pathways and Includes Peptides with a Broad Spectrum of HLA-DM Sensitivity.* J Biol Chem, 2016. **291**(11): p. 5576-95.
30. Ferragina, P. and G. Manzini, *Opportunistic Data Structures with Applications.* Proceedings of the 41st Annual Symposium on Foundations of Computer Science., 2000: p. 390.
31. Turajlic, S., et al., *Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis.* Lancet Oncol, 2017. **18**(8): p. 1009-1021.
32. Capietto, A.H., S. Jhunjhunwala, and L. Delamarre, *Characterizing neoantigens for personalized cancer immunotherapy.* Curr Opin Immunol, 2017. **46**: p. 58-65.
33. Samandi, S., et al., *Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins.* Elife, 2017. **6**.
34. Lee, S., et al., *Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution.* Proc Natl Acad Sci U S A, 2012. **109**(37): p. E2424-32.
35. Chugunova, A., et al., *Mining for Small Translated ORFs.* J Proteome Res, 2018. **17**(1): p. 1-11.

36. Fields, A.P., et al., *A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation.* Mol Cell, 2015. **60**(5): p. 816-27.
37. Palazzo, A.F. and T.R. Gregory, *The case for junk DNA.* PLoS Genet, 2014. **10**(5): p. e1004351.
38. Pearson, H., et al., *MHC class I-associated peptides derive from selective regions of the human genome.* J Clin Invest, 2016. **126**(12): p. 4690-4701.
39. Fortier, M.H., et al., *The MHC class I peptide repertoire is molded by the transcriptome.* J Exp Med, 2008. **205**(3): p. 595-610.
40. Granados, D.P., et al., *MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements.* Blood, 2012. **119**(26): p. e181-191.
41. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
42. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.
43. Bassani-Sternberg, M., et al., *Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.* Mol Cell Proteomics, 2015. **14**(3): p. 658-73.
44. Apcher, S., et al., *Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17951-6.
45. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nat Rev Genet, 2016. **17**(2): p. 93-108.
46. Halligan, D.L., et al., *Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila.* Genome Res, 2004. **14**(2): p. 273-9.
47. Li, W.H., C.I. Wu, and C.C. Luo, *A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.* Mol Biol Evol, 1985. **2**(2): p. 150-74.

48. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias.* Nat Rev Genet, 2011. **12**(1): p. 32-42.
49. Vanderperre, B., et al., *Direct detection of alternative open reading frames translation products in human significantly expands the proteome.* PLoS One, 2013. **8**(8): p. e70698.
50. Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**(3): p. 193-204.
51. D'Lima, N.G., et al., *A human microprotein that interacts with the mRNA decapping complex.* Nat Chem Biol, 2017. **13**(2): p. 174-180.
52. Kozak, M., *Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes.* Cell, 1986. **44**(2): p. 283-92.
53. Starck, S.R., et al., *Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I.* Science, 2012. **336**(6089): p. 1719-23.
54. Golovko, A., et al., *The eIF2A knockout mouse.* Cell Cycle, 2016. **15**(22): p. 3115-3120.
55. Starck, S.R., et al., *Translation from the 5′ untranslated region shapes the integrated stress response.* Science, 2016. **351**(6272): p. aad3867.
56. Kim, J.H., et al., *eIF2A mediates translation of hepatitis C viral mRNA under stress conditions.* EMBO J, 2011. **30**(12): p. 2454-64.
57. Liu, Z., et al., *GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes.* PLoS One, 2012. **7**(3): p. e34370.
58. Rechsteiner, M. and S.W. Rogers, *PEST sequences and regulation by proteolysis.* Trends Biochem Sci, 1996. **21**(7): p. 267-71.
59. Chen, X., et al., *Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites.* Bioinformatics, 2013. **29**(13): p. 1614-22.
60. Prakash, S., et al., *An unstructured initiation site is required for efficient proteasome-mediated degradation.* Nat Struct Mol Biol, 2004. **11**(9): p. 830-7.

61. van der Lee, R., et al., *Intrinsically disordered segments affect protein half-life in the cell and during evolution.* Cell Rep, 2014. **8**(6): p. 1832-44.
62. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
63. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. **21**(16): p. 3433-4.
64. Parmentier, N., et al., *Production of an antigenic peptide by insulin-degrading enzyme.* Nat Immunol, 2010. **11**(5): p. 449-54.
65. Paul, S., et al., *HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity.* J Immunol, 2013. **191**(12): p. 5831-9.
66. de Verteuil, D.A., et al., *Immunoproteasomes shape the transcriptome and regulate the function of dendritic cells.* J Immunol, 2014. **193**(3): p. 1121-32.
67. Gama-Sosa, M.A., et al., *The 5-methylcytosine content of DNA from human tumors.* Nucleic Acids Res, 1983. **11**(19): p. 6883-94.
68. Feinberg, A.P. and B. Vogelstein, *Hypomethylation distinguishes genes of some human cancers from their normal counterparts.* Nature, 1983. **301**(5895): p. 89-92.
69. Feinberg, A.P. and B. Vogelstein, *Hypomethylation of ras oncogenes in primary human cancers.* Biochem Biophys Res Commun, 1983. **111**(1): p. 47-54.
70. Dvinge, H. and R.K. Bradley, *Widespread intron retention diversifies most cancer transcriptomes.* Genome Med, 2015. **7**(1): p. 45.
71. Luo, J., N.L. Solimini, and S.J. Elledge, *Principles of cancer therapy: oncogene and non-oncogene addiction.* Cell, 2009. **136**(5): p. 823-37.
72. Gordon, D.J., B. Resio, and D. Pellman, *Causes and consequences of aneuploidy in cancer.* Nat Rev Genet, 2012. **13**(3): p. 189-203.
73. Eales, K.L., K.E. Hollinshead, and D.A. Tennant, *Hypoxia and metabolic adaptation of cancer cells.* Oncogenesis, 2016. **5**: p. e190.

74. Oakes, S.A. and F.R. Papa, *The role of endoplasmic reticulum stress in human pathology.* Annu Rev Pathol, 2015. **10**: p. 173-94.
75. Harding, H.P., Y. Zhang, and D. Ron, *Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase.* Nature, 1999. **397**(6716): p. 271-4.
76. Liang, H., et al., *PTENalpha, a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism.* Cell Metab, 2014. **19**(5): p. 836-48.
77. Reineke, L.C., et al., *Insights into the role of yeast eIF2A in IRES-mediated translation.* PLoS One, 2011. **6**(9): p. e24492.
78. Kwon, O.S., et al., *An mRNA-specific tRNAi carrier eIF2A plays a pivotal role in cell proliferation under stress conditions: stress-resistant translation of c-Src mRNA is mediated by eIF2A.* Nucleic Acids Res, 2017. **45**(1): p. 296-310.
79. Senovilla, L., et al., *An immunosurveillance mechanism controls cancer cell ploidy.* Science, 2012. **337**(6102): p. 1678-84.
80. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
81. Jamal-Hanjani, M., et al., *Tracking the Evolution of Non-Small-Cell Lung Cancer.* N Engl J Med, 2017. **376**(22): p. 2109-2121.
82. Zhang, A.W., et al., *Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer.* Cell, 2018. **173**(7): p. 1755-1769 e22.
83. Ma, X., et al., *Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours.* Nature, 2018. **555**(7696): p. 371-376.
84. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
85. Natsume, A., et al., *The DNA demethylating agent 5-aza-2'-deoxycytidine activates NY-ESO-1 antigenicity in orthotopic human glioma.* Int J Cancer, 2008. **122**(11): p. 2542-53.
86. Kirkin, A.F., et al., *Adoptive cancer immunotherapy using DNA-demethylated T helper cells as antigen-presenting cells.* Nat Commun, 2018. **9**(1): p. 785.

87. Siebenkas, C., et al., *Inhibiting DNA methylation activates cancer testis antigens and expression of the antigen processing and presentation machinery in colon and ovarian cancer cells.* PLoS One, 2017. **12**(6): p. e0179501.
88. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade.* Science, 2016.
89. Sansom, S.N., et al., *Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia.* Genome Res, 2014. **24**(12): p. 1918-31.
90. Mueller, D.L., *Mechanisms maintaining peripheral tolerance.* Nat Immunol, 2010. **11**(1): p. 21-7.
91. Kong, Y., et al., *Transposable element expression in tumors is associated with immune infiltration and increased antigenicity.* bioRxiv preprint, 2018.
92. Sheng, W., et al., *LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade.* Cell, 2018. **174**(3): p. 549-563 e19.
93. Smith, C.C., et al., *Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma.* J Clin Invest, 2018.
94. Kalina, J.L., et al., *Mutational Analysis of Gene Fusions Predicts Novel MHC Class I-Restricted T-Cell Epitopes and Immune Signatures in a Subset of Prostate Cancer.* Clin Cancer Res, 2017. **23**(24): p. 7596-7607.
95. Bocchia, M., et al., *Specific binding of leukemia oncogene fusion protein peptides to HLA class I molecules.* Blood, 1995. **85**(10): p. 2680-4.
96. Rosenberg, S.A. and N.P. Restifo, *Adoptive cell transfer as personalized immunotherapy for human cancer.* Science, 2015. **348**(6230): p. 62-8.
97. Zacharakis, N., et al., *Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer.* Nat Med, 2018. **24**(6): p. 724-730.
98. Debets, R., et al., *TCR-engineered T cells to treat tumors: Seeing but not touching?* Semin Immunol, 2016. **28**(1): p. 10-21.
99. Tey, S.K., *Adoptive T-cell therapy: adverse events and safety switches.* Clin Transl Immunology, 2014. **3**(6): p. e17.

100. Prasad, V., *Immunotherapy: Tisagenlecleucel - the first approved CAR-T-cell therapy: implications for payers and policy makers.* Nat Rev Clin Oncol, 2018. **15**(1): p. 11-12.
101. Larkin, J., F.S. Hodi, and J.D. Wolchok, *Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma.* N Engl J Med, 2015. **373**(13): p. 1270-1.
102. Hellmann, M.D., et al., *Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden.* N Engl J Med, 2018. **378**(22): p. 2093-2104.
103. Kranz, L.M., et al., *Systemic RNA delivery to dendritic cells exploits antiviral defence for cancer immunotherapy.* Nature, 2016. **534**(7607): p. 396-401.
104. Sahin, U., et al., *Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer.* Nature, 2017. **547**(7662): p. 222-226.
105. Goodenough, E., et al., *Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3′ UTR.* Proc Natl Acad Sci U S A, 2014. **111**(15): p. 5670-5.
106. Roulois, D., et al., *DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts.* Cell, 2015. **162**(5): p. 961-73.
107. Chiappinelli, K.B., et al., *Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses.* Cell, 2015. **162**(5): p. 974-86.
108. Chiappinelli, K.B., et al., *Combining Epigenetic and Immunotherapy to Combat Cancer.* Cancer Res, 2016. **76**(7): p. 1683-9.

APPENDIX I

AI. The nature of self for T cells – a systems-level perspective

Diana P. Granados[1,2], Céline M. Laumont[1,2], Pierre Thibault[1,3] and Claude Perreault[1,2]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, PO Box 6128, Downtown Station, Montreal, QC, Canada, H3C 3J7

[2]Department of Medicine, Université de Montréal, PO Box 6128, Downtown Station, Montreal, QC, Canada, H3C 3J7

[3]Department of Chemistry, Université de Montréal, PO Box 6128, Downtown Station, Montreal, QC, Canada, H3C 3J7

Correspondence to:
Claude Perreault
claude.perreault@umontreal.ca

Current Opinion in Immunology, Volume 34, pages 1-8 (November 1, 2014)

AI.1 Authors' contributions

Diana P. Granados: significantly contributed to writing.

Céline M. Laumont: prepared all figures and contributed to writing.

Pierre Thibault: contributed to design of Figure AI.1 and contributed to writing.

Claude Perreault: wrote the first draft of the manuscript.

All authors edited and approved the final version of the manuscript.

AI.2 Abstract

T-cell development and function are regulated by MHC-associated self peptides, collectively referred to as the immunopeptidome. Large-scale mass spectrometry studies have highlighted three key features of the immunopeptidome. First, it is not a mirror of the proteome or the transcriptome, and its content cannot be predicted with currently available bioinformatic tools. Second, the immunopeptidome is more plastic than previously anticipated, and is molded by several cell-intrinsic and cell-extrinsic factors. Finally, the complexity of the immunopeptidome goes beyond the 20-amino acids alphabet encoded in the germline, and is not restricted to canonical reading frames. The large amounts of ‘dark matter’ in the immunopeptidome, such as polymorphic, cryptic and mutant peptides, can now be explored using novel proteogenomic approaches that combine mass spectrometry and next-generation sequencing.

AI.3 Introduction

Recognition of self has a pervasive influence on the development and functions of the immune system because the adaptive lymphocytes of jawed vertebrates are selected on self-molecules, sustained by self-molecules, and activated in the presence of self-molecules[1]. This raises the fundamental question: what is the molecular definition of self for the adaptive immune system? We will discuss how large-scale mass spectrometry (MS) studies have recently allowed systems immunologists to unveil the molecular composition of the MHC-restricted immunopeptidome. Since most MS studies have been performed on peptides presented by MHC class I molecules (HLA class I in humans), they will represent the main focus of this article.

AI.4 The origin and role of the immune self recognized by CD8T cells

The T cell receptor (TCR) of classic adaptive CD8T cells recognizes MHC I-associated peptides (MIPs). In the absence of infection or phagocytosis, all MIPs derive from endogenous self proteins: these self MIPs are collectively referred to as the self MHC I immunopeptidome (SMII). The SMII regulates positive and negative selection in the thymus, as well as survival and reactivity of CD8T cells to non-self in the periphery[2]. The dominant paradigm holds that MIPs derive primarily from proteasomal digestion of rapidly degraded proteins (RDPs), which are degraded by the ubiquitin proteasome system during or within one hour after their synthesis[3]. In line with this, the majority of ubiquitination events in cells occurs on newly synthesized proteins[4]. A large proportion of RDPs are defective ribosomal products (DRiPs) (**Figure AI.1**)[5,6]. In theory, many processes can lead to DRiPs formation, but their relative importance remains elusive[3,5]. Most, though not all authors[7], consider that the contribution of old proteins (retirees) to the SMII is relatively modest[5], except under two circumstances: (i) following irradiation, which leads to rapid degradation of old proteins[8], and (ii) in the case of cross-presented MIPs because they derive from endocytosis of stable proteins[9]. A schematic overview of the processes involved in the genesis of the SMII is shown in **Figure AI.1**.

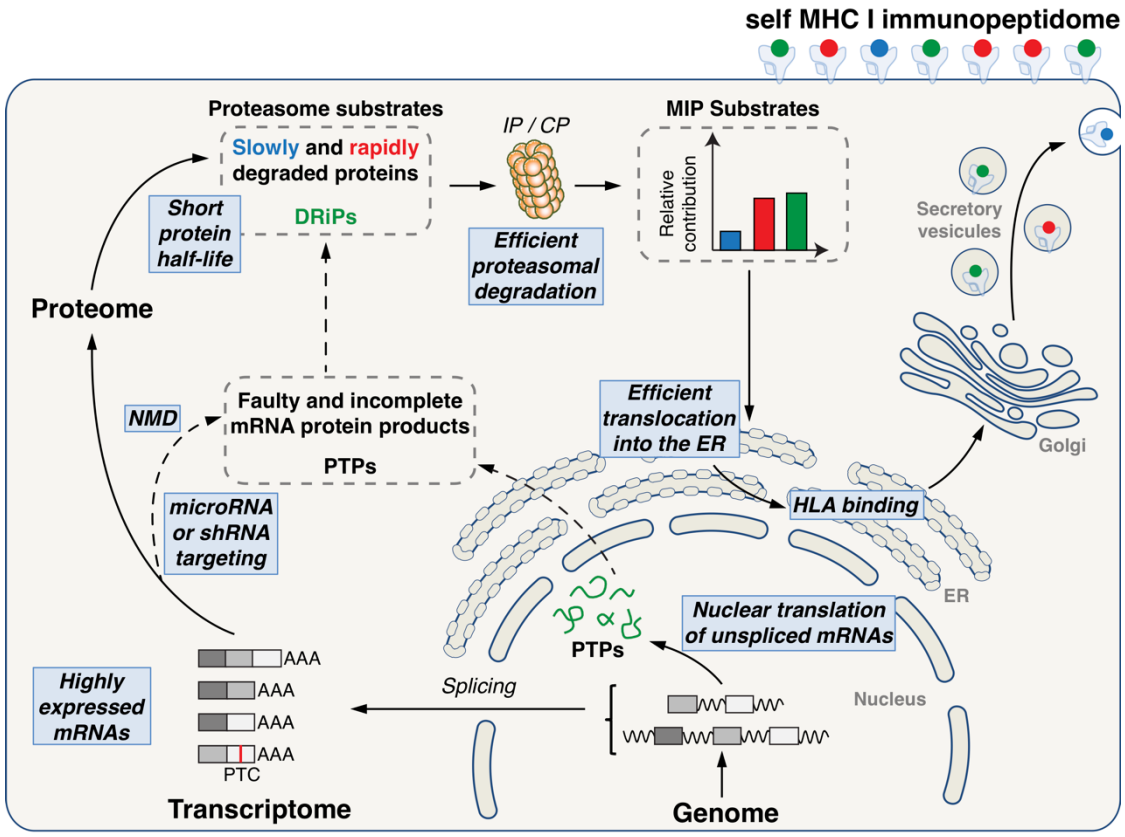


Figure AI.1 | Overview of the major processes involved in the genesis of the SMII. Text in rectangular boxes indicates features whose contribution to MIP generation is demonstrated or suspected. The role of transporters, chaperones and non-proteasomal proteases is not depicted here. CP, constitutive proteasomes; ER, endoplasmic reticulum; IP, immunoproteasomes; NMD, nonsense-mediated decay; PTC, premature termination codons; PTP, pioneer translation products.

AI.5 The SMII is not a mirror of the proteome

AI.5.1 Overview of the proteome and the SMII

Large-scale MS offers the sole direct approach to analyze the molecular composition of the proteome and the SMII[10-13]. The proteome of human cells contains about 10^4 different proteins with a mean length of 449 amino acids[12]. The number of protein copies per cell spans a concentration range of seven orders of magnitude (up to 20×10^6)[12]. In HeLa cells, the 40 most abundant proteins comprised 25% of the proteome and the most abundant 600 proteins constitute 75% of the proteome mass[11]. It has been estimated that a typical nucleated cell expresses at the cell surface about $1–5 \times 10^5$ MHC class I molecules that present about 10^4 different MIPs, 95% of which have a median length of 9 amino acids[10,14-16]. Hence, the SMII would represent at most 2% of the proteome: 10^4 MIPs @ 9 amino acids per MIP deriving from a proteome containing 10^4 proteins @ 449 amino acids/protein[10,12,14,15]. MIPs derive from a wide variety of proteins distributed in all cell compartments, encoded by genes located on all chromosomes[14,17,18]. Simply by virtue of their size, longer proteins generate more MIPs than shorter proteins[14,15,19]. For individual MIPs, the mean number of copies per cell is about 50 with extremes ranging from 1 to about 10,000[14,20,21]. Estimates based on analyses of human EBV lymphoblastoid cell lines suggest that the five most abundant MIPs already ‘fill’ 6% of the HLA class I molecules, the top 50 fill 27%, the top 300 fill 59%, and the top 1000 fill 83%[14]. MHC allotypes display more or less stringent peptide binding motifs and on average, each MHC class I allotype presents more than 1000 different MIPs, although the size of the peptide repertoire is widely variable across allotypes[22]. Further development of quantitative large-scale MS could allow the estimation of a diversity index for each particular SMII. In analogy to ecological sciences, the ‘SMII diversity index’ would represent a quantitative measure that reflects how many different MIPs compose the SMII of a given cell under specific conditions, and simultaneously take into account how evenly the number of copies per MIP are distributed.

AI.5.2 Limited overlap between the proteome and the immunopeptidome

There is only a modest correlation between the amounts of MIPs and the relative abundance of their source protein: some MIPs derive from low abundance proteins whereas some highly abundant proteins do not generate MIPs[14,19 23-25]. Two related factors explain why the immunopeptidome is not a mirror of the proteome. First, the SMII preferentially derives from RDPs whereas the most abundant proteins in the proteome have a long half- life[26]. Although long-lived proteins contribute to the SMII, their representation in the SMII is significantly lower than their relative proportions in the proteome. Indeed, each protein consists of two subpopulations: proteins that achieve a native conformation and are degraded at protein-specific rate, and DRiPs with a shorter half-life[27]. Irrespective of whether their native form has a long or short half-life, DRiPs are included in the SMII. Accordingly, by determining the relative rates of synthesis of MIPs and of their source proteins using large-scale stable isotope labeling by amino acids in cell culture, Admon's group concluded that a significant portion of the SMII is derived from DRiPs[13].

Second, different proteins have different DRiP rates. Though many processes may lead to DRiP formation, the relative importance of these processes as well as the physical nature of DRiPs remain ill-defined. In practice, various putative mechanisms of DRiP formation have been identified: nonsense-mediated decay of mRNAs carrying premature termination codons[28], and destabilization of mRNAs by miRNAs[29] or shRNAs[30]. However, the relative impact of these processes on MIP generation could depend on the cell type or metabolic state, and have yet to be assessed in large-scale experiments.

AI.6 Numerous factors enhance the complexity of the SMII

AI.6.1 Cell lineage and metabolic stage

Large-scale peptidomic studies have shown that the SMII is complex and plastic. Different tissues/cell types display different MIP repertoires[31,32]. Out of 614 MIPs detected on a mouse pancreatic b cell line, 314 (51%) were absent on thymocytes and splenocytes; treatment of b cells with interferon gamma (IFNγ) increased the number of MIPs to 883, of which 321 (36%) were absent on thymic and spleen cells[20]. Likewise, 40% of MIPs eluted from freshly harvested C57BL/6 thymocytes were absent on C57BL/6 dendritic cells (DCs)[31]. Adamopoulou *et al.* analyzed two populations of thymocyte-depleted thymic cells obtained from children undergoing corrective cardiac surgery: CD11c+ DCs and CD11c– cells (a mixture of epithelial and lymphomyeloid cells)[32]. Strikingly, 83% of MIPs were unique to CD11c+ or CD11c– cells. In B lymphocytes, the MIP-coding exome is enriched in genes implicated in immunoglobulin production[29]; in LPS-treated myeloid DCs, it is enriched in genes involved in myeloid differentiation, antigen processing and Toll-like receptor signaling[31]. Hence, the MIP repertoire conceals a cell-type-specific signature, which is closely linked to cell function.

Treatment with a selective mTORC1 inhibitor (rapamycin) led to significant changes in the abundance (2–15-fold) of 53% of MIPs present on EL4 cells[33]. Differentially expressed MIPs resulted from events at the transcriptional, translational and post-translational level, and derived from proteins tightly connected to the mTORC1 signaling pathway[33]. Notably, six MIPs detected on rapamycin-treated cells were ‘neo-MIPs’, absent on untreated cells. Immunization against two of these neo-MIPs elicited cytotoxic T cell responses. This suggests that cells can communicate their metabolic status to the adaptive immune system and that CD8T cells are reactive to neo-MIPs expressed on metabolically stressed cells. Altogether, these studies demonstrate that the SMII projects at the cell surface a representation of biochemical networks and metabolic events regulated at multiple levels inside the cell[33].

AI.6.2 Inflammation, infection and drugs

Inflammation and infection have cell-extrinsic and cell-intrinsic effects on the SMII. The inflammatory environment causes ER stress, largely via protein oxidation and deprivation of oxygen and nutrients. In turn, ER stress impinges on the SMII by decreasing protein synthesis and increasing protein degradation[33,34]. ER stress is linked to autophagy, and degradation of self proteins by autophagy generates MHC II-associated peptides[35]. Although autophagy has been reported to enhance the presentation of viral MIPs[36], it is unclear whether it can also contribute to the presentation of self MIPs. Cytokines in the inflammatory milieu can affect MIP generation. Thus, out of 883 MIPs detected in pancreatic b cells treated in vitro with IFNγ, 55% were absent on untreated cells[20]. The appearance of numerous MIPs after IFNγ treatment is probably due to the upregulation of molecules involved in MHC antigen processing, including immunoproteasomes. Indeed, immunoproteasomes significantly increase MIP abundance and diversity[31,37]. Other inflammatory cytokines most likely affect the SMII but their effect has yet to be assessed at the systems-level[38].

Viruses can affect the MHC antigen presentation pathway *per se*[39], and also have pleiotropic effects on synthesis and degradation of host proteins, two key processes that mold the SMII[33]. Hence, aside from expression of viral MIPs, infection leads to drastic changes in the profile of self MIPs expressed at the cell surface[40,41]. Wahl *et al.* observed that influenza-infected cells expressed 20 neo self MIPs (absent in uninfected cells) and over-expressed 347 other self MIPs[41]. Viral infections are frequently associated with autoimmune diseases, but the mechanistic underpinnings of this association remain controversial. An important unanswered question is whether virus-induced changes in the SMII are mechanistically linked to autoimmunity.

Finally, drugs can dramatically modify the SMII. Abacavir and carbamazepine can bind to the MIP-binding groove of specific HLA-B molecules (e.g., HLA-B*57:01 and HLA-B*15:02), thereby altering their MIP repertoire and triggering autoimmune-like reactions (reviewed in 42). A totally different mechanism is involved in the case of drugs such as aminoglycosides. Indeed, aminoglycosides such as gentamicin induce stop

codon read-through at a level sufficient to generate immunogenic MIPs derived from 30 untranslated regions[25].

AI.7 The dark matter in the immunopeptidome – the hidden side of self

The complexity of the SMII is spelled in words that go beyond a 20-amino acids alphabet encoded in the germline and restricted to canonical reading frames[43]. It is of utmost importance to understand that classically, peptide identification by MS is based on search algorithms that compare observed peptide fragments with a reference database such as Uniprot[44]. Thus, MS software tools search for a match between sample peptides and the reference proteome (encoded by canonical reading frames). Hence, standard high-throughput MS is typically blind to a whole universe of peptides which are not included in protein databases (**Figure AI.2**): (i) MIPs encoded by regions containing genomic polymorphisms[21], (ii) MIPs derived from unconventional translational mechanisms (cryptic MIPs), (iii) MIPs bearing somatic mutations (mutant MIPs). Therefore, MIPs of prime interest in transplantation and cancer immunology cannot be detected by standard MS approaches. This limitation can now be overcome with proteogenomic methods that combine transcriptomic data with MS[45,46].

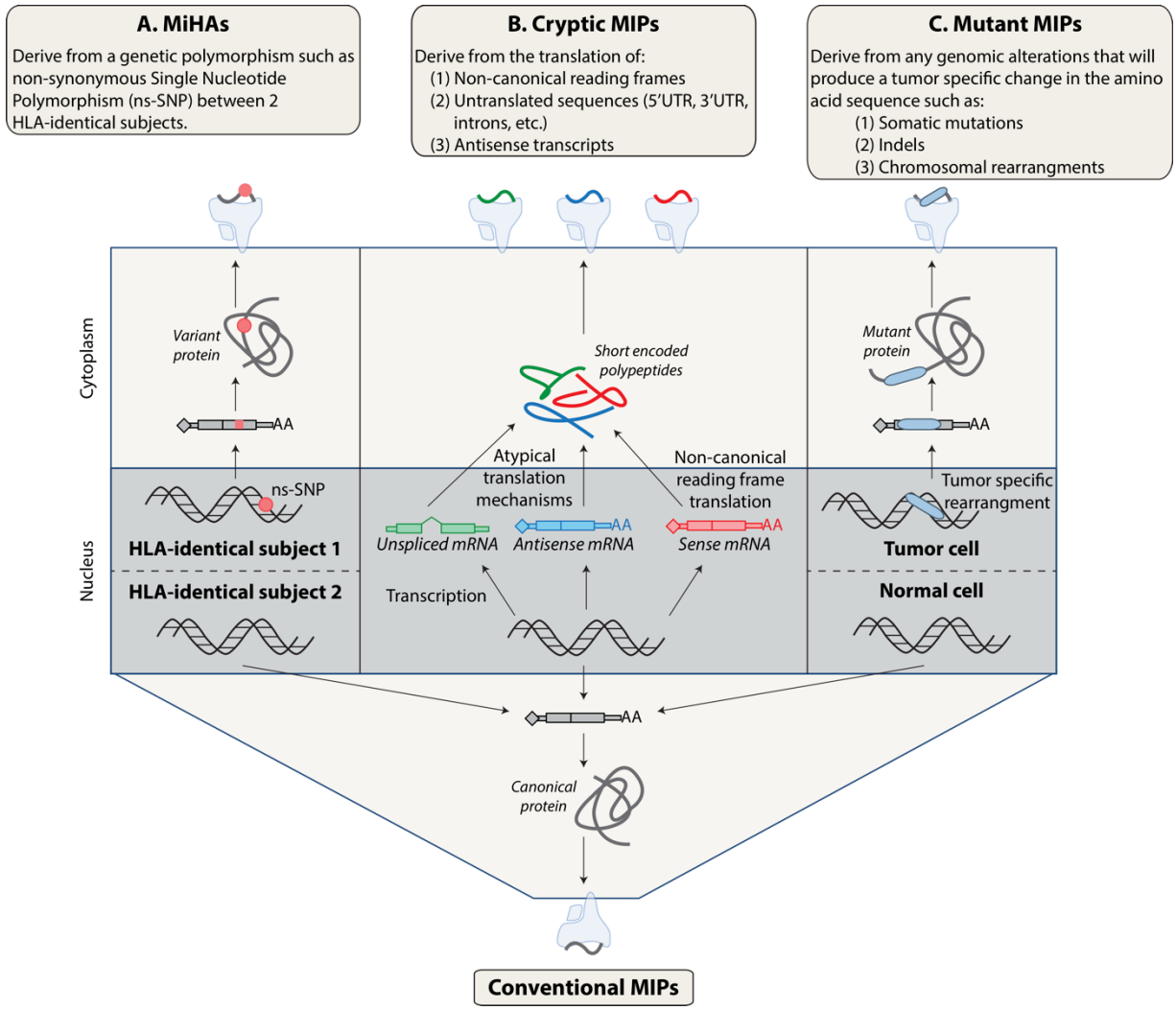


Figure AI.2 | The 'dark matter' of the SMII. Non-canonical MIPs that can be discovered using high-throughput proteogenomics include MiHAs, cryptic MIPs and mutant MIPs.

AI.7.1 MiHAs

For historical reasons, MHC-associated peptides coded by polymorphic genomic sequences are referred to as MiHAs[47]. MiHAs are essentially genetic polymorphisms viewed from a T-cell perspective. MiHA-coding alleles can be dominant or recessive at the peptide level. Thus, a non-synonymous single nucleotide polymorphism (ns-SNP) in a MIP-coding genomic sequence will either hinder MIP generation (recessive allele) or generate a variant MIP (dominant allele)[48]. MiHAs are

medically relevant because they elicit three types of alloimmune response: graft rejection, graft-versus-host disease and the allogeneic graft-versus-leukemia effect, which currently represents the most widely effective strategy for cancer immunotherapy in humans[49]. Over the last three decades, 35 human MiHAs encoded by autosomes and presented by HLA class I molecules have been discovered using arduous reductionist approaches[47,48]. Yet, due to the lack of a suitable systems-level approach[50], the global impact of non-MHC genomic polymorphisms on the SMII (i.e., what proportion of MIPs are MiHAs) remained until recently a speculative question. We addressed this question using a combination of next-generation sequencing and high-throughput MS peptide identification[19]. Whole exome and transcriptome sequencing was performed on cells from two non- twin HLA-identical siblings. Next generation sequencing data were translated in silico to create personalized protein databases that were used for MS sequencing of MIPs. Out of 4468 MIPs sequenced by MS, 34 were MiHAs coded by dominant alleles: 24 of these alleles were shared by the two subjects and 10 were unique to one subject. The number of unshared MiHAs is expected to be 1.8-fold higher in unrelated (HLA-matched) subjects relative to siblings[47,48]. The dominant importance of the MHC genotype in molding the SMII is shown by the fact that, at the population level, 88% of the MIP-coding exome is invariant, and only 12% contains ns-SNPs[18]. Nonetheless, MiHAs are attractive targets for adoptive T-cell immunotherapy of cancer, particularly hematologic cancers. In mouse models injection of T cells targeted to a single MiHA can eradicate neoplastic cells without causing any graft-versus-host disease[21,51,52]. Because of the low number of molecularly defined human MiHAs, less than 30% of patients would be eligible for immunotherapy targeted to specific MiHAs[53]. Proteogenomic-based MiHA discovery should now rapidly overcome this caveat.

AI.7.2 Cryptic MIP

When studying the molecular bases of alloreactivity, Shastri *et al.* made a startling observation. Using a lacZ-inducible alloreactive T cell clone as a probe, they screened a splenic cDNA library in transiently transfected antigen-presenting cells and isolated a cDNA clone that allowed expression of the MIP recognized by the T cell

clone. The salient finding was that this peptide derived from a non-canonical reading frame initiated with a non-AUG start codon[54]. In a series of elegant studies, Shastri's group discovered that synthesis of this peptide was initiated with a CUG codon decoded as a leucine rather than a methionine[55]. More recently, studies using ribosome profiling have shown that the proteome is more complex than anticipated[56]. Indeed, ribosome profiling, an emerging technique that uses deep sequencing to monitor in vivo translation, has revealed that ribosomes occupy many regions of the transcriptome thought to be noncoding, including 50 UTRs and long noncoding RNAs[57]. These observations strongly suggest that translation is pervasive on cytosolic transcripts outside of conserved reading frames. Accordingly, recent studies suggest that MIPs can arise from untranslated regions (UTRs or introns) or from alternate translational reading frames[43,57,58]. In addition, seminal observations provide compelling evidence that MIPs arising from 'untranslated regions' may derive from immature mRNAs translated in the nucleus[6,58-60]. A few of these so-called 'cryptic MIPs' were validated by MS sequencing[25,61]. However, since most putative cryptic MIPs have not been validated by MS, some of them must be considered with skepticism because their identification relied primarily on methods fraught with high false discovery rates[62-64]: bioinformatic predictions (that cannot predict key steps in MIP processing) and in vitro T-cell reactivity assays (which are tainted by T-cell cross-reactivity). Systems-level evaluation of the landscape of cryptic MIPs is possible but will absolutely require unbiased proteogenomic studies. Ideally, this type of study would entail transcriptome sequencing and ribosome profiling of relevant cells followed by MS analyses in which MIP spectra are assigned by searching the six-frame *in silico* translation of transcriptomic reads. This strategy has been successfully used for identification of proteins in non-model species as well as higher eukaryotes[46,65,66].

AI.7.3 Mutant MIP

Several lines of evidence suggest that antigens expressed by neoplastic cells are able to elicit protective anti-tumor responses in humans and that T cells specific for these antigens are enriched in tumor-infiltrating lymphocytes (TILs)[67-69]. What is the nature of tumor antigens that can elicit protective anti-tumor responses? Tumor

antigens belong to two main classes: (i) tumor-associated antigens (TAAs) which are qualitatively normal MIPs overexpressed on cancer cells and (ii) tumor-specific antigens (TSAs) which result from somatic gene mutations or translocations and are truly tumor-specific. Based on the assumption that T cells recognizing the most immunogenic tumor antigens are enriched in TILs, two groups assessed whether melanoma TILs recognized the 175 TAAs identified to date on melanoma cells[70,71]. These comprehensive studies led to a striking conclusion: less than 1% of CD8 melanoma TILs recognized known TAAs. The inescapable corollary is that 99% of CD8 TILs are specific for unknown antigens. Since TILs recognize the most relevant (i.e., immunogenic) tumor antigens, it can be inferred that the most relevant tumor antigens have not been discovered yet and are most likely TSAs rather than TAAs[63,72,73]. It is therefore sobering that the landscape of human TSAs remains practically unknown and that the contribution of MS to this field has been minimal[74,75]. A few TSAs have been discovered one at a time, usually by laborious screening of cDNA libraries with antigen-reactive T cells[74]. More recently, a total of ten human TSAs have been discovered by two groups using strategies based on whole-exome sequencing of melanoma tumor DNA from 4 subjects, MHC binding predictions, and in vitro T-cell assays[63,76]. Though these landmark studies represent a progress in the field of TSA discovery, they illustrate the limitations inherent to strategies that do not include high-throughput MS. First, the number of TSAs discovered (ten in four subjects) was low considering that melanoma is the type of cancer bearing the highest number of mutations[77]. Indeed, recent studies in patients with ovarian cancer or cholangiosarcoma suggest that the number of TSAs that can be discovered with this approach will be even lower in non-melanoma tumors[78,79]. Second, more than 95% of TSA candidates identified with exome sequencing and MHC- binding predictions do not elicit T-cell responses[63,76,78], and are therefore probably false positives. Finally, identification of TSAs based on T-cell reactivity without validation by MS sequencing remains tentative[80].

AI.8 Concluding remarks and perspectives

Recent progress in proteogenomics enables large-scale MS sequencing of all MIPs whose source transcript can be sequenced by RNA-seq (essentially all transcripts, including unprocessed mRNAs). In principle, the sole MIPs that can elude RNA-seq based proteogenomics are those derived from non-contiguous protein sequences *via* proteasome-mediated splicing. Indeed, after excision of intervening fragments, proteasomes can create peptide bonds between distant protein segments. Though proteasomal splicing is a low efficiency process, it can generate biologically relevant MIPs including MiHAs and TSAs[80]. Further studies are needed to determine whether refinements in de novo MS sequencing (i.e., without the assistance of database search) could allow high-throughput analysis of MIPs generated via proteasome splicing. Lastly, quantitative aspects of the SMII are still missing (number of copies for each individual MIPs)[10], and need to be studied in order to assess changes in the diversity index of the SMII under different conditions. With this in hand, and with the collaboration of chemists and bioinformaticians, systems immunologists can now explore a whole new universe: the dark matter in the immunopeptidome. It is particularly urgent to qualitatively and quantitatively characterize the landscape of human TSAs by assessing the number of TSA copies per cancer cell and the extent of inter-tumor and intra-tumor heterogeneity in TSA expression.

AI.9 Acknowledgements

Research performed in the authors' labs was supported by grants from the Canadian Institutes of Health Research (CIHR) (MOP 42384) and the Canadian Cancer Society (grant # 701564). C.P. and P.T. hold Canada Research Chairs in Immunobiology, and Proteomics and Bioanalytical Spectrometry, respectively. The Institute for Research in Immunology and Cancer is supported in part by the Canada Foundation for Innovation and the Fonds de Recherche Santé Québec. Special thanks to our colleague Sébastien Lemieux for making bioinformatics 'as simple as possible, but not simpler'.

AI.10 References

1. Davis, M.M., et al., *T cells as a self-referential, sensory organ.* Annu Rev Immunol, 2007. **25**: p. 681-95.
2. Mandl, J.N., et al., *T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens.* Immunity, 2013. **38**(2): p. 263-274.
3. Anton, L.C. and J.W. Yewdell, *Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors.* J Leukoc Biol, 2014. **95**(4): p. 551-62.
4. Kim, W., et al., *Systematic and quantitative assessment of the ubiquitin-modified proteome.* Mol Cell, 2011. **44**(2): p. 325-40.
5. Yewdell, J.W., *DRiPs solidify: progress in understanding endogenous MHC class I antigen processing.* Trends Immunol, 2011.
6. Baboo, S., et al., *Most human proteins made in both nucleus and cytoplasm turn over within minutes.* PLoS One, 2014. **9**(6): p. e99346.
7. Rock, K.L., et al., *Re-examining class-I presentation and the DRiP hypothesis.* Trends Immunol, 2014. **35**(4): p. 144-52.
8. Reits, E.A., et al., *Radiation modulates the peptide repertoire, enhances MHC class I expression, and induces successful antitumor immunotherapy.* J Exp Med, 2006. **203**(5): p. 1259-71.
9. Norbury, C.C., et al., *CD8+ T cell cross-priming via transfer of proteasome substrates.* Science, 2004. **304**(5675): p. 1318-21.
10. Mester, G., V. Hoffmann, and S. Stevanovic, *Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands.* Cell Mol Life Sci, 2011. **68**(9): p. 1521-32.
11. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Mol Syst Biol, 2011. **7**: p. 548.
12. Beck, M., et al., *The quantitative proteome of a human cell line.* Mol Syst Biol, 2011. **7**: p. 549.

13. Bourdetsky, D., C.E. Schmelzer, and A. Admon, *The nature and extent of contributions by defective ribosome products to the HLA peptidome.* Proc Natl Acad Sci U S A, 2014. **111**(16): p. E1591-9.
14. Hassan, C., et al., *The human leukocyte antigen-presented ligandome of B lymphocytes.* Mol Cell Proteomics, 2013. **12**(7): p. 1829-43.
15. Mommen, G.P., et al., *Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD).* Proc Natl Acad Sci U S A, 2014. **111**(12): p. 4507-12.
16. Berlin, C., et al., *Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy.* Leukemia, 2015. **29**(3): p. 647-59.
17. Fortier, M.H., et al., *The MHC class I peptide repertoire is molded by the transcriptome.* J Exp Med, 2008. **205**(3): p. 595-610.
18. Granados, D.P., et al., *Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides.* Nat Commun, 2014. **5**: p. 3600.
19. Hoof, I., et al., *Proteome sampling by the HLA class I antigen processing pathway.* PLoS Comput Biol, 2012. **8**(5): p. e1002517.
20. Dudek, N.L., et al., *Constitutive and inflammatory immunopeptidome of pancreatic beta-cells.* Diabetes, 2012. **61**(11): p. 3018-25.
21. Vincent, K., et al., *Rejection of leukemic cells requires antigen-specific T cells with high functional avidity.* Biol Blood Marrow Transplant, 2014. **20**(1): p. 37-45.
22. Paul, S., et al., *HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity.* J Immunol, 2013. **191**(12): p. 5831-9.
23. Milner, E., et al., *The turnover kinetics of major histocompatibility complex peptides of human cancer cells.* Mol Cell Proteomics, 2006. **5**(2): p. 357-65.
24. Croft, N.P., et al., *Kinetics of antigen expression and epitope presentation during virus infection.* PLoS Pathog, 2013. **9**(1): p. e1003129.
25. Goodenough, E., et al., *Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3′ UTR.* Proc Natl Acad Sci U S A, 2014. **111**(15): p. 5670-5.

26. Boisvert, F.M., et al., *A quantitative spatial proteomics analysis of proteome turnover in human cells.* Mol Cell Proteomics, 2012. **11**(3): p. M111 011429.
27. Qian, S.B., et al., *Tight linkage between translation and MHC class I peptide ligand generation implies specialized antigen processing for defective ribosomal products.* J Immunol, 2006. **177**(1): p. 227-33.
28. Apcher, S., et al., *Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation.* Proc Natl Acad Sci U S A, 2011. **108**(28): p. 11572-7.
29. Granados, D.P., et al., *MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements.* Blood, 2012. **119**(26): p. e181-191.
30. Gu, W., et al., *Both treated and untreated tumors are eliminated by short hairpin RNA-based induction of target-specific immune responses.* Proc Natl Acad Sci U S A, 2009. **106**(20): p. 8314-9.
31. de Verteuil, D., et al., *Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules.* Mol Cell Proteomics, 2010. **9**(9): p. 2034-47.
32. Adamopoulou, E., et al., *Exploring the MHC-peptide matrix of central tolerance in the human thymus.* Nat Commun, 2013. **4**: p. 2039.
33. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.
34. Granados, D.P., et al., *ER stress affects processing of MHC class I-associated peptides.* BMC Immunol, 2009. **10**: p. 10.
35. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nat Rev Immunol, 2011. **11**(12): p. 823-36.
36. English, L., et al., *Autophagy enhances the presentation of endogenous viral antigens on MHC class I molecules during HSV-1 infection.* Nat Immunol, 2009. **10**(5): p. 480-7.
37. Kincaid, E.Z., et al., *Mice completely lacking immunoproteasomes show major changes in antigen presentation.* Nat Immunol, 2011. **13**(2): p. 129-35.

38. Morandi, F., et al., *IL-27 in human secondary lymphoid organs attracts myeloid dendritic cells and impairs HLA class I-restricted antigen presentation.* J Immunol, 2014. **192**(6): p. 2634-42.
39. Hansen, T.H. and M. Bouvier, *MHC class I antigen presentation: learning from viral evasion strategies.* Nat Rev Immunol, 2009. **9**(7): p. 503-13.
40. Hickman, H.D., et al., *Cutting edge: class I presentation of host peptides following HIV infection.* J Immunol, 2003. **171**(1): p. 22-6.
41. Wahl, A., et al., *HLA class I molecules reflect an altered host proteome after influenza virus infection.* Hum Immunol, 2010. **71**(1): p. 14-22.
42. Illing, P.T., et al., *Human leukocyte antigen-associated drug hypersensitivity.* Curr Opin Immunol, 2013. **25**(1): p. 81-9.
43. Starck, S.R. and N. Shastri, *Non-conventional sources of peptides presented by MHC class I.* Cell Mol Life Sci, 2011. **68**(9): p. 1471-9.
44. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.
45. Woo, S., et al., *Proteogenomic database construction driven from large scale RNA-seq data.* J Proteome Res, 2014. **13**(1): p. 21-8.
46. Branca, R.M., et al., *HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics.* Nat Methods, 2014. **11**(1): p. 59-62.
47. Warren, E.H., et al., *Effect of MHC and non-MHC donor/recipient genetic disparity on the outcome of allogeneic HCT.* Blood, 2012. **120**(14): p. 2796-806.
48. Spierings, E., et al., *Phenotype frequencies of autosomal minor histocompatibility antigens display significant differences among populations.* PLoS Genet, 2007. **3**(6): p. e103.
49. Vincent, K., D.C. Roy, and C. Perreault, *Next-generation leukemia immunotherapy.* Blood, 2011. **118**(11): p. 2951-9.
50. Hombrink, P., et al., *Discovery of T cell epitopes implementing HLA-peptidomics into a reverse immunology approach.* J Immunol, 2013. **190**(8): p. 3869-77.
51. Fontaine, P., et al., *Adoptive transfer of minor histocompatibility antigen-specific T lymphocytes eradicates leukemia cells without causing graft-versus-host disease.* Nat Med, 2001. **7**(7): p. 789-94.

52. Meunier, M.C., et al., *T cells targeted against a single minor histocompatibility antigen can cure solid tumors.* Nat Med, 2005. **11**(11): p. 1222-9.
53. Bleakley, M., et al., *Leukemia-associated minor histocompatibility antigen discovery using T-cell clones isolated by in vitro stimulation of naive CD8+ T cells.* Blood, 2010. **115**(23): p. 4923-33.
54. Malarkannan, S., M. Afkarian, and N. Shastri, *A Rare Cryptic Translation Product Is Presented by Kb Major Histocompatibility Complex Class I Molecule to Alloreactive T Cells.* J Exp Med, 1995. **182**: p. 1739-1750.
55. Starck, S.R., et al., *Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I.* Science, 2012. **336**(6089): p. 1719-23.
56. Ingolia, N.T., *Ribosome profiling: new views of translation, from single codons to genome scale.* Nat Rev Genet, 2014. **15**(3): p. 205-13.
57. Ingolia, N.T., et al., *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.
58. Apcher, S., et al., *Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17951-6.
59. David, A., et al., *Nuclear translation visualized by ribosome-bound nascent chain puromycylation.* J Cell Biol, 2012. **197**(1): p. 45-57.
60. Yewdell, J.W. and A. David, *Nuclear translation for immunosurveillance.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17612-3.
61. Weinzierl, A.O., et al., *A cryptic vascular endothelial growth factor T-cell epitope: identification and characterization by mass spectrometry and T-cell assays.* Cancer Res, 2008. **68**(7): p. 2447-54.
62. Popovic, J., et al., *The only proposed T-cell epitope derived from the TEL-AML1 translocation is not naturally processed.* Blood, 2011. **118**(4): p. 946-54.
63. Robbins, P.F., et al., *Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells.* Nat Med, 2013. **19**(6): p. 747-52.

64. Sewell, A.K., *Why must T cells be cross-reactive?* Nat Rev Immunol, 2012. **12**(9): p. 669-77.
65. Castellana, N. and V. Bafna, *Proteogenomics to discover the full coding content of genomes: a computational perspective.* J Proteomics, 2010. **73**(11): p. 2124-35.
66. Evans, V.C., et al., *De novo derivation of proteomes from transcriptomes for transcript and protein identification.* Nat Methods, 2012. **9**(12): p. 1207-11.
67. Restifo, N.P., M.E. Dudley, and S.A. Rosenberg, *Adoptive immunotherapy for cancer: harnessing the T cell response.* Nat Rev Immunol, 2012. **12**(4): p. 269-81.
68. Galon, J., et al., *The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures.* Immunity, 2013. **39**(1): p. 11-26.
69. Page, D.B., et al., *Immune modulation in cancer with antibodies.* Annu Rev Med, 2014. **65**: p. 185-202.
70. Andersen, R.S., et al., *Dissection of T-cell antigen specificity in human melanoma.* Cancer Res, 2012. **72**(7): p. 1642-50.
71. Kvistborg, P., et al., *TIL therapy broadens the tumor-reactive CD8(+) T cell compartment in melanoma patients.* Oncoimmunology, 2012. **1**(4): p. 409-418.
72. Heemskerk, B., P. Kvistborg, and T.N. Schumacher, *The cancer antigenome.* EMBO J, 2013. **32**(2): p. 194-203.
73. Hinrichs, C.S. and N.P. Restifo, *Reassessing target antigens for adoptive T-cell therapy.* Nat Biotechnol, 2013. **31**(11): p. 999-1008.
74. Coulie, P.G., et al., *Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy.* Nat Rev Cancer, 2014. **14**(2): p. 135-46.
75. Rosenberg, S.A., *Finding suitable targets is the major obstacle to cancer gene therapy.* Cancer Gene Ther, 2014. **21**(2): p. 45-7.
76. van Rooij, N., et al., *Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma.* J Clin Oncol, 2013. **31**(32): p. e439-42.
77. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.

78. Wick, D.A., et al., *Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer.* Clin Cancer Res, 2014. **20**(5): p. 1125-34.
79. Tran, E., et al., *Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer.* Science, 2014. **344**(6184): p. 641-5.
80. Vigneron, N. and B.J. Van den Eynde, *Proteasome subtypes and the processing of tumor antigens: increasing antigenic diversity.* Curr Opin Immunol, 2012. **24**(1): p. 84-91.

APPENDIX II

AII. Immunogenic stress and death of cancer cells: Contribution of antigenicity vs adjuvanticity to immunosurveillance

Norma Bloy[1,2,3,4,5,6,7] | Pauline Garcia[3,6,7] | Céline M. Laumont[8,9] | Jonathan M. Pitt[1,10,11] | Antonella Sistigu[12] | Gautier Stoll[1,2,3,4] | Takahiro Yamazaki[11] | Eric Bonneil8 | Aitziber Buqué[1,2,3,4,6] | Juliette Humeau[1,2,3,4,5,6,7] | Jan W. Drijfhout[13,14] | Guillaume Meurice[15] | Steffen Walter[16] | Jens Fritsche[17] | Toni Weinschenk[16,17] | Hans-Georg Rammensee[18] | Cornelis Melief[19] | Pierre Thibault[8,20] | Claude Perreault[8,9,21] | Jonathan Pol[1,2,3,4,6] | Laurence Zitvogel[10,11,22] | Laura Senovilla[1,2,3,4] | Guido Kroemer[1,2,3,4,5,23,24]

[1]Sorbonne Paris Cité, Université Paris Descartes, Paris, France

[2]Équipe 11 labellisée Ligue Nationale contre le Cancer, Centre de Recherche des Cordeliers, Paris, France

[3]Institut National de la Santé et de la Recherche Médicale, U1138, Paris, France
[4]Université Pierre et Marie Curie, Paris, France

[5]Metabolomics and Cell Biology Platforms, Gustave Roussy Cancer Campus, Villejuif, France

[6]Institut Gustave Roussy Cancer Campus, Villejuif, France

[7]Faculty of Medicine, University of Paris Sud, Kremlin-Bicêtre, France
[8]Institute for Research in Immunology and Cancer, Université de Montréal, Montreal, Quebec, Canada

[9]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada

[10]Center of Clinical Investigations in Biotherapies of Cancer (CICBT), Villejuif, France

[11]Institut National de la Santé et de la Recherche Médicale (INSERM), U1015, Equipe labellisée Ligue Nationale Contre le Cancer, Gustave Roussy Cancer Campus, Villejuif, France

[12]Unit of Tumor Immunology and Immunotherapy, Department of Research, Advanced Diagnostics and Technological Innovation, Regina Elena National Cancer Institute, Rome, Italy

[13]Department of Immunohematology and Blood Transfusion, Leiden University, Leiden,

The Netherlands

[14]Medical Center, Leiden, The Netherlands

[15]Bioinformatic Core Facility, UMS AMMICA, INSERM US23, CNRS UMS3665, Gustave Roussy, Villejuif, France 16Immatics US, Houston, TX, USA

[17]Immatics Biotechnologies, Tübingen, Germany

[18]Department of Immunology, Institute for Cell Biology, University of Tübingen, Tübingen, Germany

[19]ISA Pharmaceuticals, Leiden, The Netherlands

[20]Department of Chemistry, Faculty of Arts and Sciences, Université de Montréal, Montreal, Quebec, Canada

[21]Division of Hematology, Hôpital Maisonneuve-Rosemont, Montreal, Quebec, Canada

[22]Center of Clinical Investigations in Biotherapies of Cancer (CICBT), Villejuif, France

[23]Pôle de Biologie, Hôpital Européen Georges Pompidou, AP-HP, Paris, France

[24]Department of Women's and Children's Health, Karolinska University Hospital, Stockholm, Sweden

Correspondence to:
Dr. Jonathan Pol, Dr. Laura Senovilla and Dr. Guido Kroemer, UMR1138 équipe 11, Centre de Recherche des Cordeliers, 15, rue de l'Ecole de Médecine, Paris, France.
and
Dr. Laurence Zitvogel, UMR1015 équipe 1, Gustave Roussy Campus Cancer, 114, rue Edouard Vaillant, Villejuif, France.
Email: laurence.zitvogel@gustaveroussy.fr

Immunological Reviews, Volume 180, pages 165-174 (October 13, 2017)

AII.1 Authors' contributions

Because this project was an international collaborative effort, I will only comment on the contributions made by our research team.

Céline M. Laumont: prepared samples and analyzed data leading to Figure AII.5, contributed to writing.

Eric Bonneil: injected samples in the mass spectrometer and general discussion.

Pierre Thibault: general discussion.

Claude Perreault: analyzed data, discussed results and contributed to writing.

All authors edited and approved the final version of the manuscript.

AII.2 Abstract

Cancer cells are subjected to constant selection by the immune system, meaning that tumors that become clinically manifest have managed to subvert or hide from immunosurveillance. Immune control can be facilitated by induction of autophagy, as well as by polyploidization of cancer cells. While autophagy causes the release of ATP, a chemotactic signal for myeloid cells, polyploidization can trigger endoplasmic reticulum stress with consequent exposure of the “eat-me” signal calreticulin on the cell surface, thereby facilitating the transfer of tumor antigens into dendritic cells. Hence, both autophagy and polyploidization cause the emission of adjuvant signals that ultimately elicit immune control by CD8+ T lymphocytes. We investigated the possibility that autophagy and polyploidization might also affect the antigenicity of cancer cells by altering the immunopeptidome. Mass spectrometry led to the identification of pep- tides that were presented on major histocompatibility complex (MHC) class I molecules in an autophagy-dependent fashion or that were specifically exposed on the surface of polyploid cells, yet lost upon passage of such cells through immunocompetent (but not immunodeficient) mice. However, the preferential recognition of autophagy-competent and polyploid cells by the innate and cellular immune systems did not correlate with the preferential recognition of such peptides *in vivo*. Moreover, vaccination with such peptides was unable to elicit tumor growth-inhibitory responses *in vivo*. We conclude that autophagy and polyploidy increase the immunogenicity of cancer cells mostly by affecting their adjuvanticity rather than their antigenicity.

Keywords: autophagy, calreticulin, endoplasmic reticulum stress, hyperploidy, immunopeptidome

AII.3 Introduction

Ever more compelling evidence indicates that cancers are not just cell-autonomous diseases that arise due to the accumulation of genetic and epigenetic alterations in malignant cells. Full-blown malignancy only develops if cancer cells are ignored by the immune system or if they manage to actively subvert the control by innate and cognate immune effectors[1]. The idea of immunosurveillance was first supported by experiments showing that carcinogen- or oncogene-induced cancers develop more frequently in immunodeficient (ID) rodents and that cancer cells from ID mice were frequently rejected upon their transfer into histocompatible immunocompetent (IC) mice, yet lost their antigenic potential upon repeated passage through such IC host in a process called 'immunoselection'[2-4]. However, the concept of immunosurveillance only became fully acknowledged by clinical oncologists following accumulating evidence that (i) those cancers that are heavily infiltrated by CD8+ cytotoxic T cells (or other positive immune effectors such as CD4+ T cells, NK cells, and activated dendritic cells [DCs]) and with relatively few FOXP3+ regulatory T cells (or other immunosuppressive cell types including myeloid-derived suppressor cells) have a comparatively good prognosis[5,6]; (ii) chemotherapeutics and targeted agents including the tyrosine kinase inhibitor imatinib are particularly efficient if they elicit a local anticancer immune response[7-10]; and (iii) so-called immune checkpoint blockers (namely antibodies targeting CTLA-4, PD-1, or PD-L1) mediate an unprecedented broad activity against multiple distinct cancer types and occasionally even cure patients with metastatic melanoma[11,12].

According to the 'danger' hypothesis[13], immune responses are optimally elicited if the stimulus has two distinctive features. First, it is important that cancer cells (or cells infected by a pathogen) are antigenically different from their normal counterparts. Tumor-associated antigens (TAAs) can arise through two distinct mechanisms, namely mutation within the coding region of a protein causing the generation of a peptide that is presented by major histocompatibility complex (MHC) molecules ('neoantigens')[14-16] or ectopic expression of antigens that are usually only expressed during early embryogenesis, and in the placenta or testis of adults ('carcinoembryonic' and 'cancer

testis' antigens)[17]. Indeed, there is a correlation between the efficacy of immune checkpoint blockers and mutation load, meaning that cancers with a high level of mutations (such as melanoma, smoking-associated non-small-cell lung cancer, and microsatellite-instable colorectal carcinoma) are particularly amenable to immunotherapy[18,19]. However, the presence of pathogen-encoded antigens or TAAs is not entirely sufficient to elicit an immune response against infected or malignant cells. In addition, such cells must emit danger signals in the form of danger-associated molecular patterns (DAMPs) that act as adjuvants[20]. In the context of infection, such DAMPs can be referred to as microbe-associated molecular patterns (MAMPs)[21,22]. However, in the context of cancer, such DAMPs are endogenous molecules that appear on the surface of the tumor cells or are released into the pericellular space as cell death-associated molecules (CDAMs)[23]. Important CDAMs include chemotactic factors for the attraction of DCs (such as adenosine triphosphate, ATP, a ligand of purinergic receptors, and the protein annexin A1, ANXA1, a ligand of formyl peptide receptor-1)[8,24,25], the 'eat-me' signal calreticulin (CALR) that is exposed on the surface of stressed/dying cancer cells and facilitates the transfer of TAAs into DCs[26,27], the DCs maturation signal HMGB1 (high mobility group B1 protein, a ligand of Toll-like receptor-4)[28], and type-1 interferons that elicit a cascade of events resulting in the local recruitment of T cells[29]. Importantly, several successful chemotherapeutics are particularly efficient in stimulating the emission of CDAMs by cancer cells, thus causing immunogenic cell death[30-33]. Indeed, the therapeutic success of chemotherapies with anthracyclines and oxaliplatin against mammary and colorectal carcinomas, respectively, appears to be largely mediated by an anticancer immune response, both in animal models and in cancer patients[8,34,35].

As discussed above, it appears that immunogenicity essentially results from the combination of two factors, namely antigenicity and adjuvanticity. We have accumulated preclinical and clinical proof that two particular cellular processes, namely autophagy and hyperploidy, can favorably influence cancer immunosurveillance. Here, we will address the question to which extent these two processes may alter the antigenicity of cancer cells, thus stimulating their immune recognition.

Autophagy is a lysosomal bulk degradation process that is characterized by the formation of autophagosomes, which typically recruit proteins from the microtubule-associated proteins 1A/1B light chain 3 (best known as LC3) family as a biomarker of ongoing autophagic activity[36,37]. Breast cancer patients that lack LC3B-positive puncta in the cytoplasm of malignant cells (and hence are considered as autophagy- deficient) are characterized by a poor ratio of CD8+ over FOXP3+ T cells and have a poor prognosis with respect to progression-free and overall survival[37,38]. Moreover, cancer cells that have been engineered to lack essential components of the autophagic machinery (such as the genes Atg5 or Atg7) become refractory to the induction of anticancer immune responses by treatment with anthracyclines or oxaliplatin, thus forming chemotherapy-resistant tumors[25,39,40]. The defective immunogenicity of autophagy-deficient cancer cells has been linked to a reduction of ATP release from dying cancer cells, presumably because autophagy is required for lysosomal ATP secretion[41]. Accordingly, measures designed to increase extracellular ATP concentrations (such as the inhibition of ATP- degrading enzymes) can re-establish the anticancer immune response[25]. Nonetheless, there is ample evidence that autophagy also affects the immunopeptidome, i.e., the repertoire of the peptides presented by MHC class I molecules at the cell surface, by affecting protein degradation and trafficking within the cell[42]. Hence, we addressed the question as to whether antigenic peptides whose presentation depends on autophagy might contribute to autophagy-dependent immunogenicity.

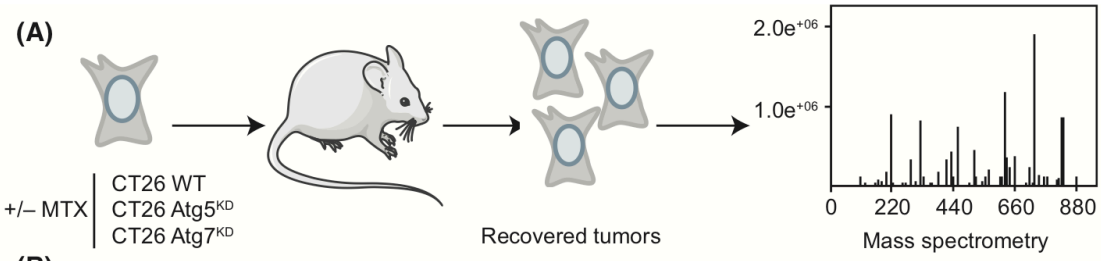
An increase in ploidy, and in particular whole-genome duplications leading to tetraploidy, strongly increases the immunogenicity of cancer cells[7]. Indeed, cancer cells that have been artificially tetraploidized *in vitro* usually do not grow upon transfer into histocompatible IC mice (although they readily take on ID hosts lacking T cells) [7,43,44]. However, if they grow in IC mice, which usually occurs after a delay, the resulting tumor cells are characterized by a reduction in ploidy as a sign of immunoselection[7,43]. In accord with this observation, there is accumulating evidence that somatic copy number alterations (that possibly reflect prior tetraploidization events) in clinically manifest human cancers correlate with poor local immunosurveillance by cytotoxic T cells[45].

Thus, carcinogenesis might involve a transient phase of tetraploidization that is only tolerated by the immune system, if the cancer cells manage to reduce their DNA content and re-acquire a sub-tetraploid (and sometimes pseudo-diploid) state leading to their escape from immune control[7,46]. We have found that tetraploidy causes an endoplasmic reticulum (ER) stress response with hyperphosphorylation of eIF2α and consequent exposure of the ER protein CALR on the cell surface[47]. Knockdown of a protein required for CALR exposure[48,49], ERp57 (also known as protein disulfide-isomerase A3, PDIA3), is sufficient to facilitate the growth of tetraploid cells in IC mice[7]. However, phosphorylation of eIF2α may be expected to affect the translation of multiple proteins and hence to impact on the immunopeptidome[50,51]. Moreover, CALR is part of the molecular machinery that loads antigenic peptides into the nascent MHC class I molecules in the ER[52,53]. Therefore, we decided to investigate the possible impact of tetraploidy on the immunopeptidome.

AII.3.1 Contribution of autophagy to immunogenicity

Autophagy-deficient cancer cells may escape from chemotherapy-induced immunosurveillance in experimental models. Indeed, mouse cancers that usually elicit a tumor growth-reducing immune response, after radiotherapy or chemotherapy with anthracyclines or oxaliplatin, fail to do so if they are manipulated to become autophagy deficient[25,54-57]. Autophagy deficiency is also associated with a poor local anticancer immunosurveillance (with an unfavorable ratio of CD8+ cytotoxic T lymphocytes to FOXP3+ regulatory T cells in the local immune infiltrate) in human breast cancer[58]. Accordingly, autophagy deficiency in breast cancer cells is associated with poor prognosis[38]. Conversely, autophagy induction by starvation or by pharmacological inducers can stimulate anticancer immunosurveillance in mouse models[39,40,59]. Mechanistically, it appears that autophagy is required for lysosomal ATP secretion, which contributes to elevating extracellular ATP concentrations in the context of caspase-dependent apoptotic cell death[41,60] Extracellular ATP then acts as a chemotactic factor on purinergic receptors (such as P2Y2 receptors) to allow for the recruitment of myeloid cells including DC precursors into the tumor bed, hence starting of a cascade that culminates in the presentation of TAAs to specific T cells[24,61].

We wondered whether autophagy might also affect the antigenic properties of cancer cells, based on the fact that this process may influence the immunopeptidome of antigen-donor cells (such as virus-infected or cancer cells) and antigen-presenting cells[42,55,62,63]. To study this possibility, we took advantage of CT26 cells, a colon carcinoma cell line derived from BALB/c mice (H-2-K^d, H-2-D^d, H-2-L^d, I-A^d, I-E^d) that had been rendered autophagy-deficient by knocking down the essential autophagy genes Atg5 or Atg7[25]. Tumors were generated by subcutaneously injecting wildtype (WT), Atg5KD, or Atg7KD CT26 cells into mice that were treated or not with the anthracycline mitoxantrone (MTX). Two days later, when WT cells (but not Atg5KD or Atg7KD) manifested autophagy in response to chemotherapy[25], the tumors were excised and subjected to immunopeptidome analyses[64] (**Figure AII.1A**). Peptides that were induced by chemotherapy in WT tumors but not in autophagy-deficient tumors were identified (**Figure AII.1B**), synthesized, and subsequently analyzed for their immunogenic properties. CT26 cells were killed by MTX *in vitro* and then injected subcutaneously into BALB/c mice following a protocol that reproducibly induces an immune response that protects mice against re-challenge with live CT26 cells and hence induces a vaccination effect[28,31,34]. This protocol induced a significant immune response against a known CT26 TAA, namely the AH1 peptide, as measured by an ELISPOT assay in which splenocytes were evaluated for AH1 peptide-induced interferon-γ (IFNγ) production (**Figure AII.2A**). However, none of the 40 peptides that were identified as MTX-induced and autophagy-dependent (**Figure AII.1B**) could stimulate a significant response when they were tested for their capacity to stimulate IFNγ production (**Figure AII.2B**). Similarly, attempts to protect mice against the growth of CT26 cells by vaccinating with these peptides failed (data not shown). In conclusion, it appears that autophagy-dependent changes in the immunopeptidome have no major impact on the immunogenicity of these cancer cells.



(B)

Peptide ID	Peptide sequence	Gene symbol	Peptide ID	Peptide sequence	Gene symbol
1	VYSQIPAAVKL	*Ms4a3*	21	SFVNCLTHL	*Xpo4*
2	IGPWIQTKM	*Actn1*	22	VGPHLRFEI	*Wbp7*
3	TYNNIFQYHQL	*Mcm8*	23	MSPQKFWGL	*Crbn*
4	AGPMAAFKW	*Glyr1*	24	SPWRGFII	*Csf1r*
5	IGAIRFFAF	*Gnpat*	25	FFLRMKCTL	*Hif1a*
6	AFLQSLRTL	*Sapcd2*	26	IFLHVLSEL	*Ttc3*
7	VYIITKPEL	*Nprl2*	27	TGPDYYSQL	*Mll2*
8	FYMFGWTHV	*Cds2*	28	SPYRVMLF	*Casp8*
9	SPVRDWLF	*Msln*	29	AYQSVGAEL	*Aldh1l2*
10	FGPCILNRL	*N/A*	30	HGAAAIFKL	*Myof*
11	LGPAHWLVF	*Stab1*	31	AGPLRVLCV	*Mmp2*
12	GGPVKIGQLV	*Zbp1*	32	KYSFMESLL	*Dhx32*
13	MGPKKVGMLW	*Eif4g1*	33	VPQSNFLDLF	*Dab2*
14	SGPAATQDF	*Abl1*	34	AGPLMADFI	*Adamtsl3*
15	SGAHRIDGL	*Bcor*	35	YLKILTGL	*Hace1*
16	FPYRLVGL	*Gramd4*	36	MLRGPGPGRLLLL	*Fn1*
17	KYCISVRTL	*Iqsec1*	37	LPSIIPWGL	*Lrrk1*
18	HGMTRFVGI	*Mrps11*	38	GKDLKPSTPRSK	*Npm1*
19	AYILTRFAL	*Ndrg1*	39	KGFLKNFLI	*Acly*
20	QGFFRSISL	*Mtor*	40	KLLRLLSLI	*Huwe1*

Figure AII.1 | Identification of MHC class I-associated peptides that are mitoxantrone (MTX)-induced and autophagy-dependent. (**A**) BALB/c mice were injected s.c. (near the thigh, 8 $\times 10^5$ cells in 100 µL of PBS) with wildtype (WT) colon carcinoma CT26 cells or CT26 cells stably expressing a shRNA specific for the essential autophagy genes Atg5 (Atg5KD) or Atg7 (Atg7KD). Before the tumor surface reached 35-45 mm2, mice (5-8 per group) received either 1 mM MTX intraperitoneally in 200 µL of PBS or an equivalent volume of PBS as a control. Two days after the treatment, tumors were carefully removed and directly shock-frozen in liquid nitrogen. At least five tumors per group were pooled to obtain 1 g of tissue. (**B**) The six different conditions (WT-MTX, WT-PBS, Atg5-MTX, Atg5-PBS, Atg7-WTX, Atg7-PBS) were each immunoprecipitated using 20-8-4-S and 34-4-20-S for capture of H-2-K^d and H-2-D^d, respectively. The 12 eluates were further analyzed by label-free LC-MS using up to 10 replicates each providing quantitative peptidomics data for CT26. Peptides presented exclusively in condition WT-MTX were used as basis for selection of 40 autophagy-associated peptides.

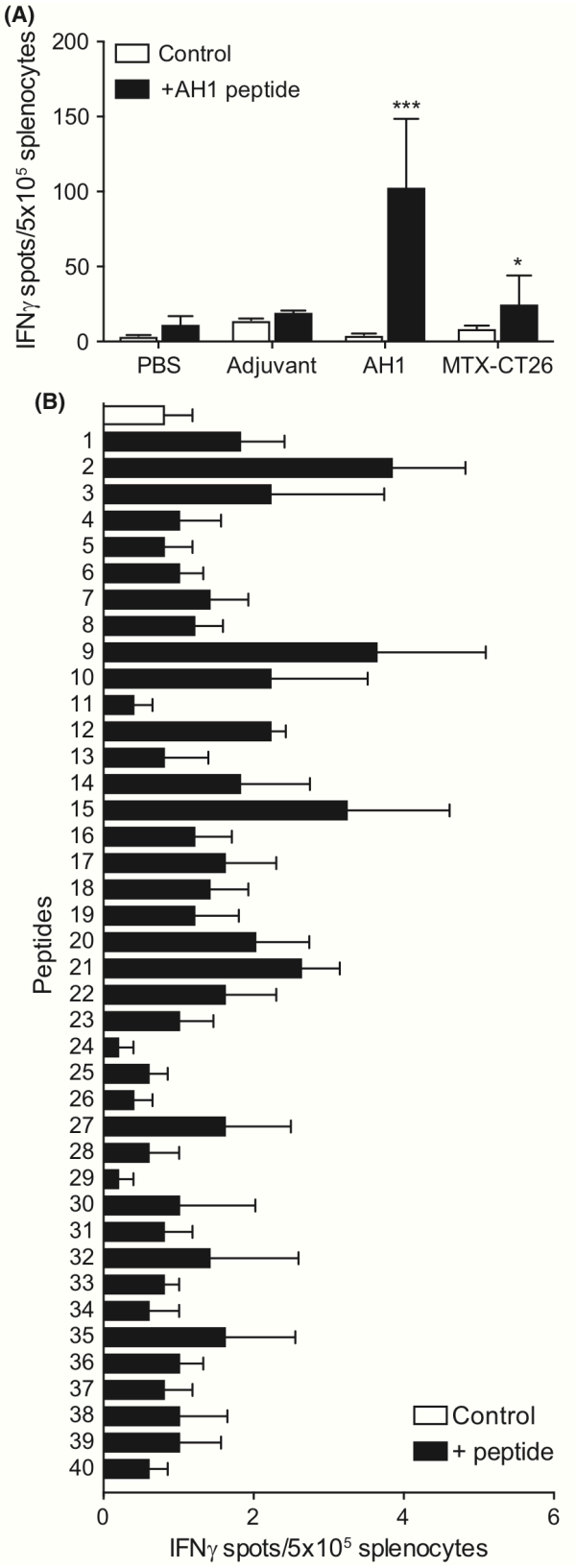


Figure AII.2 | Comparison of the immunogenic potential of different peptides from CT26 cells. (**A**) IFNγ ELISPOT analysis of splenocytes from BALB/c mice that were immunized subcutaneously at the footpad with PBS alone (n=2), adjuvant composed of 50 µg poly(I:C) and 25 µg CpG (n=2), 150 µg AH1 peptide+adjuvant (AH1, n=6) or 3×10^5 CT26 cells that had been cultured for 40 hours with 12 µM mitoxantrone (MTX-CT26, n=6). These vaccinations were performed 21 and 7 days before erythrocyte-free splenocytes were re-stimulated or not with AH1 peptide, and the frequency of interferon-γ (IFNγ)-producing cells was measured by ELISPOT. (**B**) IFNγ ELISPOT analysis of splenocytes from BALB/c mice (n=5) that were immunized with MTX-CT26 cells (as in **A**) and then were re-stimulated in vitro with 40 different MTX-induced peptides (listed in Fig. 1B). (**A**) Results are representative of two independent experiments. Error bars indicate SD (**A**) or SEM (**B**). Samples were compared using two-tailed paired Student's t-test. *P < .05, ***P < .001, compared to non-stimulated splenocytes. Note that the scales in **Figure AII.2A, B** are different. MTX, mitoxantrone.

AII.3.2 Immunosurveillance against hyperploid cancer cells – Impact on calreticulin exposure and the immunopeptidome

Hyperploid cancer cells can be recognized by the immune system in such a way that their growth is suppressed or delayed. This applies to multiple mouse models, namely CT26 colorectal cancer cells growing in BALB/c mice as well as MCA205 fibrosarcoma growing in C57BL/6 mice (H-2-K^b, H-2-D^b, H-2-L^{Null}, I-A^b, I-E^{Null})[7]. Similarly, hyperploid (H) EL4 lymphoma cells, which were generated by the transient exposure of cells to nocodazole followed by cytofluorometric purification of cells with an abnormally high (>4n) DNA content[47], are delayed in their growth when injected into IC C57BL/6 mice as compared with ID NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice (best known as NOD scid gamma or NSG mice). In contrast, parental (P) EL4 lymphoma cells proliferated equally in IC and ID mice (**Figure AII.3A-D**). Hyperploid EL4 cells were passaged through IC and ID mice and recovered from the tumors to determine their characteristics (**Figure AII.3E**). In accord with previous results obtained with other cancer cell types[7], immunoselection (i.e., passage through IC mice) yielded cells that exposed less CALR on their cell surface (**Figure AII.3F**) and exhibited a partial reduction in their ploidy (**Figure AII.3G**). In contrast, *in vivo* passage without immunoselection (through ID mice) yielded cells with high CALR exposure and high ploidy (**Figure AII.3F,G**). These findings confirm and extend our prior observation that hyperploidy is functionally linked to CALR exposure and that both characteristics are subjected to negative selection by an intact immune system[7,43,44]. Indeed, depletion of ERp57, which is required for CALR exposure[48,49], abolishes the immunoselection in favor of a reduced ploidy[7]. Moreover, stable transfection of cells with a CALR variant that is tethered to the surface of the plasma membrane is sufficient to preclude the growth of cancers in vivo[7,65].

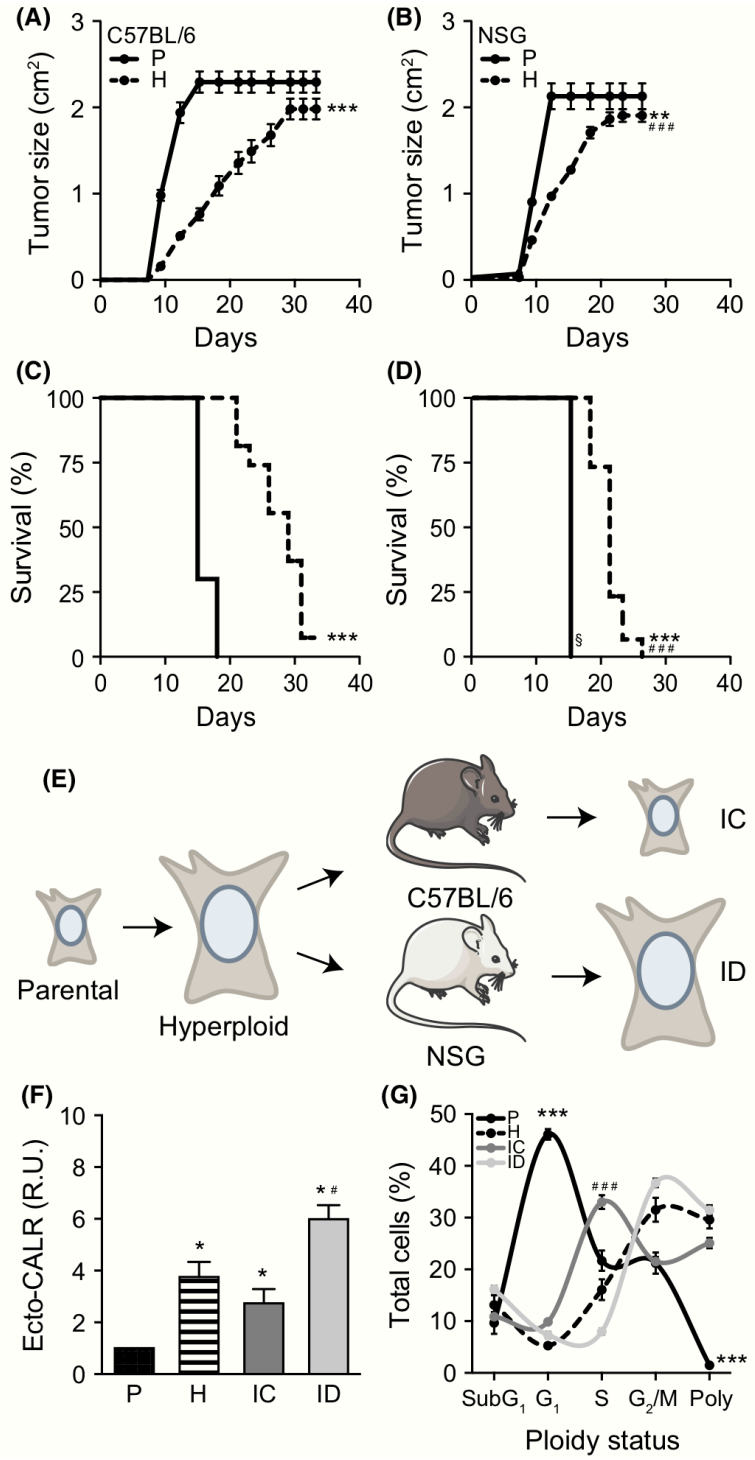


Figure AII.3 | Immunosurveillance of EL4 hyperploid cells. The parental (P) murine lymphoma EL4 cell line or two hyperploid (H) EL4 clones were inoculated into immunocompetent (C57BL/6) (**A,C**) or immunodeficient (NSG) (**B,D**) mice, and tumor growth (**A,B**) and incidence (**C,D**) were monitored. The parental cells were injected into 10 mice per group, while hyperploid clones were injected into 30 mice per group. Tumor growth curves were

analyzed using Wald's test, whereas tumor incidence was compared by log-rank test. **P < .01 H vs P inoculated into NSG mice; ***P < .001 H vs the corresponding P; ###P < .001 as compared to H inoculated into C57BL/6; §P < .05 as compared to P inoculated into C57BL/6. (**E**) Hyperploid (H) clones immunoselected in C57BL/6 mice (IC, n=18) or grown *in vivo* without immunoselection in NSG mice (ID, n=20) were recovered before the tumor surface reached 3 cm^2 and then assessed for CALR exposure by immunofluorescence staining of viable (propidium iodide-excluding) cells (**F**) or Hoechst 33342 staining to determine DNA content by cytofluorometry (**G**) Parental (P) EL4 cells and hyperploid (H) clones cultured in vitro were used as controls. Results are means ± SEM. In (**F**) two-tailed unpaired Student's t-test was used for statistical comparisons. *P < .05 as compared to parental cell line; #P < .05 as compared to hyperploid clones. In (**G**) the two-way ANOVA test was used for statistical comparisons. ***P < .001 as compared to all H, IC, and ID; ###P < .001 as compared to all P, H, and ID.

To investigate additional hyperploidy-associated changes, as well as the impact of immunoselection, we performed analyses of the transcriptome and immunopeptidome of parental and hyperploid EL4 cells, before and after *in vivo* passage of the latter through IC and ID mice. Principal component analyses of the transcriptome correctly differentiated the microarray results from 4 parental cells, 8 hyperploid clones cultured in vitro, 7 tumors obtained from the *in vivo* passage of hyperploid cells through ID mice, and 10 tumors obtained from passage of such cells through IC animals (**Figure AII.4A**). Importantly, in the second principal component (16% of explained variance), the transcriptome of ID-passaged tumor closely resembled that of their hyperploid precursors, while that of IC-passaged tumors diverged and approached that of the parental EL4 cells, a finding that is suggestive of an immunoselection-associated reversal of hyperploidy-associated traits (**Figure AII.4B**). Thus, the difference between parental vs hyperploid cells cultured *in vitro* in the expression of individual genes (black points in **Figure AII.4C**) is far more important than that between hyperploid cells passaged through ID vs IC mice (red points in **Figure AII.4C**).

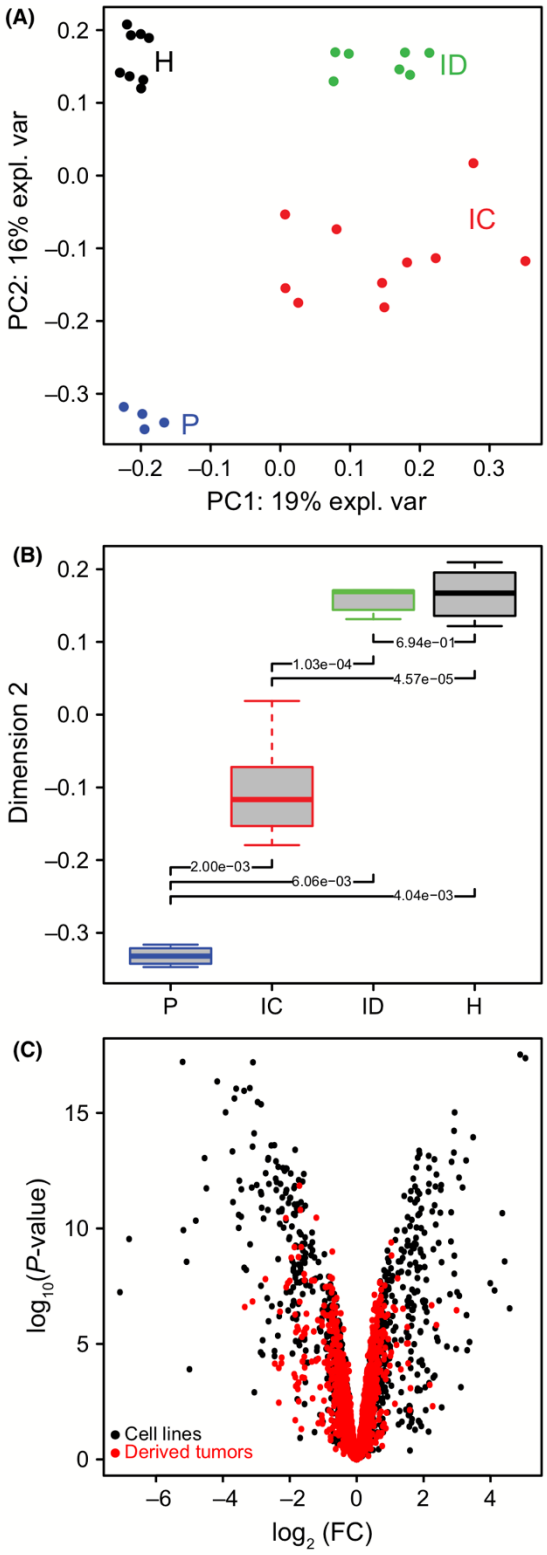
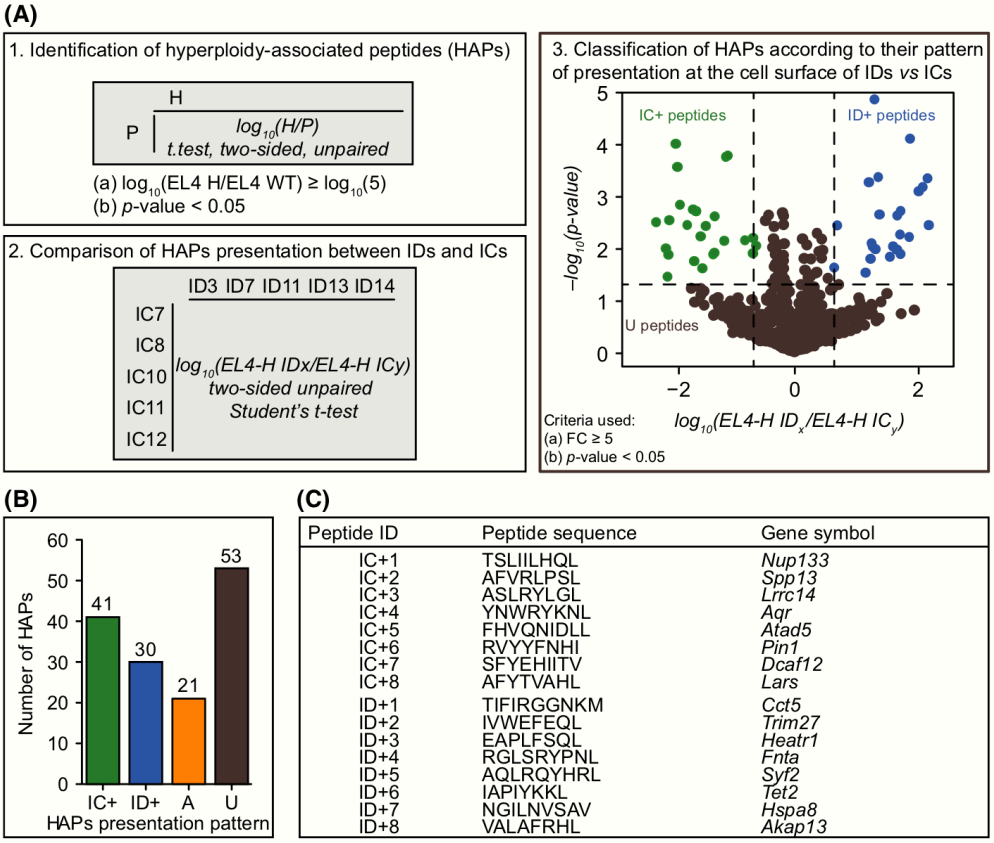


Figure AII.4 | Modulation of gene expression by ploidy status. (**A**) Whole-gene expression analysis from parental (P) EL4 cells, hyperploid clones (H), and their derived tumors from immunocompetent (IC) or immunodeficient (ID) mice. Representation of the iPCA of normalized data obtained by means of the Agilent SurePrint G3 Mouse GE 8×60K Microarray (AMADID 28005). PC1 and PC2 are the first and second principal component explaining 19% and 16% of the variability observed between transcriptomic profiles, respectively. (**B**) Boxplot representation of PC2-coordinate distribution for each group. Samples were compared using a non-parametric Wilcoxon statistical test. P values are indicated for each comparison. (**C**) Volcanoplot (log-fold change vs $\log_{10}$ of P-value) of differential expression. Samples were compared using moderated t-statistics ("eBayes" of the R limma package) for parental vs hyperploid cells (black) and immunocompetent (IC)-derived tumors vs ID-derived tumors (red).

Next, we analyzed the immunopeptidome of those cells to identify peptides that would be presented by hyperploid (but not parental) EL4 cells and that would be subjected to immunoselection, meaning that they would persist on hyperploid cells passaged through ID but not IC mice. Hence, we first identified the hyperploidy-associated peptides by mass spectrometry[66] and then determined their persistence or disappearance in 5 ID and 5 IC lines, respectively, using stringent statistical filters. Volcano plots were performed for all 145 peptides significantly associated with in vitro hyperploidization to determine their association with *in vivo* passage through IC vs ID mice (**Figure AII.5A**). This led to the classification of peptides into four classes, those that are significantly associated with IC or ID, those that are ambiguous (significantly associated with IC and ID across different comparisons) and those that do not change significantly ($P \geq .05$ or $FC < 5$) (**Figure AII.5A,B**). The eight peptides that were most strongly associated with IC and ID passage were listed (**Figure AII.5C**) and synthesized. Several among these peptides induced strong T-cell responses (determined by ELISPOT IFNγ quantitation on splenocytes) upon their inoculation into C57BL/6 mice in the presence of adjuvant (poly I:C and CpG). These responses were quantitatively as strong as those observed after vaccination with the protein ovalbumin (OVA, which contains class I and class II peptides) (**Figure AII.6A**) and were significant ($P < .05$, unpaired Student's t-test) for 5 out of 7 IC-associated (**Figure AII.6B**) and 7 out of 8 ID-associated peptides (**Figure AII.6C**). However, only vaccination with OVA conferred protection against the growth of an OVA-expressing EL4 cell line (**Figure AII.6D**). ID- and IC-associated peptides were equally inefficient in eliciting a growth-inhibitory immune response against hyperploid EL4 cells injected into mice individually (**Figure AII.6E,F**) or as a mixture (data not shown). In conclusion, it appears that ID-associated peptides (which appear on the surface of hyperploid cells and are underrepresented upon immunoselection of such cells) are unable to elicit a protective immune response against hyperploid EL4 clones.



Peptide ID	Peptide sequence	Gene symbol
IC+1	TSLIILHQL	*Nup133*
IC+2	AFVRLPSL	*Spp13*
IC+3	ASLRYLGL	*Lrrc14*
IC+4	YNWRYKNL	*Aqr*
IC+5	FHVQNIDLL	*Atad5*
IC+6	RVYYFNHI	*Pin1*
IC+7	SFYEHIITV	*Dcaf12*
IC+8	AFYTVAHL	*Lars*
ID+1	TIFIRGGNKM	*Cct5*
ID+2	IVWEFEQL	*Trim27*
ID+3	EAPLFSQL	*Heatr1*
ID+4	RGLSRYPNL	*Fnta*
ID+5	AQLRQYHRL	*Syf2*
ID+6	IAPIYKKL	*Tet2*
ID+7	NGILNVSAV	*Hspa8*
ID+8	VALAFRHL	*Akap13*

Figure AII.5 | Ploidy status modulates the immunopeptidome. (**A**) 1,Workflow used to identify and compare the abundance of hyperploidy-associated peptides (HAPs) by quantitative high-throughput mass spectrometry. MHC I-associated peptides were distinguished from contaminants using four criteria: (i) length of 8–15 amino acids, (ii) peaks score ≥20, (iii) binding score for H-2-K^b or H-2-D_b ≤1000 nM (as predicted by NetMHCcons v.1.0) and (iv) reproducible detection (in at least three replicates of 1 out of the 12 samples). 2, The immunopeptidome of a hyperploid (H) EL4 clone and EL4 parental (P) cell line was compared to identify HAPs significantly more abundant at the cell surface as defined by a fold change (FC) ≥5 and a P-value <.05 (P < .05, two-sided unpaired Student's t-test). The abundance of those HAP was then compared between hyperploid (EL4-H) clone after passage into immunodeficient (NSG, IDs, n=5) vs immunocompetent (C57BL/6, ICs, n=5) mice. Using the same FC and P-value threshold than for the EL4-H vs EL4-P comparison, those HAPs were classified by their pattern of presentation into IC+ peptides, significantly more presented in at least one IC cell line; ID+ peptides, significantly more presented in at least one ID cell line; ambiguous (A) peptides, significantly presented in one IC and one ID cell line; and unchanged (U) peptides, not differentially presented between ID and IC cell lines. (**B**) Classification of HAPs according to their detection status. IC+, overexpressed in immunocompetent (IC) cell line; ID+, overexpressed in immunodeficient (ID) cell line; A, ambiguous; and U, unchanged peptides. (**C**) Best 8 IC+ HAPs potential immunogenic candidates (first eight peptides) and eight ID+ HAPs peptides controls (last eight peptide), as defined by a FC ≥5 in EL4-H (ICs-IDs) ≥6, when then selected for *in vivo* vaccination experiments.

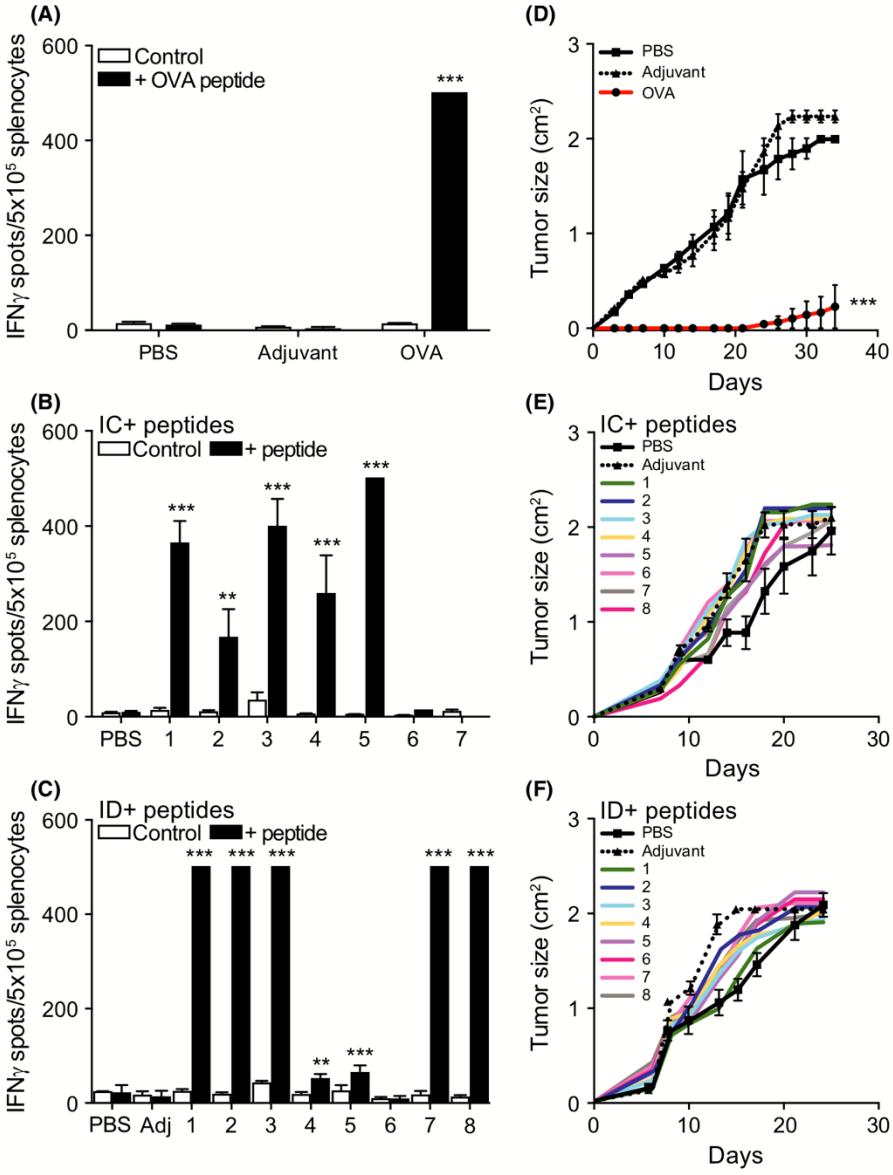


Figure AII.6 | Failure of hyperploidy-associated peptides to induce a protective immune response against hyperploid EL4 cells. C57BL/6 mice were injected in the footpad with PBS, adjuvant (poly(I:C) plus CpG, Adj) alone or in combination with ovalbumin (OVA) protein in (**A**), eight peptides overexpressed in immunocompetent mice (IC+ peptides, listed in Figure 5C) in (**B**), or eight peptides overexpressed in immunodeficient mice (ID+ peptides, listed in Figure 5C) in (**C**), 21 and 7 days before spleens were harvested and IFNγ production from re-stimulated splenocytes was analyzed by ELISPOT. At least two spleens per condition were used in duplicate (n=2 for PBS and adjuvant groups in A, and all groups in (**C**); n=4 all groups in (**B**), n=6 OVA group in (**A**)). Error bars indicate SD (**A,C**) or SEM (**B**). Samples were compared using a non-parametric two-way ANOVA test, **P < .01, ***P < .001, as compared with non-stimulated condition. In addition, C57BL/6 mice were vaccinated with ovalbumin (OVA, n=10) protein as positive control in D, IC+ peptides (n=5 for P1, P3, P5, P7, P8 n=5; n=6

for P2; n=3 for P6) in (**E**), or ID+ peptides (n=9 for each peptide) in (**F**) on days 21 and 7 days before inoculation of the EL4 OVA-expressing cell line (EG7, (**D**)) or hyperploid clones (**E,F**). As a negative control, mice were sham-vaccinated with PBS: n=5 in (**D**), n=10 in (**E**), n=9 in (**F**), n=5 in (**A,B**) and n=9 in (**C**). Error bars indicate SEM. Overall tumor growth curves were analyzed by means of the Wald's test. ***P < .001 as compared to PBS.

AII.4 Concluding remarks

Here, we examined the possibilities that two cellular phenomena that increase tumor cell immunogenicity, namely (i) autophagy and (ii) polyploidy, might increase tumor cell antigenicity via an effect on the MHC class I-restricted immunopeptidome. Although sophisticated mass spectrometry methods identified MHC class I-associated peptides exposed only on autophagy-competent (but not on autophagy-deficient) cancer cells stimulated with the anthracycline MTX, such peptides were barely detected by T cells from mice that had been vaccinated with MTX-treated cancer cells. Thus, the maximal frequency of T cells producing IFNγ in response to such autophagy-dependent peptides was one order of magnitude less frequent than that induced by a dominant TAA. Moreover, vaccination with such peptides was unable to elicit a protective anticancer immune response. In a similar fashion, we identified peptides that specifically appear on the surface of lymphoma cells that had been polyploidized *in vitro,* persisted on such cells upon their passage through ID mice, yet disappeared upon immunoselection of such cells in IC hosts. Although most of these peptides were immunogenic *in vivo*, meaning that they elicited high frequencies of antigen-specific IFNγ-producing T cells, vaccination with such peptides again failed to confer protection against the growth of hyperploid tumor cells. As a caveat, this experimental failure might be attributed to the fact that the hyperploidy-associated peptides chosen here were not neoantigens (i.e., derived from non-synonymous mutations) or that the vaccination protocols might have been suboptimal. For instance, we only vaccinated with class I-restricted peptides in the absence of class II-restricted 'helper' peptides. However, it should be noted that immunization with the same batch of adjuvant using a tumor-relevant antigen (such as OVA in the case of OVA-expressing EG7 cells, including class I- and class II-restricted epitopes) did confer protection and that the vaccination with peptides was highly efficient in eliciting specific IFNγ responses. Thus, the failure of stress induced class I-restricted peptides alone to induce protection could also be the lack of concurrent CD4 responses. As a further caveat, it is possible that the hyperploid clones analyzed here might have accumulated variants during *in vitro* passage, rendering them heterogeneous with respect to their antigenic characteristics.

In this case, only a fraction of the tumor cells generated *in vitro* would be eliminated following a specific immune response. However, we did not detect a single case in which vaccination with a peptide (or a pool of peptides) would have caused an even partial delay in tumor growth, again arguing against this possibility.

In conclusion, it appears that autophagy and polyploidy increase the immunogenicity of cancer cells mostly—if not exclusively—through an effect on adjuvanticity, at least in the models that we have studied thus far. This interpretation should spur a renewed interest in studying the precise nature of the immunoadjuvant signals. Genomic and transcriptomic methods can be useful for the identification of neoantigens, yet are not suitable for studying adjuvant signals. Hence, the repertoire of technologies necessary for therapy-relevant cancer immunogenicity must be enlarged to methods that adequately detect the potential of tumor cells to emit adjuvant signals.

AII.5 Acknowledgements

Animal experiments were in compliance with the EU Directive 63/2010 and were approved by the Ethical Committee of Gustave Roussy (2016_062_ex047_5251 and 2017032916211041v3 pro- tocols, Villejuif, France) (CEEA IRCIV/IGR no. 26, registered at the French Ministry of Research). GK is supported by the Ligue contre le Cancer (équipe labelisée); Agence National de la Recherche (ANR) – Projets blancs; ANR under the frame of E-Rare-2, the ERA-Net for Research on Rare Diseases; Fondation ARC pour la Recherche sur le Cancer; Cancéropôle Ile-de-France; Institut National du Cancer (INCa); Institut Universitaire de France; Fondation pour la Recherche Médicale (FRM); the European Commission (ArtForce); the European Research Council (ERC); the LeDucq Foundation; the LabEx Immuno- Oncology; the RHU Torino Lumière, the SIRIC Stratified Oncology Cell DNA Repair and Tumor Immune Elimination (SOCRATE); the SIRIC Cancer Research and Personalized Medicine (CARPEM); and the Paris Alliance of Cancer Research Institutes (PACRI). LS is sup- ported by the Fondation ARC pour la Recherche sur le Cancer (ARC, PJA20151203519). AS is supported by Associazione Italiana per la

Ricerca contro il Cancro (AIRC, Start-Up 2016 #18418) and Ministero Italiano della Salute (RF_ GR-2013-02357273). Taxe d'apprentissage TA PAGA_24 2013 to perform whole-genome expression arrays. NB owns a PhD fellowship from Fondation ARC pour la Recherche sur le Cancer, and JH owns a PhD fellowship from Fondation Philanthropia.

AII.6 Conflict of interest

The authors declare no conflict of interest.

AII.7 References

1. Palucka, A.K. and L.M. Coussens, *The Basis of Oncoimmunology.* Cell, 2016. **164**(6): p. 1233-1247.
2. Vesely, M.D., et al., *Natural innate and adaptive immunity to cancer.* Annu Rev Immunol, 2011. **29**: p. 235-71.
3. Schreiber, R.D., L.J. Old, and M.J. Smyth, *Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion.* Science, 2011. **331**(6024): p. 1565-70.
4. Zitvogel, L., A. Tesniere, and G. Kroemer, *Cancer despite immunosurveillance: immunoselection and immunosubversion.* Nat Rev Immunol, 2006. **6**(10): p. 715-27.
5. Fridman, W.H., et al., *The immune contexture in human tumours: impact on clinical outcome.* Nat Rev Cancer, 2012. **12**(4): p. 298-306.
6. Becht, E., et al., *Immune Contexture, Immunoscore, and Malignant Cell Molecular Subgroups for Prognostic and Theranostic Classifications of Cancers.* Adv Immunol, 2016. **130**: p. 95-190.
7. Senovilla, L., et al., *An immunosurveillance mechanism controls cancer cell ploidy.* Science, 2012. **337**(6102): p. 1678-84.
8. Vacchelli, E., et al., *Chemotherapy-induced antitumor immunity requires formyl peptide receptor 1.* Science, 2015. **350**(6263): p. 972-8.
9. Rusakiewicz, S., et al., *Immune infiltrates are prognostic factors in localized gastrointestinal stromal tumors.* Cancer Res, 2013. **73**(12): p. 3499-510.
10. Zitvogel, L., et al., *Immunological off-target effects of imatinib.* Nat Rev Clin Oncol, 2016. **13**(7): p. 431-46.
11. Sharma, P. and J.P. Allison, *Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential.* Cell, 2015. **161**(2): p. 205-14.
12. Topalian, S.L., et al., *Immunotherapy: The path to win the war on cancer?* Cell, 2015. **161**(2): p. 185-6.

13. Matzinger, P., *The danger model: a renewed sense of self.* Science, 2002. **296**(5566): p. 301-5.
14. Gubin, M.M., et al., *Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens.* Nature, 2014. **515**(7528): p. 577-81.
15. Schumacher, T.N. and R.D. Schreiber, *Neoantigens in cancer immunotherapy.* Science, 2015. **348**(6230): p. 69-74.
16. Yadav, M., et al., *Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.* Nature, 2014. **515**(7528): p. 572-6.
17. Woo, S.R., L. Corrales, and T.F. Gajewski, *Innate immune recognition of cancer.* Annu Rev Immunol, 2015. **33**: p. 445-74.
18. Snyder, A., et al., *Genetic basis for clinical response to CTLA-4 blockade in melanoma.* N Engl J Med, 2014. **371**(23): p. 2189-99.
19. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade.* Science, 2016.
20. Bianchi, M.E., *DAMPs, PAMPs and alarmins: all we need to know about danger.* J Leukoc Biol, 2007. **81**(1): p. 1-5.
21. Mackey, D. and A.J. McFall, *MAMPs and MIMPs: proposed classifications for inducers of innate immunity.* Mol Microbiol, 2006. **61**(6): p. 1365-71.
22. Zitvogel, L., et al., *Microbiome and Anticancer Immunosurveillance.* Cell, 2016. **165**(2): p. 276-87.
23. Zitvogel, L., O. Kepp, and G. Kroemer, *Decoding cell death signals in inflammation and immunity.* Cell, 2010. **140**(6): p. 798-804.
24. Elliott, M.R., et al., *Nucleotides released by apoptotic cells act as a find-me signal to promote phagocytic clearance.* Nature, 2009. **461**(7261): p. 282-6.
25. Michaud, M., et al., *Autophagy-dependent anticancer immune responses induced by chemotherapeutic agents in mice.* Science, 2011. **334**(6062): p. 1573-7.
26. Gardai, S.J., et al., *Cell-surface calreticulin initiates clearance of viable or apoptotic cells through trans-activation of LRP on the phagocyte.* Cell, 2005. **123**(2): p. 321-34.

27. Obeid, M., et al., *Calreticulin exposure dictates the immunogenicity of cancer cell death.* Nat Med, 2007. **13**(1): p. 54-61.
28. Apetoh, L., et al., *Toll-like receptor 4-dependent contribution of the immune system to anticancer chemotherapy and radiotherapy.* Nat Med, 2007. **13**(9): p. 1050-9.
29. Sistigu, A., et al., *Cancer cell-autonomous contribution of type I interferon signaling to the efficacy of chemotherapy.* Nat Med, 2014. **20**(11): p. 1301-9.
30. Krysko, D.V., et al., *Immunogenic cell death and DAMPs in cancer therapy.* Nat Rev Cancer, 2012. **12**(12): p. 860-75.
31. Kroemer, G., et al., *Immunogenic cell death in cancer therapy.* Annu Rev Immunol, 2013. **31**: p. 51-72.
32. Kepp, O., et al., *Consensus guidelines for the detection of immunogenic cell death.* Oncoimmunology, 2014. **3**(9): p. e955691.
33. Galluzzi, L., et al., *Immunogenic cell death in cancer and infectious disease.* Nat Rev Immunol, 2017. **17**(2): p. 97-111.
34. Casares, N., et al., *Caspase-dependent immunogenicity of doxorubicin-induced tumor cell death.* J Exp Med, 2005. **202**(12): p. 1691-701.
35. Tesniere, A., et al., *Immunogenic death of colon cancer cells treated with oxaliplatin.* Oncogene, 2010. **29**(4): p. 482-91.
36. Kabeya, Y., et al., *LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing.* EMBO J, 2000. **19**(21): p. 5720-8.
37. Choi, A.M., S.W. Ryter, and B. Levine, *Autophagy in human health and disease.* N Engl J Med, 2013. **368**(19): p. 1845-6.
38. Ladoire, S., et al., *The presence of LC3B puncta and HMGB1 expression in malignant cells correlate with the immune infiltrate in breast cancer.* Autophagy, 2016. **12**(5): p. 864-75.
39. Pietrocola, F., et al., *Caloric Restriction Mimetics Enhance Anticancer Immunosurveillance.* Cancer Cell, 2016. **30**(1): p. 147-160.
40. Pietrocola, F., et al., *Autophagy induction for the treatment of cancer.* Autophagy, 2016. **12**(10): p. 1962-1964.

41. Martins, I., et al., *Molecular mechanisms of ATP secretion during immunogenic cell death.* Cell Death Differ, 2014. **21**(1): p. 79-91.
42. Munz, C., *Autophagy proteins in antigen processing for presentation on MHC molecules.* Immunol Rev, 2016. **272**(1): p. 17-27.
43. Bloy, N., et al., *Morphometric analysis of immunoselection against hyperploid cancer cells.* Oncotarget, 2015. **6**(38): p. 41204-15.
44. Boileve, A., et al., *Immunosurveillance against tetraploidization-induced colon tumorigenesis.* Cell Cycle, 2013. **12**(3): p. 473-9.
45. Davoli, T., et al., *Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy.* Science, 2017. **355**(6322).
46. Vitale, I., et al., *Mitotic catastrophe: a mechanism for avoiding genomic instability.* Nat Rev Mol Cell Biol, 2011. **12**(6): p. 385-92.
47. Senovilla, L., et al., *Image Cytofluorometry for the Quantification of Ploidy and Endoplasmic Reticulum Stress in Cancer Cells.* Methods Mol Biol, 2017. **1524**: p. 53-64.
48. Panaretakis, T., et al., *The co-translocation of ERp57 and calreticulin determines the immunogenicity of cell death.* Cell Death Differ, 2008. **15**(9): p. 1499-509.
49. Panaretakis, T., et al., *Mechanisms of pre-apoptotic calreticulin exposure in immunogenic cell death.* EMBO J, 2009. **28**(5): p. 578-90.
50. Harding, H.P., et al., *An integrated stress response regulates amino acid metabolism and resistance to oxidative stress.* Mol Cell, 2003. **11**(3): p. 619-33.
51. Pakos-Zebrucka, K., et al., *The integrated stress response.* EMBO Rep, 2016. **17**(10): p. 1374-1395.
52. Pamer, E. and P. Cresswell, *Mechanisms of MHC class I--restricted antigen processing.* Annu Rev Immunol, 1998. **16**: p. 323-58.
53. van Endert, P.M., *Genes regulating MHC class I processing of antigen.* Curr Opin Immunol, 1999. **11**(1): p. 82-8.
54. Michaud, M., et al., *Subversion of the chemotherapy-induced anticancer immune response by the ecto-ATPase CD39.* Oncoimmunology, 2012. **1**(3): p. 393-395.

55. Ma, Y., et al., *Autophagy and cellular immune responses.* Immunity, 2013. **39**(2): p. 211-27.
56. Ko, A., et al., *Autophagy inhibition radiosensitizes in vitro, yet reduces radioresponses in vivo due to deficient immunogenic signalling.* Cell Death Differ, 2014. **21**(1): p. 92-9.
57. Rao, H.S. and M. Kamalraj, *Synthesis and characterization of 4-aryl-4H-chromenes from H-cardanol.* Nat Prod Commun, 2014. **9**(9): p. 1333-40.
58. Ladoire, S., et al., *Combined evaluation of LC3B puncta and HMGB1 expression predicts residual risk of relapse after adjuvant chemotherapy in breast cancer.* Autophagy, 2015. **11**(10): p. 1878-90.
59. Galluzzi, L., et al., *Activating autophagy to potentiate immunogenic chemotherapy and radiation therapy.* Nat Rev Clin Oncol, 2017. **14**(4): p. 247-258.
60. Martins, I., et al., *Chemotherapy induces ATP release from tumor cells.* Cell Cycle, 2009. **8**(22): p. 3723-8.
61. Ma, Y., et al., *Anticancer chemotherapy-induced intratumoral recruitment and differentiation of antigen-presenting cells.* Immunity, 2013. **38**(4): p. 729-41.
62. Tey, S.K. and R. Khanna, *Autophagy mediates transporter associated with antigen processing-independent presentation of viral epitopes through MHC class I pathway.* Blood, 2012. **120**(5): p. 994-1004.
63. Oliveira, C.C. and T. van Hall, *Alternative Antigen Processing for MHC Class I: Multiple Roads Lead to Rome.* Front Immunol, 2015. **6**: p. 298.
64. Walter, S., et al., *Multipeptide immune response to cancer vaccine IMA901 after single-dose cyclophosphamide associates with longer patient survival.* Nat Med, 2012. **18**(8): p. 1254-61.
65. Chen, X., et al., *Calreticulin promotes immunity and type I interferon-dependent survival in mice with acute myeloid leukemia.* Oncoimmunology, 2017. **6**(4): p. e1278332.
66. Caron, E., et al., *The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.* Mol Syst Biol, 2011. **7**: p. 533.