Université de Montréal

Improving anti-cancer therapies through a better identification and characterization of non-canonical MHC-I associated peptides

Par

Maria Virginia Ruiz Cuevas

Programme de bio-informatique,

Département de Biochimie et Médecine moléculaire

Faculté de Médecine

Thèse présentée en vue de l'obtention du grade

Philosophiae Doctor (Ph.D.) en Bio-informatique.

Décembre 2022

© Maria Virginia Ruiz Cuevas, 2022

Université de Montréal

Programme de Bio-informatique, Faculté de Médecine

Ce thèse intitulée

Improving anti-cancer therapies through a better identification and characterization of non-canonical MHC-I associated peptides

Présenté par

Maria Virginia Ruiz Cuevas

A été évaluée par un jury composé des personnes suivantes

Adrian Serohijos

Président-rapporteur

Claude Perreault

Directeur de recherche

Sébastien Lemieux

Codirecteur

David Knapp

Membre du jury

Mathieu Lavallée-Adam

Examinateur externe

El Bachir Affar

Représentant du doyen

Résumé

Les preuves de plus en plus nombreuses de la traduction des protéines non canonique ont suscité l'intérêt pour leur identification et leur caractérisation en vue de leur utilisation dans les immunothérapies. En outre, des études récentes sur le répertoire des peptides associés au complexe majeur d'histocompatibilité de classe I (CMH-I, connus sous le nom de MAPs ou immunopeptidome), ont suggéré que les MAPs dérivés de ces traductions sont des cibles potentielles pour l'immunothérapie du cancer. L'objectif de cette étude était donc d'évaluer l'impact de ces MAP dans le cancer en développant des méthodes pour faciliter leur identification et leur validation en tant que cibles potentielles pour l'immunothérapie.

Afin de faciliter l'identification des protéines non canoniques, nous avons développé Ribo-db, une approche protéogénomique qui combine le séquençage de l'ARN, le profilage ribosomal et la spectrométrie de masse. Cette approche permet de générer des bases de données spécifiques visant à inclure la diversité des protéines. Notre analyse avec Ribo-db d'échantillons de lymphome diffus à grandes cellules B (DLBCL) a révélé qu'environ 10% des MAP étaient dérivés de protéines non canoniques. Ces protéines avaient des propriétés distinctes par rapport à celles dérivées de protéines canoniques. Elles étaient plus courtes et avaient une stabilité plus faible, mais une plus grande efficacité dans la génération de MAPs. Fait important, nous avons constaté un chevauchement limité entre les protéines non canoniques détectées dans l'immunopeptidome et celles détectées dans le proteome entier, ce qui suggère l'existence de deux répertoires distincts de protéines non canoniques.

Sachant que les MAP non canoniques peuvent être des cibles efficaces pour l'immunothérapie du cancer, nous avons développé BamQuery, un outil permettant d'évaluer leur expression dans les tissus afin de déterminer s'ils peuvent être utilisés dans un vaccin. BamQuery vise à prédire la probabilité de présentation au CMH-I de chaque MAP dans différents tissus sur la base de son expression ARN. En utilisant BamQuery, nous avons découvert que des antigènes tumoraux (TA) précédemment identifiés seraient fortement exprimés dans les tissus sains, ce qui en fait de mauvais candidats pour l'immunothérapie. En outre, nous avons également

identifié des cibles immunothérapeutiques très potentielles dans DLBCL qui étaient dérivées de traductions non canoniques. Ces cibles se sont révélées prometteuses car elles étaient peu exprimées dans les tissus normaux mais fortement exprimées et partagées dans les échantillons tumoraux. Ainsi, BamQuery s'est avéré être un outil utile pour identifier et hiérarchiser les cibles immunothérapeutiques potentielles.

Dans l'ensemble, nos recherches ont indiqué que les régions non canonique du génome augmentent la diversité des MAPs qui peuvent être reconnues par les cellules T. De plus, l'expression des MAPs dans les tissus peut être utilisée comme un prédicteur de leur présentation au CMH I afin d'identifier des cibles fiables pour l'immunothérapie, ce pour quoi BamQuery est un outil efficace.

Keywords: protéines non canonique, antigènes tumoraux, immunopeptidome, complexe majeur d'histocompatibilité de classe I, protéogénomique, séquençage du profilage ribosomique, spectrométrie de masse, lymphocytes T.

Abstract

Increasing evidence of non-canonical protein translation has sparked interest in their identification and characterization for use in immunotherapy. In addition, recent studies on the repertoire of major histocompatibility complex class I (MHC-I) associated peptides (MAPs or immunopeptidome), have suggested that MAPs derived from these translations are potential targets for cancer immunotherapy. Therefore, the aim of this study was to assess the impact of these MAPs in cancer by developing methods to facilitate their identification and their validation as potential targets for immunotherapy.

To facilitate the identification of non-canonical proteins, we developed Ribo-db, a proteogenomic approach that combines RNA sequencing, ribosome profiling and mass spectrometry. This approach enables the generation of specific databases aimed at including protein diversity. The use of Ribo-db to analyze diffuse large B-cell lymphoma (DLBCL) samples revealed that approximately 10% of MAPs were derived from non-canonical proteins. These proteins had distinct properties compared to those derived from canonical proteins. They had shorter lengths and lower stability, but greater efficiency in generating MAPs. Importantly, we found limited overlap between the non-canonical proteins detected in the immunopeptidome and those detected in the whole proteome suggesting the existence of two distinct non-canonical protein repertoires.

Knowing that non-canonical MAPs can be effective targets for cancer immunotherapy, we developed BamQuery, a tool to assess their expression in tissues to determine whether they can be used in a vaccine. BamQuery aims to predict the probability of MHC-I presentation of each peptide in different tissues based on its RNA expression. Using BamQuery, we found that previously identified tumor antigens (TA) would be highly expressed in healthy tissues, making them poor candidates for immunotherapy. In addition, we also identified highly potential immunotherapeutic targets in DLBCL that were derived from non-canonical translations. These targets showed promising as they were poorly expressed in normal tissues but highly expressed

and shared in tumor samples. Thus, BamQuery proved to be a useful tool for identifying and prioritizing potential immunotherapeutic targets.

Overall, our research indicated that non-canonical regions of the genome increase the diversity of MAPs that can be recognized by T cells. Furthermore, the expression of MAPs in tissues can be used as a predictor of their presentation to MHC I to identify reliable targets for immunotherapy, for which BamQuery is an effective tool.

Keywords: non-canonical proteins, tumor-antigens, immunopeptidome, Major histocompatibility complex class I, proteogenomics, ribosome profiling sequencing, mass spectrometry, T cells.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Abbreviations

A

aa	Amino acid
aeTSA	Aberrantly expressed TSA

B

B-LCL	B-lymphoblastoid cell line

C

CRISPR	Clustered regularly interspaced short palindromic repeats
CTA	Cancer-testis antigen
CTLA-4	Cytotoxic T-lymphocyte-associated protein 4

D

Db	Database
DC	Dendritic cell
DLBCL	Diffuse large B-cell lymphoma.
DRiP	Defective ribosomal product

E

ER	Endoplasmic reticulum

ERE	Endogenous retroviral element

F

FDR	False discovery rate

H

HLA	Human Leucocyte antigen

L

LC-MS/MS	Liquid chromatography-MS/MS
lncRNAs	long non-coding RNAs

M

m/z	Mass-to-charge ratio
MAP	MHC I-associated peptide
MCS	MAP coding sequence
MHC	Major histocompatibility complex
MHC I	Major histocompatibility complex class I
MHC II	Major histocompatibility complex class II
mRNA	Messenger RNA
MS	Mass spectrometry
mTEC	Medullary thymic epithelial cell

mTSA	Mutated TSA

N

ncMCS	non-canonical MCS
ncMAP	non-canonical MAP
ncRNA	non-coding RNA
nt	Nucleotide

O

ORF	Open reading frame

P

PD-1	Programmed death-1
PD-L1	Programmed death-ligand 1
PSM	Peptide-spectrum-match

R

RDP	Rapidly degraded protein
Ribo-seq	Ribosome profiling
RNA-Seq	RNA sequencing
rphm	Read-per-hundred-million

S

SLiPs	short-lived proteins
SNVs	Single nucleotide variants

T

TA	Tumor antigen
TCGA	The Cancer Genome Atlas Research Network
TCR	T-cell receptor
TME	Tumor microenvironment
tpm	Transcripts per million
TSA	Tumor-specific antigen

U

UTR	Untranslated region

W

WES	Whole-exome sequencing
WGS	Whole-genome sequencing

« One day will be the day for those that

said « someday ... » »

and the day has come!

To my mother in heaven, the person who has given me the most love by giving her life to give it to me. Every day, until my last day, I will try to make you proud and make you feel it was worth it. I love you back.

To my mother on earth, although we do not share blood you have been the one who has inspired me the most to be ambitious and to decide for myself my future.

To Diego, my lifesaver, thank you for being there in my darkest and most terrifying moments. I could never have gotten this far or anywhere without you holding my hand to pull me through.

To my 5 year old self, it's okay to be afraid, embrace yourself because you are already strong enough to face it. Along the way you will find that you will face so many tough storms that after a few you will start to realize that you are becoming the strongest person you have ever known. And one day, someday, you will realize that you have even been strong enough to do the long, hard work that others never needed or could not do to swallow the world. There, in that moment embrace yourself again, because you will once again remember that your life has been worth living and that you have an incredible story to tell. And since you have proven it all to yourself, you will have no choice but to believe in yourself and abandon all the fears that once served you, because now you will be aware that you were always above them!

Acknowledgements

We are all ports

We are like ports that wait for and hold ships. Throughout our lives, people come and go like ships in a harbor. Some arrive like abrupt waves; others arrive gently; some out of curiosity and others by mere chance. It doesn't matter how, why, or when, but from all of them, from all those who have come and gone and those who have stayed, a lesson will be kept. Invisible bonds will be made that will cling to our memory and spread throughout our existence like indelible traces. They will leave us good or bad memories that will give meaning to our life, that will define us, that will shape us. And in the end, we will be just that, the sum of all of them, of all those who once came to our port.

And we are also voyageurs, we dock in different ports to leave our own mark, our memory. And there we will remain or not, like the waves that come and go.

Me

Thank you to all the ships that have crossed by my port, I am a bit of each of you!

And unexpectedly writing these last lines became a challenge. I have too many things to be grateful for and too many people to thank. I am the sum of much, I am the result of much. And appreciating this mere moment of finishing writing my thesis makes me realize once again how far I have come since I was a kid playing with chickens in the countryside. I can't help but realize, when I look back, how steep my road has been, but how beautiful the landscape has become. So, I hope to say thank you for everything and everyone for helping me lay all the blocks that have built the edifice of my life. And if by some bad chance I have forgotten someone please believe me it is not bad intentioned, it is just that like chickens I also have a bad memory.

First, I would like to thank my supervisors **Claude Perreault** and **Sébastien Lemieux** for betting on me and giving me the opportunity to cross the ocean to Canada. Probably, like you, I did not imagine

that one day, five years later, I would be writing "acknowledgement" lines in a thesis. Because to be honest, during all this time I was almost never certain that I would ever write one. And here I am and here it is, inconceivable[1]! I hope you believe me when I tell you that during these five years, I learned so much, but so much! So much that I think I even learned to love the cold of Canada and to appreciate how my lungs each winter clear as they freeze.

I'd like to start by telling you, **Claude**, a little anecdote. When the pandemic hit in 2020 and we were all locked in, I watched the Last Dance documentary[2] that tells the story of one of my childhood heroes, Michael Jordan[3]. This would have no relevance today in my thesis if it weren't for the fact that during episode 7 of the series, I couldn't help but stop thinking about you. And don't worry it's not because of something weird, it's not like I saw you playing basketball or gambling or anything like that, but because at the end of the episode I suddenly realized that you were the Michael Jordan of my life, the superstar I decided to follow! And I assure you that realization made me appreciate how far you've taken me and how resilient you've taught me to be. You wanted to win, but you wanted me to win too and be a part of it. So, I want to thank you deeply and sincerely for pushing me to succeed. I feel like I can never thank you enough but thank you truly! You have taken me to another level that even I could not have expected.

Seb, I don't know where to start but thank you so much. I must thank you for always being a mental challenge for me, I always enjoyed our debates because I had to think hard to find something intelligent to say. I really feel lucky to have had you as a PI, because you have allowed me to be myself, and that is priceless. I am aware that nowadays it is not within everyone's reach to give and receive that gift, so I thank you very much for giving that privilege to me. Thank you for letting me be as childish as I allowed myself to be without making me feel uncomfortable or judging me. Maybe because you saw me trying to live my life backwards at the Benjamin button's[4] style, or I don't know what but thank you. You've allowed me to collect beautiful memories in the lab! Like the virtual game night with the whole lab during the pandemic where I kept screaming and laughing out loud and dying the whole time!! Or the last lab Christmas party!! Ohhhh my goshhh! I will keep that beautiful moment so close to my heart. Thanks to you and everyone in the lab for that day letting me be a karaoke rock star. I know I was practically scratching the walls with all my singing/screaming. So, thank you truly for your patience all these years and I hope I can count on that for many more.

I would like to express my gratitude to the members of my thesis committee, **Adrian Serohijos** and **Pierre Thibault**. Thank you very much for taking the time to listen to me, advise me and support me during this thesis. You are the best members anyone could wish to have on their committee. Thank you very much.

I also want to thank you, **Elaine Meunier**. You gave me a nice gift that day I went to your office, and you read through me the anxiety I was trying to keep inside. Thank you for that nice conversation, I really enjoyed learning more about you, you are super inspiring!! You gave me so much that day, you can't imagine how much, you made me forget my storm, what a powerful thing to give. You are great, and I have no words to understand how patient you are. Because I know all the time, I am messaging

you to send you files that I forget to attach, and you have always been kind to me no matter what. I can't imagine being in your situation if there are so many other dumb students like me who forget to subscribe to the university on time, or don't send the right documents, etc. You are great for that and so much more. Thank you for your patience and all your kindness all these years, I always loved all your good vibes, you are an extraordinary woman.

And so, the time has come to thank all my colleagues, lab members of these two teams! To all the past and present members, thank you! You are all exceptional scientists to whom I take my hat off! Thanks to those who were good to me and to those who were not so good! You all became life lessons, and like all life lessons, they became good over time! Above all, I would like to thank **Qing, Anca, Assya, Greg, Caro labella, Safia, Marjo, Jeremy, Tom, Nandita, Caro, Marie-Pierre, Catherine**, I know you all have always been good to me, and I am deeply grateful to you for that! The world would be a better place for everyone if there were more people like you, you guys are the top!

To my dear **Qing** from the bottom of my heart, you were the first and only one for a long time who supported me when I came to the lab! You are the one responsible for me writing this thesis today. I remember like it was yesterday that we both dreamed of finishing our PhDs at the Agora, and I blinked and today you are done, and I am about to follow the pace. This is really happening? haha! Thank you so much my dearest, I think you deserve ALL but only the BEST, I love you so much!

To **Anca**, my dear! You are by far one of the most generous, reliable, and brilliant people I have ever had the pleasure of knowing. Thank you for considering me your friend all these years. I will always be there for you wherever I am as you have always been there for me. Thank you for boosting my confidence when I needed it so badly, thank you for rehearsing presentations with me and giving me so much advice on so many aspects of my life! I want you to always remember that for me you are someone exceptional!

To **Assya** ma cherie, I've already told you how much I admire you! Gosh you are so amazing; I wish I was good at imitation games so I could fake it until I made it to be as great as you, but I'm not good at any game, so...no hope! No but seriously, how is it possible that you know soooo many things? It is not normal!! Thank you **Assya**, for being you and letting me be myself. You are like a mentor to me; you will never know how much I have learned from you and how much you have taught me and how much I miss youuuuu. Thank you for being there, listening to me and pushing me with courage to continue. Thank you for those moments, preparing pumpkins, those memories are some of the bests at IRIC. I love you very much, and when you win a Nobel, or a Turing award remember that I was the first one to predict it!

To **Greg**, your dance moves have had me stunned all these years, I don't know how to move like a spaghetti like you do, when the most I can do is walk haha! Thank you for always making me feel good, you gave me a sense of belonging, a feeling of brotherhood. Thanks for all your great advice always very well appreciated. I want you to know that I have really enjoyed working with you at BamQuery, you are someone anyone can learn a lot from, as you are an exceptional scientist. I admire your style of reasoning; I definitely learned a lot from you! I wish you all the best opportunities

in the world because for me you deserve only the best. I hope someday we will cross paths again and go watch another hockey game where the Canadiens[5] win for once. Thank you so much again, thank you for being the best!

To **Caro labella**, belladonna! Thank you bella mia for being a formidable person always ready to help. Thank you for those long and nice conversations that we had, you always inspiring me to be more organized and to make lists! Sharing with me teas and recipes that I will never make because you know I am a danger in the kitchen. Thank you for teaching me how to do so many things, I will never forget that you took the time to explain to me how to make the first images for my first article, you are one of a kind! You always were patient with me, even when I was doing the silliest things! Now, you know you count on me for life, I will be there whenever you need me, and that: Mi casa tu casa, always!! You got me for life bella! Love you my Caro and Thank you for everything!

To **Safia** dear, it takes one to recognize one! You are a strong woman with a beautiful heart. You are such an inspiring person, I admire you deeply and you are among my heroines!! You will conquer the world no doubt! You are so strong, so smart such a good person and I can only be grateful for your existence in my life! And yes, I didn't understand for a long time, but now I understand why you were so excited to see that Avengers movie[6] when we were in Boston, it was because of baby groot!!! Now I understand that nobody can resist the I am groot[7]!! Haha! Love you my Cherie, and doesn't matter if someday you will move to the other side of the planet I will go to visit because I am not going to lose you!

To **Marjo** ma cocotte d'amour, oh mon amour, ma Cherie, ma soulmate! Have I told you how much I love you, silly? Ma **Marjo**, I wish I could narrow at this very moment the oceans and half the world that separate us just to tell you to your face, I love you so, so much!! You are one of the most beautiful souls, one of the most evolved persons I have ever met. I can't believe someone at your age could be that way. I swear I am so happy finishing my PhD but if I could go back in time, I would love to sneak into any of those times we were together at IRIC to laugh as loud as we used to. I would love to go back to any of those long nights at IRIC that we would spend until midnight or 2am working in sandals, eating whatever cookies we managed to get out of those freaking vending machines and listening to the reggae or African music you loved so much, the happy hippie life quoi!. I loved everything, any moment near you. You are a blessing in my life ma cocotte, thank you for giving me so much of you, for giving me back my faith in humanity! I want you to achieve all your dreams especially the ones related to good jobs and money... because you know I count on you to become rich so I can finally be sponsored by you to travel the world... still waiting and getting frustrated... haha! I love you my love and I promise you that one day I will go to the other side of the world to tell it to your pretty face!!

To **Jeremy** and **Tom**, it was always a pleasure to meet you and work alongside you. I can see that you will fly as high as you want because you are but flooded with intelligence, kindness and humility and people will see that. Thank you because you were always willing to answer any of my questions, and help with anything I need!

To **Nandita** my sweet beeee, my beautiful friend inside and out! Nandita, I feel so honored to have walked these last years by your side. you are a wonderful person who only teaches kindness and goodness to all the people around you! I wish you a bright future, so you keep flooding the world with so much love! Thank you for all your advice, for listening to me, for helping me to be better and to improve! I love you my sweet beeee and always!

To **Caro** ma Cherie, thank you for all the chocolates you always had ready to share! Thank you, ma Cherie, for taking care of those cells some time ago but most of all thank you for all your genuine kindness. I always appreciated the conversations with you when you asked me How was I doing because I always felt your genuine desire to know the truth and not the typical ruthless politeness, thank you for so being real!

To **Catherine** ma belle! swing danse, what a nice discovery eh!?! I am so glad we did it together and I enjoyed it so much with you! Thank you for being this beautiful human being to be around! I am so grateful to have crossed my path with yours, because contrary to many people who came to Quebec and who will leave without having growth a friendship with a Quebec person, and I am proud to say it was not my case! You are one of the two Quebecers that I will carry from now and forever in my heart. I love you ma belle and I am so happy for you and for your beautiful family, you deserved it and I wish you all the blessings!

And last but not least, to **Marie-Pierre! MP**, ma Cherie I promise I have tears in my eyes just thinking of you and thinking what I'm going to write to make it as fair as possible! You know what you meant to me, you were a lifesaver **MP**, and it would never be enough to tell you how grateful I am. I admire you, hopefully someday I will be at least a little bit of the excellent scientist you are! You are one of the smartest and most amazing people in science I have ever met. You surprise me every time, my God, what do you eat to be so smart? But on a serious note, and with my heart in my hand, thank you forever, for being the mentor I always wanted for myself! Thank you, ma Cherie, for being that one rock in my PhD, this thesis should have your name all over it, because I owe you a big one! Without you I wouldn't be here finishing! Whatever I become, it will be because I knew you one day, and I trust it will be great.

To **Patrick Gendron**, I couldn't leave you out of this thesis even though I don't like you very much!!! Haha!!! No... not true!!! Oh, my cher Pat, I hope you will miss me as much as I'll miss writing you all those messages on Slack to report you mostly problems! Hahaha. I know you hate me, and maybe you are developing PTSD with the notification sound in Slack thinking it's me again but it's ok, It's tough love! Anyways, thank you, thank you, thank you for helping me with the BamQuery page and having the patience to explain to me I don't know how many times the back and forth strand colors of IGV, by the way I still don't quite get it haha !!

To **Mathieu Courcelles**, I could not leave you out of this thesis either. Thank you very much for all your help, for all your patience in answering all my questions. I know you know I was lying all the time when I said: "quelques petites questions! They were never "petites", they were long, one after the

other, a cascade quoi! But a thousand thanks for all your patience and for wanting to share your knowledge with me! I will never forget that!

To all my friends at IRIC, but especially to **Amandine, Alizee, Chloe, Eloise, Lea, Swati, Unain, Charles, Lara, Laia, Layane, Lina, Anais, Gaby, Roger, Maria, Pedro,** you are the best in the chaos, like flowers in the desert, like radiant sunshine in winter and you know what I mean by that. You keep it real, and you are truly beautiful souls! Thanks for all the scientific and non-scientific conversations, for the lunches by the windows, for all the coffees! Thanks for making IRIC a better place for me and everyone around you - you guys rock!

And to my dear friends, you mean the whole world to me! You know that I love each one of you so much that you are at the top of my Hall of Fame of all times, my life support! I LOVE YOU DEEPLY.

Eevaaaaa, mon italieenne favorite de tous-tous le temps, je t'aime ma Cherie pour toute la vie, tu signifie une soeur pour moi !!

Marta mi bruja, usted no se puede deshacer de mi nunca, por que usted es como el aire que respire, la amo mi boba.

Saralina, corola, my Beyonce de la vida, ya tu sabes lo que significas para mi, hermana de otra madre, mi amor lindo! En donde sea que estes y que yo este, me tienes a mi. Te amo mi corola!

Amanda mi Mana, tu eres mas Buena que el pan, mana! Ya sabes como te adoro, pues te sigo adrando como siempre y mas!

Tiph ma cherie, Meli-Meli mon Coeur de melon et ma belle Chloe, si aujourd'hui je suis une dre c'est grâce à votre soutien, à vos encouragements et à votre merveilleuse amitié. Merci pour tout, Jussieu est un de plus beaux souvenirs grâce à vous. Vous faites partie de ces Français que j'aime, que j'admire et que je suis fière d'avoir rencontré. Je vous aime pour toujours.

Nika my sister, tu es mon âme sœur ! Merci de m'avoir ouvert ton cœur, tu es l'un des êtres humains les plus beaux et les plus purs que j'ai eu le plaisir de rencontrer. Tu es une source d'inspiration dans ma vie et je veux que tu saches que je suis et serai toujours là pour toi. Je t'aime!

Caro cherie, Pascalle, Darwin, Pixel, Claire ma Cherie and my baby Koala, Emilie, ma belle famille de la rue Baldwin, vous êtes mon ancrage à Montréal, au Canada. Je vous aime tous et je vous remercie pour tous les beaux moments passés ensemble, toutes les célébrations, pour tous nos jeux, nos bbqs, les potlucks. Vous allez me manquer irréparablement !

Marta my amora & Alex, os quiero un monton mis queridos mios! Ustedes son mas buenos que el pan y no os voy a perder de vista! Mi casa tu casa como siempre, en donde sea que yo este!

Oumnia, Abdel my little brothers et fcg Jade all in all my familytrip, je vous aime, mais vous le savez déjà, mais encore je vous aime! Merci pour toutes nos aventures, pour tous les exploding kittens, pour tous nos fous rires, pour nos voyages ! J'ai hâte de repartir à l'aventure avec vous, car je ne me lasse pas de vous !

Elisabeth ma maman de Poitiers, s'il y a une bêtise à faire, nous sommes toutes les deux là ! haha! Je t'aime ma cherie, merci d'avoir été la première personne à nous accueillir à bras ouverts en France et d'être là après toutes ces années ! Ta fille que t'aime a l'infini et qui adore chaque moment avec toi!

Cristiane & Daniel mes chers Parents, depuis que je vous ai rencontrés, je vous connais comme une belle famille, je vous admire tellement. Merci de m'avoir intégrée dans cette belle famille, tous les beaux moments partagés et tout votre amour à ce jour remplissent mon cœur de joie!

Sure, I could write pages to tell you thank you for all you are being with me and it would be even more pages than this thesis! But you all know me well, you all know me best, you all know that you are the foundation of my building, you all know that you have me for life, and you all know that I mean it. Although you may not be happy... you'll never get rid of me... Muahaha! And to all of you, if this thesis were a poem or a song, I would dedicate it to you! Thank you for your support all these years, you have helped me to go far and do well, my love and devotion to you! I know how unbearable I can be at times, complaining, moaning and sometimes just being my dumbest self... and you have stuck with me through it all, inconceivable, but thank you! I love you until the end of my existence and even further than that because I plan to hunt you down as a Casper... and I know most of you know that I know hunting tactics after watching all the horror movies on Netflix[8]. So, get ready and have eggs and rice for me... because I know some of you know that's my favorite thing to eat... eggs.... hahaha !

A mi Familia en Colombia, a mis dos **madres**. Gracias por imprimir tanta fuerza en mí, tanta rebeldía, tantas ganas de salir adelante, de estudiar y de ser alguien en la vida. Espero que teniendo una tesis las dos se sientan orgullosas de mí, que ha sido el único punto por el que he querido venir tan lejos. ¡Gracias a mi **padre**, que, aunque no tengo los mejores recuerdos de infancia porque fue tan jodidamente traumática todos ellos me ensenaron a ser aún más fuerte de lo que ya tenía en las venas! Dicen que en los tiempos difíciles se conocen a las personas, y todos esos tiempos me permitieron a reconocerme la fuerza que llevo adentro y el fuego santandereano que me quema. A mis **hermanos**, chinos ahí estaré siempre, ¡los quiero!

Ma, don José, Diego coshi, ustedes siempre serán mi familia. Gracias por adoptarme como suya, ya sé que las palabras no serán nunca lo suficiente para agradecerles por estar ahí cuando todo estaba oscuro alrededor de mí y que solo veía el final. ¡Gracias por no dejarme allí en ese lugar! Los amo eternamente y no importa ni el tiempo ni la distancia que se ponga entre nosotros, mi amor les corresponde a 1000%. ¡Gracias desde lo profundo de mi amor, gracias! Los amo hasta el infinito y mas alla!

A **Omar Zambrano** y **Gloria Su**, no estaría hoy aquí si no fuera por ustedes en mis inicios. Ustedes fueron de los primeros que apostaron por mí. Gracias Omar por esos dos años en que me diste trabajo en tu empresa cuando yo no sabía ni pio de nada y que vivía en una burbuja en mi cabeza. Lo único que sabia era quería estudiar y salir adelante porque eso fue lo que me dijo mi mama que debía hacer. ¡Y gracias por que fuiste el primero que me ayudo a construir este sueño! ¡Gracias mil gracias! ¡Gloria Su, la mejor jefecita del mundo! Nunca me podre cansar de darte las gracias por todo el soporte cuando estaba en el banco y cuando fue el accidente. Tu eres de las personas más buenas del planeta que haya conocido, gracias mil gracias que no me abandonaste cuando más necesitaba. ¡Yo sé que no guardamos un contacto cercano pero que la vida te pague mil veces en millones para ti y tu familia todo lo que hiciste por mí! ¡Muchas gracias del fondo de mi corazón!

Estas gracias no estarían completas si no les diera las gracias a los medicos que me salvaron la vida el 2 diciembre 2005. Oh my, son tan grandes esas palabras y significan tanto y más en este momento, porque ustedes me salvaron literalmente la vida. Gracias eternas por ello. Hoy no estaría caminando, yendo y viniendo como si un bus no me hubiera pasado por encima, si no fuera por ustedes. Es que a veces yo ni me lo creo, que mi cuerpito pequeño haya sido tan fuerte para soportar todo lo que ha tenido que soportar. Y no me creo tampoco que hoy en dia haya dado ya casi la vuelta al mundo y ahora este terminando un doctorado! Gracias por salvarme la vida y por ser las magníficas personas que fueron allí en el hospital, por hacerme reír cuando no tenía más que razones que ponerme a llorar. ¡Gracias principalmente a **Juancho, Andres Pinzon y Efrain Leal**! Verdaderos héroes.

To **Badr**, my love, my cheri, my cuchurummi. Merci beaucoup cheri for believing in me, always more than me. Thank you for always pushing me out of my comfort zone, I couldn't have walked with anyone else these last 3 years! You are a great guy, a beautiful person inside and out and I feel so lucky to be sharing this moment with you and my life. I love you so much! Thank you for letting me see all your sides. I have learned so much from you, I have evolved so much thanks to you. It has been an amazing and wonderful journey, finding you has been a blessing! You are just the right fit to be the adventurer I always wanted to be. Thank you for all the trips we have done Alberta, Gaspésie and Machu-Pichu that I could not have done with anyone else but you. I can't wait to plan the next ones. I love us discovering things, places, thoughts! Cheri, thank you for helping me get up at moments when I thought I couldn't anymore, thank you for being willing to fight battles with me that are not your own. Thank you for being so patient, so loving and so concerned about me. Thank you for making me breakfast, or lunch or dinner so many times when I was with my head in the clouds.... Thank you for always trying to make me feel like I was the best chicken in town, I love you

so, so, so, so much, to the Andromeda and back that I can't wait to start my next adventurous life by your side!

And finally, this may be the first and the last time that someone has cited Snoop Dogg[9] in his/her thesis... but I think his speech delivered when he got a star on the Hollywood Walk of Fame on November 19, 2018, described very well my feeling...

Last but not least, I wanna thank **me**
I wanna thank me for believing in me
I wanna thank me for doing all this hard work
I wanna thank me for having no days off
I wanna thank me for, for never quitting
I wanna thank me for always being a giver
And tryna give more than I receive
I wanna thank me for tryna do more right than wrong
I wanna thank me for just being me at all times...
Snoop Dogg

[1] Inconceivable: word that has stuck with me since I saw the movie The Princes Bride
[2] Last Dance documentary on Netflix
[3] Michael Jordan, the greatest basketball player of all times.. Sorry LeBron
[4] Refers to the beautiful film: The Curious Case of Benjamin Button
[5] Canadiens: Montreal hockey team that most of the time loses games
[6] Avengers: marvel universe movie that Badercito made me watch and that I liked in the end
[7] A talking tree in the Avengers movie and all it says is: I am Groot!
[8] From my experience I advise to not watch Hereditary alone at home... you wont sleep
[9] Snoop Doog: cool rapper who loves to smoke weed

Overview

Cancer immunotherapy is part of the cancer treatment's arsenal, which involves stimulating the immune system to better fight cancer. Cancer can occur and prevail despite the ability of the immune system to detect and destroy it. As we age, the accumulation of mutations can impede our natural ability to defend ourselves against the onset of cancer. Once established, cancer cells can develop various mechanisms to evade the immune system. They can manipulate the normal cells surrounding tumors, turning them into barriers to counteract immune responses. Alternatively, cancer cells may remain undetected by altering genes responsible for antigen processing and presentation or by modifying cell surface-anchored proteins that impede immune recognition.

Currently, there are several types of cancer immunotherapy that, depending on the type of cancer, can be combined with other types of treatment, including conventional approaches such as surgery, chemotherapy, and radiotherapy. Among cancer immunotherapy strategies, cancer vaccines for patients already suffering the disease aim to boost the immune system by educating it to recognize and react against cancer cells carrying specific antigens. These antigens must have critical characteristics to ensure the success of the vaccine: they must be highly specific for cancer cells, and they must be able to stimulate immune cells sufficiently to promote their destruction.

However, the design of cancer vaccines is not trivial. The nature of cancer-specific antigens still needs to be elucidated to identify the best therapy-actionable antigenic targets. In recent years, immunotherapy has benefited from advances in next-generation sequencing, high-throughput mass spectrometry and bioinformatics to manage and analyze vast amounts of data. Taking advantage of these breakthroughs, several studies have helped to elucidate the presence of a hidden proteome in cancer cells. This non-canonical proteome was shown to be a rich source of cancer-specific antigens.

Therefore, the work presented in this thesis addresses the design of bioinformatics approaches to analyze genomic and proteomic data intending to understand the non-canonical

proteome that may be at the source of suitable antigenic targets. In addition, it presents a tool for *in-silico* validation of antigen specificity and immunogenicity to allow prioritization of antigens for subsequent vaccine development.

This doctoral thesis is presented in 4 chapters. Chapter 1 introduces the mechanism used by cells to communicate to the immune system their internal health status and how this mechanism can be used for the development of immunotherapies. Chapter 2 presents an article published in the journal Cell reports in which the presence of a non-canonical proteome in lymphoma cell lines is highlighted. Chapter 3 presents an article being reviewed by Genome Biology in which a software tool for the prioritization of antigens based on their RNA expression in normal and cancerous tissues is presented. Chapter 4 presents the conclusions and discusses the work presented in this thesis as well as perspectives for future work.

Chapter 1 – Introduction

The immune system protects our body from external aggressions through two lines of defense that work together to prevent the threat: the innate and the adaptive immune system. When the body encounters an invader, the innate immune system is activated as the first line of defense. This system involves a range of defenses, including anatomical barriers like the skin and mucous membranes, as well as physiological mechanisms that regulate temperature, maintain low pH levels, and produce chemical mediators. In addition to these barriers, phagocytic and endocytic mechanisms are employed to contain the pathogens, while inflammatory responses are also triggered to help fight off the infection[1]. Within the range of cellular mechanisms used to contain pathogens, neutrophils and macrophages are particularly effective at eliminating threats. They secrete highly destructive substances, such as enzymes that break down pathogen proteins and reactive chemicals that kill to then engulf and digest the damaged pathogens (phagocytosis)[2]. When this initial attack is insufficient to eradicate the infection, lymphocytes are activated to initiate a specific response against the pathogen. This activation marks the beginning of the second line of defense: the adaptive immune system. The adaptive immune system relies on the coordinated efforts of B and T lymphocytes, each armed with unique receptors designed to detect and respond to antigens[3]. These receptors are designed to recognize antigens bound to major histocompatibility complex molecules (MHCs), which are found on the surface of cells infected by pathogens[4]. By binding to these antigens, B and T lymphocytes initiate a targeted immune response to eliminate the infected cells and combat the pathogens. The pathogen-specific response provides long-lasting protection through the generation of immune memory cells conferred by the T lymphocytes. Although both B and T lymphocytes are essential for the proper functioning of the adaptive immune system, here we will focus on how T lymphocytes explore and recognize the set of peptides presented on the surface of cells, which together constitute what is known as the immunopeptidome.

1.1 The immunopeptidome

The immunopeptidome refers to the repertoire of peptides displayed on the surface of nucleated cells through their association with major histocompatibility complex class I (MHC-I) molecules, hereafter referred to as MHC-I-associated peptides (MAPs). MAPs are short peptides (8 to 11 amino acids in length) generated through the process of antigen processing and presentation[5]. During this process, proteins undergo degradation, yielding a diverse array of short peptides. Some of these peptides are then loaded onto MHC molecules and transported to the cell surface, where they are presented for recognition by immune cells. Human MHC-I molecules, also known as human leukocyte antigens (HLA), are transmembrane glycoproteins encoded from the most polymorphic genes (HLA-A, HLA-B and HLA-C) of the genome. Such polymorphisms mainly occur at the binding site of the MHC-I molecule (groove) and the peptide, resulting in allelic variations[6]. This enables each molecule to harbor a large diversity of peptides in its groove and thus allow the cell to have a broad representation of its proteome on the cell surface. However, the number of peptides that can be presented is limited by two factors. Firstly, the number of MHC-I molecules available which varies between ~$1x10^4$ to ~$5x10^5$ depending on the cell type[7]. For instance, on average only about ~$2x10^5$ MHC-I molecules are expressed at the cell surface of normal B and T cells[8], and ~$5x10^4$ to ~1.5 $x10^5$ on ovarian cancer epithelial cells[9]. Secondly, only peptides with the required sequence motifs for HLA binding are presented, leading to a selective sampling of available peptides[10]. While it has been hypothesized that each cell expressing approximately ~$2x10^5$ MHC molecules can present around ~$1x10^4$ different MAPs[11], the specific criteria for their selection remains uncertain. Pearson et al[12]. showed that a restricted number of genes (~58%) contribute to the production of MAPs in normal cells. This finding indicates that the selection of genes for MAP presentation is not random, with abundance being an important factor, albeit not the sole determinant. Thus, the authors proposed a model suggesting that MAP presentation is regulated by a shared group of proteins that can generate MAPs with motifs compatible with most of MHC-I allotypes. Such proteins might possess specific characteristics (protein abundance, translation efficiency, protein length, degradation rate, protein structure), enabling their selective entry into the antigen presentation pathway.

The immunopeptidome is therefore a dynamic repertoire. It has been observed that changes in the cell, such as those induced by treatments[13], diseases[14] or infections[15] can alter the composition of the immunopeptidome and lead to the rapid presentation of these changes on the cell surface[16]. Under such conditions, increased synthesis of the affected proteins results in a modified set of peptides that are selected for presentation by MHC-I molecules. The flexibility of MHC-I molecules to associate a broad but selected set of peptides favors the communication of such changes between the cell and the immune system through its immune surveillance process. As a result, T lymphocytes can recognize and distinguish normal from abnormal cells.

T lymphocytes such as CD8+ T cells mediate antigen recognition through their surface T cell receptors (TCRs) that are able to recognize more than a million distinct MAPs[17]. When the TCR of a CD8+ T cell encounters a foreign MAP, the TCR binds to the MAP derived from pathogen or infected cells and mount an immune response. Upon detection, the CD8+ T cell either directly kills the abnormal cells or secrets small proteins called cytokines to attract inflammatory cells that attack the abnormal cells[18, 19]. Once the abnormal cells are controlled, the repertoire of antigen-specific T lymphocytes, which expanded during the immune response, is stabilized by the process of apoptosis. As a result, most of these T cells are eliminated, while a small number are retained to ensure immunological memory[20]. Immune memory ensures long-term protection against the same antigen, i.e., in case of re-exposure, the memory conferred by the preserved T-lymphocytes can eliminate the threat. This is the principle of vaccination, in which selected antigens are introduced into the body to induce an immune response that protects against a pathogen on subsequent exposure[21].

Taken together, the interactions of CD8+ T cells with MHC-I molecules are key to defining the health of an organism. After their training in the thymus, CD8+ T cells act as patrols that scan MHC-I molecules for abnormal MAPs, thereby eliminating potential threats. Vaccines aim to take advantage of this defense mechanism by mounting immune response to preserve a long-lasting immune memory for future exposure. Thus, it is paramount to prioritize the safest, highly immunogenic antigens that preserve the health of normal cells, eliminating only the abnormality. The use of incorrect or not specific antigens can have detrimental effects on tissues, posing a significant risk to the patient's life. In the context of cancer therapies, there is a focused effort to

identify tumor-specific MAPs for the development of antitumor immunotherapy, particularly in the form of vaccination. Further details on this topic will be explored in Section **1.2.1**.

1.1.1 Generation of the immunopeptidome

The immunopeptidome is the set of short peptides that are generally produced through the degradation of ubiquitinated proteins within cells. This process occurs through the action of the ubiquitin-proteasome system responsible for breaking down proteins into their constituent peptides. Selected peptides resulting from the degradation process are transported to the endoplasmic reticulum (ER), where they are filtered according to their size and affinity for available MHC-I molecules and then loaded onto them. Peptide-loaded MHC-I molecules are transported through the Golgi to the cell surface, where they are anchored to communicate with the immunosurveillance system[22] (Figure 1).

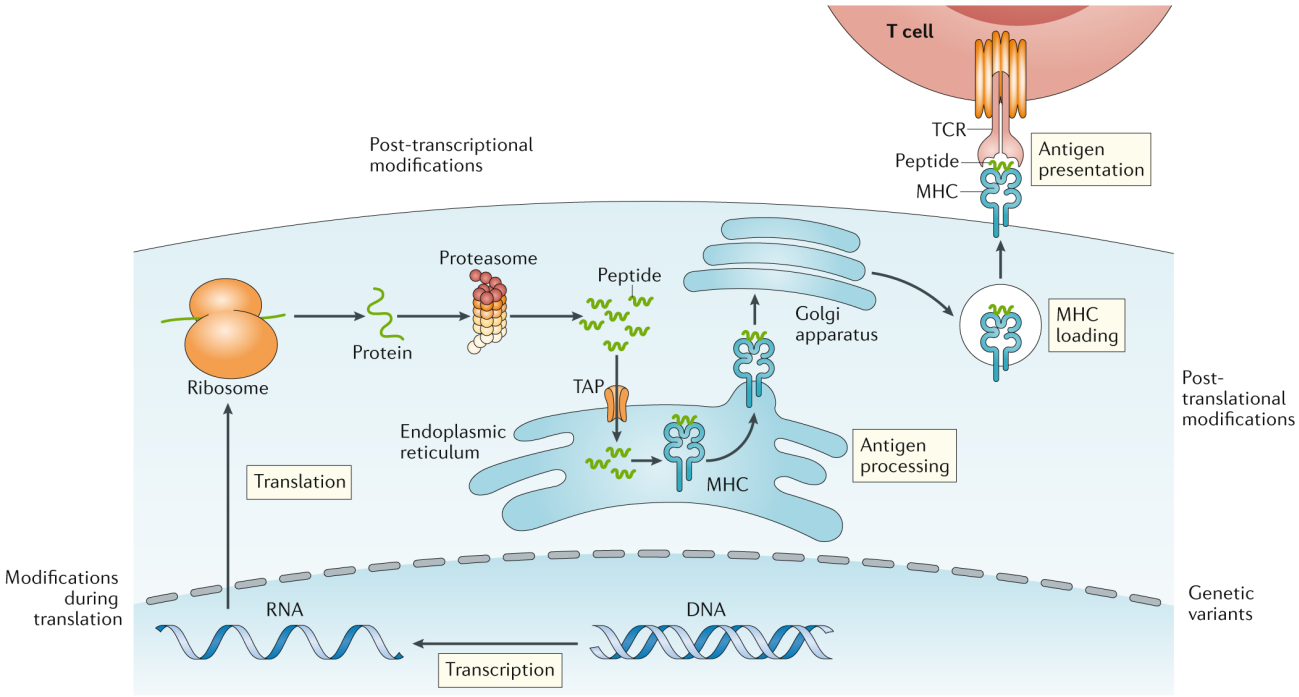


Figure 1. – Antigen processing and presentation by MHC class I molecules.
Peptides are generated from the degradation of ubiquitin-tagged, mistranslated proteins or end-of-life proteins. The peptides are further processed (trimming) by peptidases and delivered to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP). In the ER, peptides are loaded onto available MHC class I molecules, to which they have a strong binding affinity. Finally, the loaded MHC class I molecules migrate through the Golgi to the cell surface.

Adapted with permission from Springer Nature: Nature Reviews Clinical Oncology (Haen, S.P., Löffler, M.W., Rammensee, HG. et al.)[23] © 2020.

1.1.1.1 MAPs source proteins: Self vs non-Self

Under normal conditions, the set of MAPs presented on the cell surface is derived from normal constitutive proteins or "self" proteins. CD8+ T lymphocytes are destined not to react to self-MAPs as they were selected after passing the autoreactivity test in the thymus. Indeed, the thymus plays a key role in the development of the immune system by establishing central tolerance, process by which the immune system becomes tolerant to self-proteins. Medullary thymic epithelial cells (mTECs) and dendritic cells mediate this process by generating a diverse repertoire of self-MAPs from genes expressed in normal tissues[24, 25]. Self-MAPs are then presented to naive CD8+ T cells to test their reactivity. CD8+ T cells that recognize and react against self-MAPs are eliminated. Those that show little or no reaction to the large number of self-MAPs to which they were exposed become functional or mature CD8+ T cells[26-28]. Thus, while the establishment of central tolerance explicitly aims to eliminate CD8+ T cells susceptible to react against normal proteins, it implicitly trains CD8+ T cells to react exclusively against abnormal protein peptides. By establishing central tolerance, the thymus helps to ensure that the immune system can effectively protect the body from foreign invaders, while at the same time preventing it from attacking the body's own cells and tissues.

In pathological conditions, alongside self-MAPs, it is also possible to observe on the cell surface MAPs derived from abnormal or "non-self" proteins. The immune system is designed to recognize and attack substances that it has not developed tolerance for, such as infections. Cells infected by intracellular bacteria, parasites, or viruses may exhibit on their surface foreign or "non-self" MAPs that can be recognized by CD8+ T cells. When CD8+ T cells encounter infected cells that express foreign peptides, they are activated and launch an immune response to eliminate the infected cells and prevent the spread of the infection. In the context of cancer, the occurrence of epigenetic alterations and somatic mutations can lead to defects in DNA, which in turn translate into abnormal proteins that generate abnormal MAPs. The immune system is designed to identify "non-self" MAPs, but antigenic changes in cancer cells caused by individual mutations can be subtle. This presents a challenge for the immune system to distinguish these

altered MAPs from self-MAPs[29]. Furthermore, cancer cells can become invisible to the immune system due to alterations that result in the loss of antigen processing function or major histocompatibility complex (MHC) class I proteins to present antigens, thereby acquiring the ability to grow progressively[30]. Also, cancer cells can evade the immune system by promoting the establishment of an immunosuppressive environment that favors their growth and survival[31]. They achieve this by producing cytokines that not only promote their own growth, but also attract immune regulatory cells that function as effectors of immunosuppression[32]. An example of such cytokines secreted by tumors are vascular endothelial growth factor (VEGF) and transforming growth factor-β (TGF-β). VEGF is a potent cytokine that stimulates the formation of new blood vessels, which provides tumors with the necessary nutrients and oxygen to support their growth, while TGF-β has the ability to suppress the activity of immune cells, such as natural killer (NK) cells and CD8+ T lymphocytes[33]. Finally, the immune system may be compromised or weakened by chronic inflammation, aging, or the presence of other diseases, reducing its ability to recognize and attack cancer cells[34].

1.1.1.1.1 Nature of MAP source proteins

It was long thought that the immunopeptidome was generated solely from the degradation of normal proteins that reached their natural lifespan (retirees). Recent research has suggested that the immunopeptidome may also be derived from rapidly degraded proteins (RDPs)[35]. This idea is supported by the fact that CD8+ T cells can eliminate virus-infected cells within a relatively short period of time post-infection (45 minutes)[16]. This suggests that, although the half-life of a typical protein can range from 9h[36] to 48h[37, 38], depending on various regulatory mechanisms, viral proteins can be rapidly translated and degraded to generate MAPs. Further studies have also shown that MAPs tend to originate from efficiently translated proteins that are prone to degradation, such as disordered proteins or proteins with a high number of ubiquitination sites[12]. Therefore, in addition to the natural lifespan of proteins, their structural conformation may also influence the immunopeptidome.

MAPs are therefore derived from the protein pool of retirees and RDPs[39]. Quantification analysis of synthesized RDPs have shown that they represent between 25[35, 40]- 30%[41] of cell proteins and are distinguished into two groups: short-lived proteins (SLiPs)[4] and defective

ribosomal products (DRiPs)[42]. SLiPs are proteins naturally having a high turnover and undergoing degradation with an average half-life on the order of 8 minutes[38]. DRiPs are defined as proteins that, due to premature termination of translation or misfolding, do not reach a functional conformation and are subject to degradation by the proteasome[43]. Although no DRiPs have yet been isolated in vivo, MS characterization of vaccinia-infected cells demonstrated that MAPs were detected half an hour post-infection[44], probably as result of rapid translation and degradation of viral proteins. Similarly, lymphocytic choriomeningitis virus nucleoprotein-derived NP118 peptide was rapidly detected before its parent protein reached its functional lifetime after infection (>3 days)[45]. Thus, the substantial contribution of DRiPs to the immunopeptidome[46] appears to be critically necessary, as they may be the main source of viral peptides to alert the immune system and rapidly prevent the spread of infection.

1.1.1.1.2 Canonical and non-canonical proteome

The MAPs sources protein (retirees, DRiPS and SLiPs)[47] are encoded by genes located on all chromosomes[48]. It had been assumed that MAPs were derived from conventional genes, i.e., genes for which a specific region in the genome and its function is known. Conventional genes are translated into **canonical** proteins, which according to the Universal Protein Resource (UniProt)[49], are functional, conserved, widely expressed, long and human-verified proteins.

To date, there are more than 20,000 human canonical proteins reviewed in the UniProt database representing only ~2% of our genome. However, the cumulative coverage of transcribed regions assessed with RNA sequencing (RNA-Seq) showed that 74.7% of the human genome can be transcribed[50]. Further studies have revealed that our understanding of transcription and translation is incomplete, as evidenced by the observation of translation occurring beyond the boundaries of annotated protein-coding regions[51]. Thousands of open reading frames (ORFs) have been identified in putative non-coding regions of eukaryotic genomes, such as intergenic regions[52, 53]. In humans, transcriptomics[54-57] (complete set of RNA transcripts produced in the cell) and translatomics[51, 58-60] (complete set of translated transcripts), studies have confirmed widespread expression of proteins deriving from non-coding regions. Some of these proteins have been verified by proteomics (identification of proteins at a point in time)[57, 61-66]. The overwhelming evidence of **non-canonical** translation events often referred to as cryptic proteins,

rules out the idea that such observations are the result of purely random events or experimental artifacts.

The non-canonical proteome is translated from: out-of-frame translation of coding exons (frameshift)[63-66] and alternative ORFs located in allegedly non-coding regions. The alternative ORFs are located in UTRs[57, 59, 62, 63, 65, 66], non-coding RNAs as pseudogenes[61], long non-coding RNAs (lncRNAs)[51, 60, 62, 66, 67] and from intergenic regions[63]. These non-canonical proteins have distinctive properties compared to their canonical counterparts. They tend to be shorter proteins (median size ~57 aa versus ~400 aa for canonical proteins)[57, 63, 65], have lower transcriptional and translational rate[60] and initiate translation on near-cognate codons (differing from AUG by a single nucleotide)[63, 66]. Also, they are predicted to be short-lived as they have a less stable conformation in-vivo[60]. These properties are significantly divergent from those of canonical proteins, leading to misclassification of these novel alternative coding RNAs into non-coding RNA categories[57].

The extent to which non-canonical proteins contribute to the proteome and their functional roles are not yet well defined. Recent studies have suggested that non-canonical proteins may have essential roles in metabolism and cellular regulation[66]. For example, non-canonical proteins orchestrate mucosal immunity during infection[68], enhance muscle activity[69, 70], act in heart function[62] and as regulators of transcription and translation[57], and be essential for cancer cell survival[71]. In addition, studies have confirmed the presence of non-canonical proteins as sources of MAPs, meaning that the non-canonical proteome undergoes the same antigen processing and presentation pathway as canonical proteins. Thus, the cell surface is populated by MAPs derived from both canonical proteins and non-canonical proteins, the latter contributing approximately ~5-10% of the total MAPs [72-76].

The non-canonical proteome has gained a lot of attention lately as it may be instrumental in treating cancer. A recent study has suggested that MAPs derived from non-canonical proteins may be promising targets for cancer immunotherapy[77]. In this study, researchers identified cancer-specific non-canonical MAPs in leukemic cells and immunized mice with these MAPs. This led to strong anti-leukemic responses and reduced leukemic cell growth in the vaccinated mice.

When the researchers examined primary human leukemia and lung carcinoma samples, they also found non-canonical MAPs that were potentially conserved between individuals and could potentially be used as targets for immunotherapy. These findings suggest that non-canonical MAPs may be a promising avenue for developing effective cancer vaccine treatments.

Overall, understanding the origin and regulation of non-canonical proteins, derived from putative non-coding regions of the genome, is important for characterizing immunopeptidome. By elucidating the properties and functions of non-canonical proteins, researchers can better understand not only the role of non-canonical proteins in various biological processes but also their contribution to the immunopeptidome, therefore, their practicality in immunotherapy.

1.1.2 Identification of the immunopeptidome with proteogenomics

Recent advances in high-throughput sequencing and mass spectrometry (MS) have facilitated the efficient characterization of the immunopeptidome (reviewed[78, 79]). Researchers have harnessed genomic and proteomic information for the identification of MAPs, leading to the emergence of a new field called proteogenomics[80]. Briefly, upon sample preparation, immunopeptidome identification follows three major steps: MAPs isolation, MS, and MAPs identification (Figure 2).

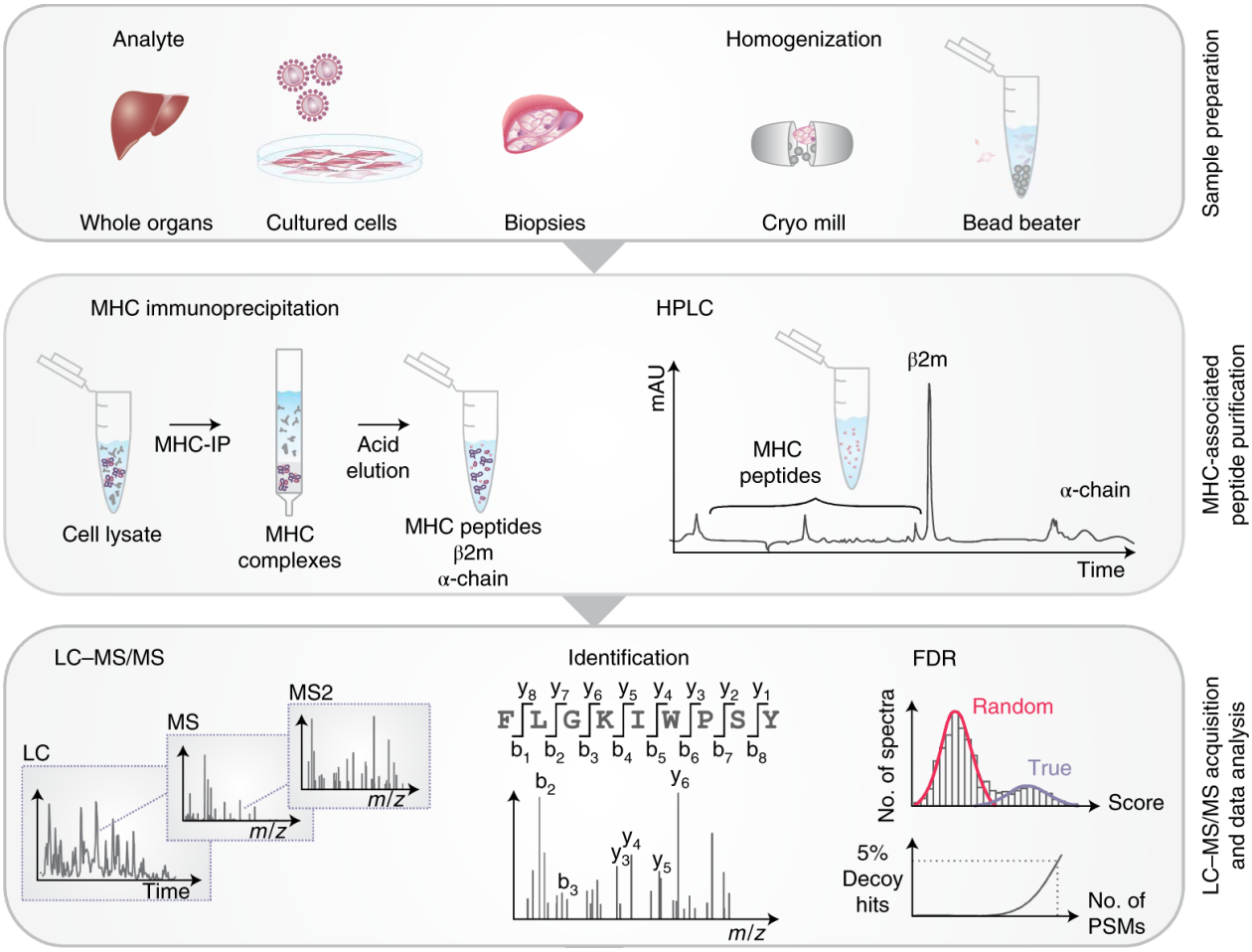


Figure 2. – Immunopeptidome identification steps.
Top panel: sample preparation. MHC I complex purification from whole organs, cultured cells, or biopsies, followed by a homogenization process. Middle panel: purification of MAPs. Immunoprecipitation (using antibodies for MHC molecules) followed by liquid chromatography (LC) for the separation of peptides and MHC molecules. Bottom panel: LC-MS/MS acquisition and data analysis. After LC, the peptides are injected into the mass spectrometer. Peptide sequence identification is performed using database search engines. Finally, the number of peptides is reduced after FDR calculation to control false positive identifications. Adapted with permission from Springer Nature: Nature Protocols (Purcell, A.W., Ramarathinam, S.H. & Ternette, N)[78]© 2019.

1.1.2.1 MAPs Isolation

MHC-I molecules are purified from whole organs, cultured cells, or biopsies, followed by a homogenization process to break down the tissue structure and create a suspension of cell fragments (Figure 2, top panel). Isolation of the processed natural MAPs from the cell lysate is typically performed by a process called MHC-I immunoprecipitation, which provides large

numbers of isolated MAPs[81, 82]. Antibodies specific for MHC-I molecules are used for isolation, followed by several washing steps to remove the unbound mixture, and then acid elution is applied to dissociate the MAPs from the MHC-I molecules. High-performance liquid chromatography (HPLC) is applied to separate the MAPs from the MHC components: α-chain and β2-microglobulin (β2m) (Figure 2, middle panel).

1.1.2.2 Mass spectrometry (MS/MS)

After being prepared, the MAPs are injected into a liquid chromatography (LC) column, where they undergo a separation process that is influenced by the chemical interaction between the MAPs and the stationary and mobile phase of the column. Over time, this interaction affects the migration of the MAPs through the column, leading to a separation of the mixture (Figure 2, bottom panel). The effluent from the LC column is directed into a mass spectrometer (MS), which is used to analyze the mixture of peptides. The MS measures the intensity and mass-to-charge ratio (m/z) of ions present in the effluent to retrieve information about their abundance and peptide sequence.

While there are several types of mass spectrometers available for peptide analysis, LC-MS/MS methods typically perform the following tasks[83, 84]:

a) Ionization: the mass spectrometer has an ionization source where the LC column effluent is nebulized, desolvated and ionized creating charged particles.
b) MS1 scan: ionized peptides, hereafter referred to as precursor ions, are selected one by one to pass through an electromagnetic field to measure their intensity and m/z. As result, an MS1 scan of each peptide is generated, which is a plot depicting the intensity (y-axis) versus m/z (x-axis). From MS1 scan, all that is known about the peptide is its mass and charge.
c) MS2 scan: each precursor ion is fragmented in a collision cell with an inert gas to obtain ion fragments, i.e., the molecule is broken down into ions to identify its amino acid composition. Ions fragments of each precursor ion are then passed through a mass analyzer to measure their intensity and m/z resulting in a MS2 scan. In the MS2 scan, the fragment ions are represented as pairs of b and y ions. The y ions represent the

fragment charge retained at the C-terminus and the b ions the charge retained at the N-terminus. The MS2 scan is then used to identify the amino acid sequence of the peptide.

1.1.2.2.1 MS data acquisition

Depending on the needs of the study there are three main modes to acquire data on a mass spectrometer: data-dependent acquisition (DDA), targeted and data-independent acquisition (DIA) mode. The selection of a data acquisition mode has a direct impact on the quality of peptide/protein identification and quantification, owing to the distinctive characteristics associated with each mode.

The DDA mode is commonly used for peptide/protein discovery (shotgun) in a sample, with the aim of achieving comprehensive coverage. This mode works by first conducting an MS1 scan to identify the most abundant precursors (10-20), which are then individually isolated and fragmented in sequential MS2 scans that are analyzed to determine the peptide sequence[84]. This strategy allows the identification of thousands of peptides/proteins, which provides a broad proteomic overview of the sample, however, it is worthy to mention that this approach also has several drawbacks. First, it has low reproducibility, several peptides can co-elute and appear in a single MS1 scan and DDA stochastically samples only the most abundant peptides. Second, due to the wide dynamic range in protein abundance, low abundant peptides are not identified. Lastly, DDA deliberately samples each peptide only once or twice to enhance peptide discovery within a reasonable timeframe, resulting in relative instead of absolute peptide quantification[84]. Yet, shotgun proteomics is frequently used for comparisons of peptide abundances between samples of interest. This strategy known as label-free shotgun proteomics, involves measuring the relative abundances (MS1 scan intensity) of peptides independently in each sample after injection into the mass spectrometer. Although batch effects may be introduced due to variations in run conditions such as temperature, column conditions, or experimenter during MS runs[85, 86], the intensities of MS1 scans of the same peptide in the samples are compared to identify differences in abundance. In contrast to label-free shotgun, label-based shotgun proteomics are advanced techniques that enable the isobaric or isotope labeling of proteins in each sample. The labeling technique enables a single MS injection of all samples, facilitating the measurement of their

relative quantification. This approach helps minimize the introduction of batch effects, enhancing the accuracy and consistency of the analysis. Isobaric labels refer to chemical groups with identical mass, but different distribution of heavy isotopes in their structure[85]. On the other hand, isotopic labels refer to chemical groups with the same structure, but different masses. Some common examples of label-based approaches include SILAC[87], which uses metabolically isotopic labeling, and TMT[88] or iTRAQ[89], which use chemically isobaric tagging to compare up to 10 different samples simultaneously.

In targeted mode, selected/multiple reaction monitoring (S/MRM)-MS and parallel reaction monitoring (PRM)-MS are two techniques that enable the relative and absolute quantification of pre-defined set of peptides. These peptides must have been previously identified with MS to determine their peptide sequences, their elution time and their m/z value of the precursor ion, which is necessary information in the targeted strategy[90]. Indeed, the mass spectrometer selectively detects only those precursor ions and fragments that correspond to the m/z and elution time of the pre-defined peptides. These techniques allow quantification of low abundance targets and circumvent the imprecision and reproducibility limitations of DDA peptide/protein discovery proteomics[86]. In MRM, peptides are analyzed on a triple quadrupole mass spectrometer to separate ions based on their mass-to-charge ratio. The first quadrupole is responsible for selecting the precursor ions which are then fragmented in the second quadrupole. From the resulting ions, predefined ions are selected for acquisition of the partial MS2 scan by the detector[91]. Unlike MRM, which only measures selected fragment ions, Parallel Reaction Monitoring (PRM) measures all the resulting fragment ions[92]. PRM is typically performed using mass spectrometers such as Orbitrap or Time of Flight (ToF), which can measure all fragmented ions. Both MRM and PRM involve selecting and fragmenting the same precursor ions multiple times to obtain more accurate quantification of a smaller number of peptides, as compared to DDA.

Finally, the DIA mode aims to integrate the strengths of the two previous modes to achieve accurate and reproducible quantification. Contrary to DDA which stochastically fragments only the most abundant peptides to generate MS2 scans, DIA acquires and fragments all precursor ions, thus being able to detect and identify even peptides at lower concentrations[93]. In DIA mode,

in each time cycle the mass spectrometer focuses on all precursors in a wider m/z window and acquires a single MS2 scan of all precursors in that window. The resulting MS2 scan is considered a multiplexed scan, as it contains all fragment ions of the peptide mixture in the respective window. Once a mass window has been analyzed, the mass spectrometer slides to the next window to collect the MS2 scan in that window[93]. Thus, DIA is a powerful method for sampling all peptides, including less abundant ones within selected mass windows, and can offer superior accuracy and reproducibility compared to DDA. However, the massive amount of data generated and the multiplexed nature of the MS2 scan make data analysis in DIA challenging. Compared to DDA, DIA requires more elaborate and sophisticated algorithms to deconvolute the complex MS2 scan generated for subsequent peptide identification and quantification[94]. Currently, the most common method for generating DIA data is Sequential Windowed Acquisition of All Theoretical Fragment Ions (SWATH), in which the mass spectrometer divides the mass range into small m/z windows of 20 or 25 Daltons (Da) and slides with 2-4 seconds cycle time by precursor acquisition window[95].

1.1.2.3 MAPs identification

To identify the immunopeptidome with MS, the final step involves determining the corresponding MAPs sequences. There are three primary methods used for analyzing and interpreting MS2 scans: de novo sequencing, database searching, and a hybrid approach that combines de novo sequencing with database searching.

The *de novo* sequencing method allows discovering the peptide amino acid sequence by calculating the mass differences between the intensity of neighboring peaks (b and y ions) in the spectra. Each mass difference is compared with a table of the molecular mass of the known amino acid to deduce the amino acid that best fits each of the given positions until the peptide sequence is identified. In recent years, several de novo peptide sequencing algorithms have emerged, which enable the direct identification of peptide sequences from MS2 scans. These algorithms utilize a diverse range of algorithmic techniques, including probabilistic graphical models, as seen in pepNovo[96]; hidden Markov models (HMMs), as utilized by pepHMM[97], dynamic programming, as used by PEAKS[98]; and neural networks, as implemented in DeepNovo[99]. Regardless of the potential of de novo sequencing methods, they are not yet widely used in the proteomics

community as their effectiveness is often hindered by various challenges[100]. These include incomplete fragmentation of peptides in the MS2 scan, the presence of ions other than b and y ions, the inability to differentiate between b and y ions, the presence of isomeric amino acids (such as leucine and isoleucine), and the similarity in mass between certain amino acids (such as lysine and glutamine)[83, 100, 101]. Accordingly, it is reasonable to state that the success of these approaches for peptide sequence determination is highly dependent on the quality of fragmentation in the MS2 scan.

The second approach is database searching, in which peptides are identified using an integrated set of algorithms known as database search engines. Tools such as Mascot[102], Peaks[103] or MaxQuant[104] query known protein sequence databases which may come from public repositories such as UniProt. The search consists of selecting peptide sequences from the database whose theoretical mass matches to a precursor ion mass obtained by MS. From the selected peptides, theoretical MS2 scans are generated and compared with the MS2 scan obtained from the precursor ion using a similarity function. This comparison allows for the calculation of a similarity score, which provides a measure of how closely the theoretical MS2 scan matches the experimental MS2 scan. Thus, the highest score defines the peptide-scan-match (PSM) i.e., the peptide sequence to be assigned to each precursor ion. Finally, as a check for correct identifications, the acquired MS2 scan is also searched against the inverted sequence database (decoy database) to determine the peptide false discovery rate (FDR)[105, 106]. The FDR is calculated from the percentage of misidentifications (PSM in the decoy database) at user-specified thresholds and is used to control the number of false positive identifications, resulting in fewer PSM being accepted[107]. The FDR is computed as the ratio of the number of identifications made in the decoy database and the number of identifications made in the target database: FDR $= \frac{decoys}{targets}$. While the target-decoy search strategy is commonly used for estimating correct peptide sequencing, it's important to note that the size of the database directly affects FDR calculation. This is because the larger the database (target and decoy), the greater the probability that the best match to the MS2 scan is incorrect[106]. The use of a larger database significantly expands the pool of potential candidate peptides, increasing the likelihood of encountering peptides with similar or overlapping mass-to-charge ratios and fragmentation patterns. This higher degree of

similarity among candidates raises the risk of finally select the incorrect peptide sequences. Consequently, the result is a low number of peptide identifications that have a high probability of being false.

Finally, a hybrid approach for the identification of peptide sequences in proteomics has been proposed, which aims to integrate the best of the other two other approaches: de novo sequencing and database search methods[108]. The hybrid approach is a tag-based approach as it uses de novo sequencing to identify and label well-resolved short sequence fragments as short as one amino acid[109]. The tagged short sequences are matched against the database to identify and consider only those peptide sequences containing the tag along with the correct flanking masses to identify the best candidate[100]. A current widely used hybrid approach is PEAKS DB[103] which proposes the use of *de novo* sequencing as the initial step in peptide identification. The goal of this strategy is to identify a short list of proteins (up to 7000) that encompasses most peptide tags. As a result, the short list decreases the search space and is used for scoring and validation involving the target-decoy approach. However, a slight change has been introduced to the target-target strategy, termed decoy fusion. This approach consists of generating a decoy protein for each target protein and combining them into a single entry. In this way, the resulting short list of proteins is free of any bias toward the target or decoy entries. In addition, peptide identifications from scans with high confidence *de novo* sequencing tags but no matches in the target database, are obtained from novel or modified peptides complementing the database searching. Yet, the de novo search lacks FDR control; therefore, additional validations will need to be applied to ensure the quality of the resulting peptide identification.

In summary, the quality of the MS2 scan is a crucial factor in all three approaches to MAPs identification. De novo sequencing is particularly dependent on scan quality, while the completeness of the sequence database is critical for peptide identification through database search. Regardless of the method used, such limitations can significantly impact the accuracy and number of identifications that can be made. The completeness of the database depends on the presence of splice variants, single amino acid variations (SAAVs) and post-translational modifications (PTMs) inherent to the sample studied and which therefore may not be included in standard reference databases such as UniProt[100]. PTMs are chemical modifications that occur in

proteins capable of altering their structure, function, and interaction with other proteins in the cell. There are several types of PTMs, including phosphorylation, acetylation, methylation, glycosylation, and ubiquitination, among many others. The database search can be parameterized to be tolerant to mass errors to consider PTMs as variable modifications. However, database searches only include a restricted set of common PTMs preventing the identifications of less prevalent modifications challenging[110]. In turn, the process of identifying the sequence of modified peptides becomes much slower when many PTMs are involved. This is because fragment ion mass shifts caused by each specific PTM must be considered, which can make the search space larger. Furthermore, peptides carrying PTMs may not follow the same rules for breaking bonds between amino acids, thus presenting an additional challenge for their sequence identification. As a result, identification of the specific PTM sites can be more difficult than identifying unmodified peptide sequence based on the mass and intensity[111]. To sum up, PTMs greatly augment the complexity of the peptide identification process, which could be addressed with the recent advancements in deep learning for de novo sequencing, ultimately leading to more precise detection of these modifications[103, 112].

In immunopeptidomics analysis, MAPs have previously been identified bearing PTMs[113, 114]. Some of these PTM-bearing MAPs can influence immunogenicity as their presence or absence informs the immune system of inner cell changes. For instance, changes in the phosphorylation status can be caused by inflammation, infection or oncogenesis[115-120]. However, the detection of MAPs carrying PTMs might be restricted by the limited detection capability of mass spectrometry instruments. For example, detection of phosphorylated MAPs is challenging due to the dynamic range of mass spectrometers, as these MAPs represent only ~1% of the immunopeptidome and are therefore less abundant than their unmodified counterparts[110, 114]. Thus, enriching modified MAPs before injection to the mass spectrometer could significantly enhance both the sensitivity and throughput of their identification[121].

1.1.2.3.1 The non-canonical proteome and the emergence of proteogenomics

To date, immunopeptidomics identification thought MS is mainly based on database searching. However, protein databases such as UniProt only contain canonical proteins that have been comprehensively collected and reviewed[49]. Gene annotation evaluates the transcripts and

proteins detected by using various assumptions that rule out or misclassify non-canonical proteins (often annotated as non-coding RNAs). Homology and conservation evaluation of observed ORFs are needed for their classification, along with other properties such as (1) the presence of the ATG initiation codon thought to be the unique codon recognized by the ribosome; (2) several exons in the structure; and (3) have a minimum ORF length (e.g., length requirement of at least 100 amino acids—aa)[122]. In general, gene annotations remain a difficult task to unambiguously annotate RNA as protein-coding or non-coding due to its confounding properties[123]. As result, the inclusion in standard databases of novel proteins from the non-canonical proteome that mostly translates from short ORFs derived from canonical genes in different frames and non-coding regions is being prevented[78].

Proteogenomics aims to bridge the gap between transcription and translation for protein identification by generating customized protein sequence databases from genomic and transcriptomic information. With this approach, robust identification of canonical and non-canonical peptides from MS proteomic data can be used to provide evidence for translation at the protein level helping to refine gene models[80]. Conventional databases such as UniProt (commonly used for peptide identification) are replaced by customized databases that aim to integrate the whole transcriptome landscape in the sample. Identification of novel peptides is performed from MS searches of databases containing both the canonical and non-canonical proteome predicted from high-throughput RNA sequencing[80]. Thus, proteogenomics has led many studies to identify proteins that were previously overlooked.

1.1.2.3.2 Database Generation from Sequencing Data

Although the inclusion of non-canonical proteins in the database is mandatory to facilitate the identification of previously unnoticed proteins, this process should be carried out with caution. To identify non-canonical translations, previous studies have used RNA sequencing data to generate custom databases that included translation of 6 frames of each RNA seq read[72]. The decision to translate 6 frames in the case of unstranded RNA seq reads is explained by the fact that RNA sequencing lacks information on the start and termination of ORFs and the translation frame. Therefore, the custom databases ended up oversized (~1 GB or $55x10^6$ sequences compared to the size of the known human proteome reference ~25 Mb or $1x10^5$ sequences), as

many sequences, at least 5 of the 6 translation frames, could be false translations[124]. Inflated databases hinder the accuracy of MS peptide identification as the risk of false discoveries increases with the size of the database[125]. Indeed, the more theoretical MS/MS spectrum scores for experimental MS/MS the higher the probability that the best match would be incorrect[106].

Aware of this limitation, researchers have developed strategies to generate the customized databases in a manageable size for database search engines. Some strategies involve the use of transcriptome assembly tools such as StringTie[126], Cufflinks[127] or Trinity[128] to reconstruct the transcriptome taking into account alternative transcripts and splicing from RNA sequencing. The transcriptome reconstruction aims to reduce the space of sequences to be translated into 3 or 6 frames, depending on the type of RNA-seq library, to be included in the database. This strategy has been used, for example, to detect the adenovirus proteome in infected human cells[129]. The authors reconstructed first the transcriptome of infected human cells that was further filtered by transcript abundance and length. The transcripts that passed the filters were then translated in 6 frames translations to finally composed the customized database. Although the size of the resulting custom database should be smaller than if each of the RNA reads were translated at 6 frames; it is still populated with irrelevant sequences from the reserved transcripts translated into 6 frames.

In the context of cancer immunopeptidome, attempts have been made to reduce the size of the database to include non-canonical sequences and thus discover suitable antigenic targets. The identification of virtually suitable antigens was carried out for example from the definition of the cancer cell-specific immunopeptidome. In these studies, the databases included only canonical and non-canonical sequences translated in 3 or 6 translation frames, depending on the event from which the transcript was identified[130-132]. Other studies, for example, identified first cancer-specific RNA-seq reads from subtraction of RNA-seq reads from normal cells (mTEC)[77, 133-135]. Small custom databases were generated from 3-frame translations of cancer-specific RNA-seq reads concatenated into contigs. These approaches have identified true potential targets that significantly improved survival in mice[77] and had good predicted survival in humans[133, 134].

Although proteogenomics has proven useful for the direct identification of immunogenic antigens, personalized databases continue to be populated with irrelevant sequences from 3- or 6-frame translations. Thus, it is still necessary to develop other strategies to warrant the accuracy of the identified peptides.

1.1.2.3.3 The advent of ribosome profiling

Ribosome profiling (Ribo-seq) is a deep sequencing technique for ribosome-protected mRNA fragments that elucidates the actual transcriptome being synthesized into proteins[136]. The ribosome-protected fragments enclose a short portion of RNA (~30 nucleotides) which corresponds to the ribosome messenger RNA (mRNA) template. Using different translation inhibitors, ribosomes can be stalled at initiation (Harringtonine) or elongation (Cycloheximide) sites to elucidate the initiation codons or the body of the proteins being translated, respectively[137, 138]. Thus, Ribo-seq provides quantitative information about mRNA transcripts concerning whether they are highly or poorly translated, their reading frames, their start and stop codons and their translation rate[139]. Ribo-seq leads to identify the precise regions of the human and mouse genomes being translated, exposing the pervasiveness of translation outside of the annotated protein-coding genes[66, 140]. Therefore, the integration of Ribo-seq for the generation of customized databases facilitates the inclusion of the canonical and non-canonical proteome actually translated by restricting the size of the search space. The resulting customized databases would therefore have desirable sizes to be manageable by the database search engines (Peaks, Mascot, MaxQuant); thus, improving the number and quality of peptides identified.

Different strategies have already been developed to help identify the set of translated sequences from Ribo-seq to generate customized databases[76, 141, 142]. As lately the immunopeptidomics studies are performed in immunotherapeutic contexts, the use of proteogenomic approaches requires robust and accurate identifications. Customized databases generated from Ribo-seq enabled the identification of high-confidence MAPs derived from non-canonical proteins. In fact, Ribo-seq itself provides an additional layer of translational evidence for peptides identified in MS experiments[73-76]. Proteogenomic approaches that leverage Ribo-Seq, RNA-seq, and MS of matched normal and tumor samples hold great promise for the discovery of safe antigens for cancer immunotherapy. When no matching ribosomal profiling is

available, immunopeptidomics can now benefit from recent efforts to standardize ORF annotations from Ribo-seq data. This new catalog includes more than seven thousand human Ribo-seq ORFs in reference databases[143]. These Ribo-seq annotations within existing reference gene annotations, including Ensembl/GENCODE, HUGO Gene Nomenclature Committee (HGNC), UniProtKB, HUPO/HPP, and PeptideAtlas, would facilitate searches in a more comprehensive protein database. Ribo-seq has been used in cancer research, including our own studies, to generate protein databases that enable peptide identification[73, 74, 144]. Therefore, it would be beneficial to employ this Ribo-seq data to normalize ORF annotations encompassing aberrant cancer-specific ORFs that would ultimately facilitate the discovery of useful antigens for immunotherapy.

In general, despite the relative low throughput in MS-based approaches, they allow physical confirmation of the MAP presentation at the molecular level. However, the use of MS for immunopeptidome identification must consider some main precautions[23, 145]. Firstly, to improve the coverage of identifications it is necessary to use a significant amount of material (cells, tissue). Second, some peptides may need additional validations, as there may be confounding factors for peptide sequence identification, such as post-translational modifications or isobaric amino acids (leucine and isoleucine amino acids cannot be differentiated). Third, the search database must meet two important conditions: it must be complete and of manageable size to avoid false positive identifications[80]. Accordingly, in **Chapter 2**, a proteogenomic approach using Ribo-Seq, RNA-seq and MS aiming to unravel the proteome of cancer cells presented on the cell surface through MHC I molecules is presented. In this study, we conclude that cancer cells present MAPs derived from non-canonical proteins that are only accessible at the level of the immunopeptidome and not at the level of the proteome.

1.2 Immunotherapy

The immunopeptidome is the repertoire of MAPs produced through the degradation of proteins to reflect the intracellular state. Therefore, each MAP is a direct representation of intracellular events occurring in healthy or malignant cells. Recognition of MAPs reflecting intracellular infection, mutations, and disease by CD8+ T cells ensure a specific immune response to eliminate the threat. In cancer, cells present abnormal MAPs (hereafter referred to as Tumor Antigens - TA) derived from protein alterations caused by mutations, normal transcripts overexpression or unconventional transcripts expression. Therefore, cancer immunotherapy aims to harness the power of the immune system to enhance antitumor responses through vaccines targeting cancer-specific MAPs.

1.2.1 Cancer vaccines

Cancer vaccines treat cancer by educating the body's natural defenses by stimulating T cells to recognize and destroy cancer cells. Thus, the main challenge for this type of immunotherapy lies in identifying specific and highly immunogenic antigens. Antigen specificity in cancer is crucial, as CD8+ T cells are unable by themselves to differentiate between benign and malignant tissues if both tissues express the antigen. When antigen specificity for cancer cells is lacking, CD8+ T cells may cause unwanted toxicity in benign tissues presenting the same antigens, which is why the uptake of this type of therapy remains limited[146].

To maximize the clinical benefits of vaccines for cancer treatment, several challenges must be addressed, in addition to the fact that cancer cells must express sufficient levels of antigens capable of stimulating an immune response[147]. Cancer cells undermine the potential of immunotherapy to induce long-lasting protection by altering the expression of the antigen processing machinery, establishing immune escape mechanisms, and utilizing the immunosuppressive tumor microenvironment (TME). In fact, cancer cells often reduce or eliminate the Major Histocompatibility Complex Class I (MHC I) antigen presentation machinery, since MHC I molecules are not essential and can actually be harmful to their survival[30]. Cancer cells may also produce cell surface-anchored proteins called checkpoint proteins. Binding of proteins such as PD-L1 or B7-1/B7-2 to cell surface-anchored T-cell proteins, such as PD-1 and

CTLA-4, respectively, slows down T cells, thus impeding the immune response. Finally, the TME, which surrounds and supports the growth of the tumor, can also hinder the effectiveness of immunotherapy by causing activated T cells to become exhausted or dysfunctional rendering them incapable of reaching and effectively attacking the tumor[31].

Therefore, the success of vaccines for cancer treatment depends on the identification of tumor antigens capable of eliciting a strong immune response, as well as the use of strategies that facilitate their delivery[148]. In recent years, cancer vaccines have shown great promise when used in combination with immune checkpoint blockade therapies such as anti-CTLA-4 and anti-PD-1[149]. These therapies work by targeting and blocking proteins that help to regulate the immune system's response to infections and abnormal cells, thereby boosting the immune system's ability to recognize and attack cancer cells. A recently completed study (November 2021), patients with human papillomavirus-induced tumors treated with nivolumab, an antibody directed against the immune checkpoint PD-1, and tumor-specific antigen vaccines. As result, patients had a 33% overall response rate and a median overall survival of 17.5 months (clinical trial NCT02426892)[150] compared to treatment with PD-1 inhibition alone in similar patients. More targeted efforts against various cancers are currently underway to evaluate the combination therapy efficacy of various checkpoint blockade modulators with specific vaccines[151, 152]. In melanoma, randomized clinical trials such as NCT04526899 and NCT03897881, aim to evaluate the efficacy, tolerability, and safety of tumor-associated antigens in combination with the anti-PD-1 antibodies cemiplimab or pembrolizumab. Recently, Moderna and Merck (last December 20, clinical trial NCT03897881), announced that stage III/IV melanoma patients treated with specific antigens and pembrolizumab had a statistically and clinically significant improvement in reducing the risk of recurrence or death by 44% versus control patients treated with pembrolizumab alone. Hence, these recent and promising results demonstrate the desirability of personalized therapies in conjugation with checkpoint inhibitors.

1.2.2 Antigenic targets presented on the surface of cancer cells.

The main actors of cancer vaccines are tumor antigens that should be capable of stimulating a rigorous immune response. Tumor antigens or TAs can be further classified into two categories

according to their genetic mechanism of expression and recognition by T cells: Tumor-Specific Antigens (TSAs) and Tumor-Associated Antigens (TAAs).

1.2.2.1 Tumor-Specific Antigens – TSAs

TSAs are MAPs that derive from the following types of translations: ORFs with nonsynonymous mutations in cancer (hereafter referred to as mutated TSAs or mTSAs); ORFs aberrantly expressed in cancer but normally silent in normal cells or from genes restrictedly expressed in immune-privileged cells (hereafter referred to as aberrantly expressed TSAs or aeTSAs) (Figure 3.a).

1.2.2.1.1 Mutated TSAs or neoantigens – mTSAs

mTSAs are very interesting targets as they can provoke tumor regression[153]. In many tumors, nonsynonymous somatic mutations resulting from genomic instability of cancer cells are introduced into proteins. MAPs encoded by these mutated sequences unique to malignant cells are called mTSAs. Research groups have primarily focused their attention on mTSAs as highly non-synonymous mutation burden in tumors has been associated with response to immune checkpoint inhibitors and CD8+ tumor infiltration; thereby suggesting that mTSAs are important targets of CD8+ T cells[154]. Although mTSAs have potential as CD8+ T cell targets, they suffer from two major drawbacks. Firstly, their identification is scarce[155], and many mTSAs may not be sufficiently translated to generate effective CD8+ T cell targets. For example, in melanoma cancer, despite having a high mutational load[156], very few mTSAs have been described that can be recognized by T cells[157]. Secondly, mutations arising from tumor-specific DNA alterations, such as single-nucleotide variants, insertions, deletions, or fusions, are often associated with certain genes expressed in cancers but are transient mutations that produce patient-specific antigens[23]. Furthermore, due to tumor heterogeneity, the quality of the antitumor response is influenced by mTSAs present in clonal cells rather than those present only in subclonal cells[158], which adds complexity to the mTSA selection method. Consequently, successful cancer immunotherapy involving mTSAs would most likely require the development of personalized cancer vaccines, as they may offer greater specificity and immunogenicity as long as these strategies include the enhancement of T-cell reactivity to mTSAs[158-160]. In this regard, in 2020, Ott et al.[161] presented a clinical trial of personalized vaccines (a cocktail of mTSAs) combined with an immune checkpoint

inhibitor (nivolumab, to correct PD-1 immunosuppressive barriers). mTSA-specific T cell responses were observed post-vaccination in all the patients, meaning that T cells were able to enter the tumor and mediate cell killing with no severe adverse reactions. These results favor the development of personalized vaccines as they demonstrate that mTSAs are safe and immunogenic targets, even though mTSAs may be difficult or impossible to find in tumors with low mutational load[162].

1.2.2.1.2 Aberrantly expressed TSAs – aeTSAs

aeTSAs are MAPs derived from cancer-specific translations resulting from genetic or epigenetic alterations absent or lowly expressed in normal cells. Alterations in transcription and translation factors, signaling pathways and ribosomal proteins can impair the translation process and lead to changes in the entire proteome[163, 164].

aeTSAs can be coded by any region of the genome, including non-canonical regions, which appears to be the main source of targetable TSAs[77]. Therapies targeting these aeTSAs, which are expected to be non-immunotolerant and to elicit high affinity and avidity in T lymphocytes, emerge to be highly effective. Unlike mTSAs which are likely to be a personalized immunotherapy, aeTSAs are more numerous and can be shared between tumors[77]. Therefore, aeTSAs appear to be a desirable target for a single vaccine that could target various types of cancer, including those with a low mutational burden. Further classification of aeTSAs includes some MAPs referred to as cancer-testis antigens (CTAs) or cancer-germline antigens (CGAs).

1.2.2.1.3 Cancer Testis Antigens – CTAs

CTAs are non-mutated aeTSAs resulting from germline genes specifically expressed in tumors and germline cells due to DNA demethylation[147]. In cancer, these antigens can be recognized by T cells and be safe targets since their source germline genes are not expressed in normal tissues. In germline cells, the expression of germline genes is of no consequence due to their lacking expression of HLA genes which renders these cells unable to present antigens to T cells. Such is the case of the melanoma antigen family A1 gene (MAGE-A1) which was the first gene identified in the human genome encoding an antigen able to be recognized by the T cells[165]. This gene is

expressed in a variety of tumors but is not expressed in normal tissues other than male germ line and trophoblast cells.

Owing to their aberrant expression in up to 40% of cancers[166] and thus their potential use as biomarkers and immunotherapeutic targets, cancer testis (CT) genes have been getting more attention. CT genes have been further classified according to their heterogeneous gene expression to identify the CT genes most suitable for cancer vaccines[167]. In general, CTAs represent promising targets, as they can be shared between different tumors without affecting normal tissues. In this regard, great caution must be exercised when validating these antigens or any other type of antigen to ensure that they are not present in healthy tissues, as a substantial risk of adverse side effects may be incurred[168].

1.2.2.2 Tumor-associated Antigens – TAAs

TAAs are MAPs that are mainly characterized by being derived from i) genes expressed only in tumor cells and in the normal tissue of origin, referred to as differentiation antigens; or from ii) genes overexpressed in cancer cells compared to normal cells (Figure 3b). Therefore, they are "self" MAPs as they are presented in normal tissues but are expected to be presented at higher levels in tumors. TAAs are the product of transcription or translation changes in cancer induced by neoplastic transformation, and visibly are less attractive targets due to their lack of tumor specificity[23]. Expression comparisons between normal and malignant cells for the identification of TAAs are often based on transcriptome data and MS confirmation of peptide presentation in the samples. T lymphocytes are thought to recognize more rapidly and efficiently the more abundant antigens presented on the cell surface[169]. Hence, tumor cells are expected to present many more of these antigens than normal cells, providing an opportunity for T cells to attack only tumor cells. Sor far, most of the differentiation antigens documented have been found on melanoma cells, where T cells recognize and attack tumor cells and normal melanocytes[170]. Consequently, patients with melanoma may develop vitiligo due to spontaneous T-cell response to differentiation antigens, often with a non-severe effect and associated with a good prognosis[171, 172]. While some immunotherapies targeting TAA have shown clinical response with no serious side effects[173], others have exhibited target toxicity caused by low target expression in normal tissues[174].

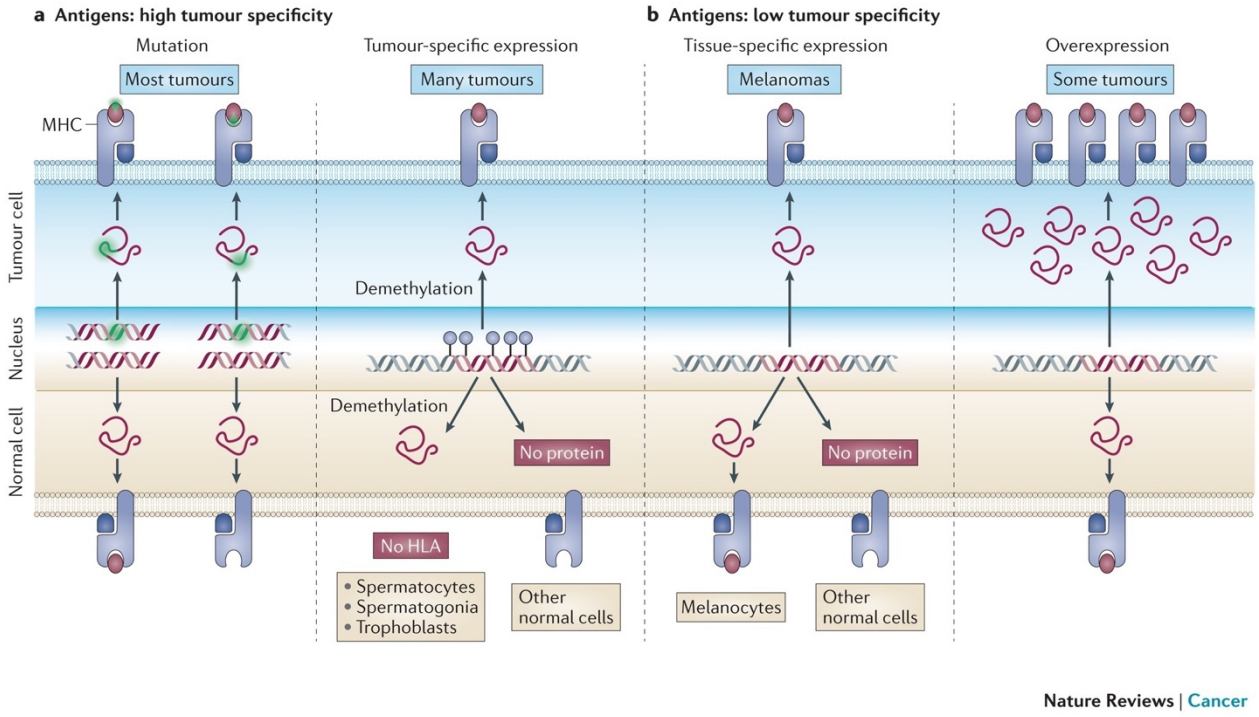


Figure 3. – Tumor Antigens classification.

a) TAs can be Tumor-Specific Antigens (TSAs) that derived from the following translations: mutated DNA sequences not or poorly expressed in normal cells (mTSAs); from the aberrant expression of transcripts normally silent in normal cells (aeTSAs).

b) TAs can also be Tumor-Associated Antigens (TAAs). These antigens are derived from differential expression between tumors and normal tissue of origin. Also, they can derive from genes overexpressed in tumors compared to normal tissues.

Adapted with permission from Springer Nature: Nature Protocols Reviews Cancer (Coulie PG, Van den Eynde BJ, van der Bruggen P and Boon T)[147] © 2014

1.2.3 Identification of Tumor-Specific Antigens using proteogenomics

Proteogenomics has revolutionized cancer immunotherapy research by leveraging transcriptomes (RNA-seq), translatomics (Ribo-seq) and proteomics (MS) in the search for tumor antigens (Figure 4). These approaches are even considered for use in the clinical laboratory to assist with the characterization of cancer biology and facilitate clinical proteogenomics to match effective treatment[175].

Furthermore, recent studies have carried out proteogenomics-based immunopeptidomics analysis to discover selectable antigens in melanoma[74, 157, 176], ovarian cancer[134], leukemia[77, 133, 176], lung cancer[77], B cell lymphoma[176], among others. Even though most efforts to find TSAs

initially focused on MAPs encoded by mutated exons, few mTSAs have been validated by mass spectrometry. Indeed, most DNA mutations are not shared across tumors[134, 157] suggesting that mTSAs are the ideal targets for the development of personalized treatments. In contrast, recent studies have identified aeTSAs derived from non-canonical non-mutated proteins that present themselves as attractive targets as they are highly shared and overexpressed in the tumor, leading to encouraging preliminary results in preclinical models[77, 133]. Thus, aeTSA emerge as worthy to concentrate efforts for the development of vaccines against cancer.

Although these studies have in common the use of proteogenomics, each group follows its heuristic methods and arbitrary assumptions for the detection, selection, and prioritization of the most suitable targets. Hence, validation of a MAP as a real TSA in the discovery process should be a widely available and accepted consensus that may require several lines of evidence for specificity and immunogenicity.

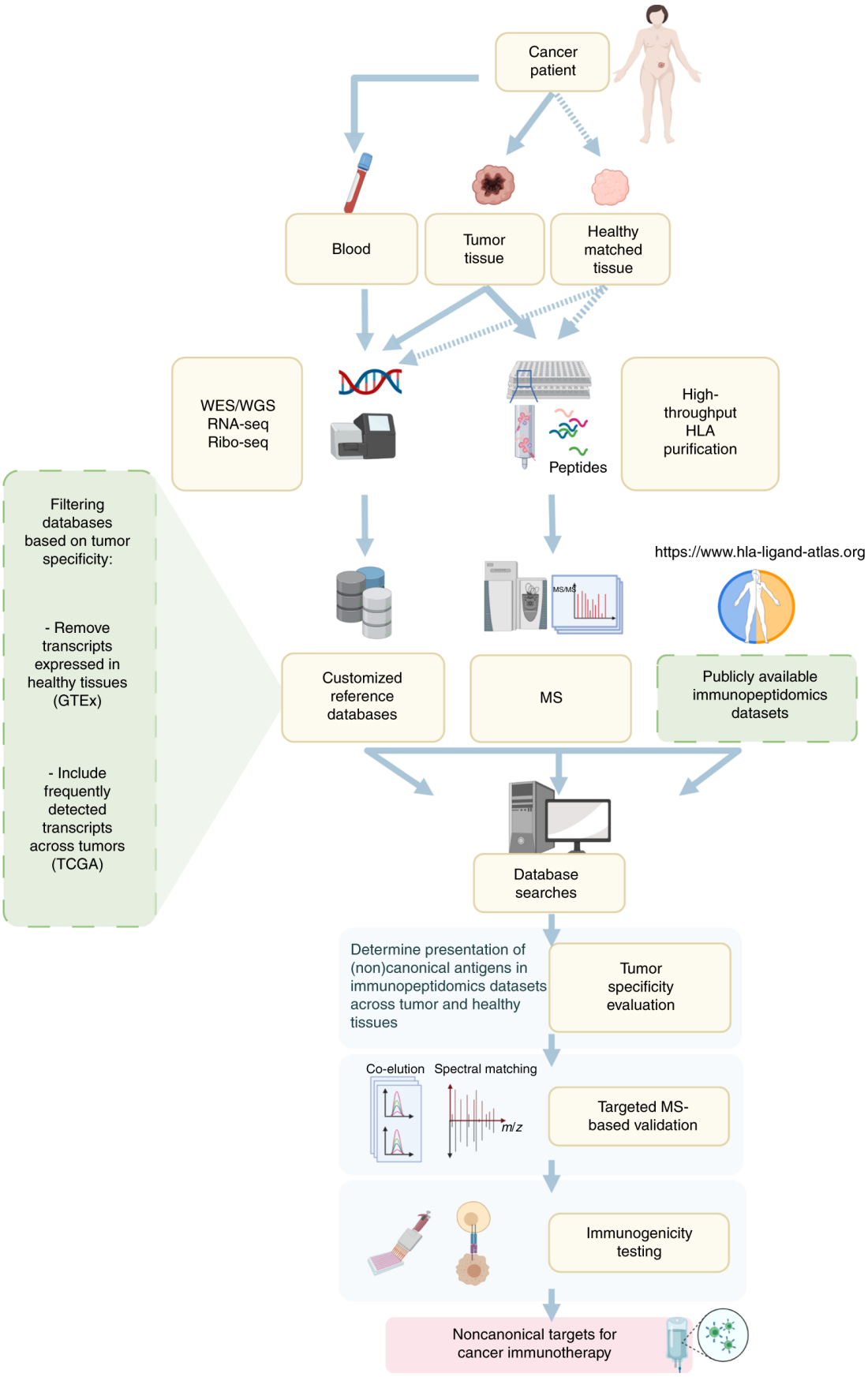

Cancer patient
Blood
Tumor tissue
Healthy matched tissue
WES/WGS
RNA-seq
Ribo-seq
Peptides
High-throughput HLA purification
Filtering databases based on tumor specificity:
- Remove transcripts expressed in healthy tissues (GTEx)
- Include frequently detected transcripts across tumors (TCGA)
https://www.hla-ligand-atlas.org
MS/MS
Customized reference databases
MS
Publicly available immunopeptidomics datasets
Database searches
Determine presentation of (non)canonical antigens in immunopeptidomics datasets across tumor and healthy tissues
Tumor specificity evaluation
Co-elution
Spectral matching
m/z
Targeted MS-based validation
Immunogenicity testing
Noncanonical targets for cancer immunotherapy

Figure 4. – Proteogenomic approach for the identification of canonical and non-canonical immunotherapeutic targets.

Tumor tissue (and, if available, healthy tissue) from the patient is collected to be processed for MAP extraction after MAP isolation, immunoprecipitation, LC and MS processes. In parallel, tumor and healthy tissue are sequenced (WES, WGS, RNA-seq/Ribo-seq). These data are used to build a customized database that will include canonical and non-canonical sequences in cancer, also including tumor-specific non-synonymous mutations. The database is used to identify the peptide sequence of MAPs in the cancer sample, but additional validation is necessary to detect tumor-specific peptides. To gather further evidence supporting the presentation of MAPs, confidence in their accurate sequencing must be assessed using targeted MS approaches. Finally, MAPs that pass all filters should be exposed to immunogenicity assays for the design of effective vaccines. Adapted with permission from Springer Nature: Nature Biotechnology (Chong, C., Coukos, G. & Bassani-Sternberg, M.)[177]© 2021

1.2.3.1 Tumor-Specific Antigen Validation

The identification and prioritization of TSAs for clinical applications should follow thorough validations. In clinical studies involving the adoptive transfer of antitumor T cells, another type of immunotherapy, some examples illustrate why antigen selection and validation require great caution. Nine patients (7 with metastatic melanoma, 1 with synovial sarcoma, and 1 with oesophageal cancer) were treated with T cells engineered to react against a peptide (CTA) from the MAGE-A3 gene; three patients presented severe brain toxicity and two of them died[168]. Three important conclusions emerged from this study. First, engineered T cells in the mouse were harmful in humans as they were able to react against a peptide from another human gene (MAGE-A12) that differs in one amino acid. Second, the homology of the MAGEA family in the mouse model differs from that in humans, which precludes observing the same reaction in the mouse and thus alerting to the possible outcome in humans. Third, it revealed the unrecognized expression of MAGE-A12 in the human brain causing the TCR-mediated inflammatory response that killed patient's neuronal cells and led to lethal toxicity. In another study, two patients (with myeloma and melanoma) treated with T cells engineered against a peptide from the MAGE-A3 gene developed cardiogenic shock and died. In this case, the T cells recognized an unrelated peptide derived from the striated muscle-specific protein titin[178]. These two examples highlight first the importance of including rigorous validation steps to ensure on-target specificity, and second the need to improve methods for defining TCR specificity.

Thus, for therapeutic vaccination the ideal TA should possess the following characteristics: (1) a high level of tumor specificity, (2) high levels of presentation in the tumor, (3) high affinity for a particular MHC I molecule, (4) high affinity for T cell receptors (TCRs) to demonstrate immunogenicity, and (5) be shared among patients. Antigens that meet all these criteria have the potential to yield the best therapeutic outcomes.

Although most of these characteristics can be easily assessed by prediction algorithms, they must be rigorously verified in the subsequent in vivo or in vitro validation phases. For instance, the high affinity to MHC I molecules is currently easily and widely assessed by using algorithms, such as NetMHCpan[179]. This algorithm predicts from the amino acid sequence the affinity of the binding between the peptide and the MHC I molecules grooves. Similarly, using neural network algorithms, such as Repitope[180], immunogenicity can be predicted by giving an immunogenicity score. The higher the score the higher the predicted immunogenicity for the peptide is.

Conversely, the assessment of tumor specificity and intra- and interindividual presentation levels remains a major challenge, as all kinds of confounding factors undermine the accuracy in this validation step[181]. A first and simplistic approach to address tumor specificity might consider assessing whether the candidate TSA sequence has not previously been reported as a "normal" MAP in public immunopeptidomics studies of normal tissues. Indeed, a currently available public immunopeptidomics atlas of healthy tissues can be used to determine the tumor specificity of a particular peptide[182]. However, the use of this atlas only provides the certainty to stop considering a TSA as a candidate when its sequence has already been detected in some normal tissue, but not otherwise. This means that the absence of the peptide in normal tissues does not imply a direct confirmation that the TSA candidate is a true TSA. In fact, the atlas is expected to remain incomplete due to the limitations of mass spectrometry-based approaches, so other lines of evidence need to be added to assess specificity.

Tumor specificity can also be assessed by evaluating and comparing the expression of antigens in normal, normal adjacent and cancerous tissues. This can be done at the immunopeptidome level by MS peptide quantification assays following immunoprecipitation of

MHC-I complexes[82]. Although this may be feasible, it is unlikely to be applied in a TSA validation setting due to logistical issues: MS requires large amounts of material, is costly, time-consuming and has low throughput. Concerning the expression assessment in adjacent tumor tissues, it would be necessary to first differentiate between normal and pre-neoplastic adjacent tissues to obtain an accurate estimation of the TSA expression[183]. Accordingly, a recent study examined the immunopeptidome from matched healthy and tumor lung tissues and performed a direct comparison of the peptides identified. Only 10 non-canonical peptides were identified exclusively in tumor lung tissue; however, based on RNA level assessment, only one peptide was found not to be expressed in GTEx[74]. This shows that the comparison between the tumor and adjacent healthy tissue immunopeptidomics is not sufficient and therefore further validation lines (expression in normal tissues) should be added.

Therefore, analysis of transcript abundance and immunopeptidome properties provides crucial information on the internal state of healthy and unhealthy cells, allowing the determination of responsiveness to treatments[184]. Typically, proteogenomic approaches defined a unique MAP origin, most of the time selecting the genomic location with the highest RNA expression in cancer. Consequently, the biotype (classification based on the annotation of genomic regions in the reference database) assigned to a given MAP is based on that genomic location. The selected locations are then used to quantify the MAPs RNA expression in large RNA sequencing databases (normal GTEx[185] and cancerous https://www.cancer.gov/tcga), thus allowing to predict MAP presence in normal and cancer cells[73, 74, 77, 133, 134]. Yet, a peptide can be produced by multiple RNA sequences since 75% of the genome can be transcribed and thus potentially translated[50]. This suggests that the expression of MAPs based on a single RNA sequence may underestimate the actual expression and thus lead to misclassification. Accordingly, in **Chapter 3**, we present BamQuery, a tool that aims to assign a comprehensive RNA-seq expression to any MAP to use this information as a predictor of MAP presentation in any tissue. To facilitate the validation of antigens as potential targets for immunotherapy, BamQuery is utilized to assess the RNA-seq expression of all genomic regions capable of generating a specific MAP.

Altogether, the immune system can be modulated to attack cancer cells from a variety of angles. Vaccines for cancer treatment are a type of immunotherapy that stimulates T cells in vivo to recognize and destroy cancer cells. Treatment with such vaccines should consider the inclusion of other therapies, such as immune checkpoint blockade, to enhance the efficacy of immunotherapy and provide lasting and safe benefits. To date, mass spectrometry-based proteogenomic approaches offer the most plausible and accurate way to discover the best candidates present in tumor cells. Many types of tumor antigens (aeTSAs, mTSAs, CTAs, TAAs) that mostly derived from atypical translations (non-canonical proteins) remain to be discovered. However, aeTSAs appear to be the most interesting targets for designing cancer vaccines due to their specificity and their share ability among patients. After identification of TSAs by proteogenomics, several steps need to be addressed to achieve successful therapy. First, the validation and prioritization step to evaluate the extent to which TSA targets provide a potent and specific immune response. Second, affinity and avidity testing of T cells in vitro and in vivo. Third, the definition of the optimal form for the administration of tumor antigens: vectorized or biochemically defined antigen formulations[186]. Finally, combination of the cancer vaccine with other methods to improve efficacy such as radiotherapy, chemotherapy, immune checkpoint inhibitors or other vaccine adjuvants[187].

1.3 Objectives of the thesis

Recently, several studies have demonstrated the presence of a hidden proteome translated from the putative non-coding fraction of the human genome, from “untranslatable” regions within protein-coding genes and from different reading frames in known proteins. For a long time, these proteins, collectively coined as non-canonical proteins, went unnoticed because they have physical and chemical properties very different from those of known or canonical proteins. For instance, they are shorter, initiate from codons other than the AUG, are predicted to be less stable, and have lower rates of transcription and translation. However, advances in genomics and proteomics have facilitated their detection with high resolution and reliability. Notably, Ribo-seq has enabled to demonstrate at the codon level the pervasiveness of translation throughout the genome, resulting in non-canonical proteins.

In cancer, the non-canonical proteins appear to be up-regulated due to the strong inherent deregulation of cancer cells, making them the main source of actionable antigens for the design of cancer vaccine treatments. In this regard, proteogenomics, which combines sequencing data (RNA-seq, Ribo-seq) and proteomics (mass spectrometry), has been key to the recent revolution in the field of immunopeptidomics. Indeed, proteogenomics allows screening the immunopeptidome at the molecular level facilitating the search of non-canonical antigens capable of activate T lymphocytes. Consequently, candidate antigens for vaccine treatment need further validation by evaluating their expression in healthy tissues to ensure their exclusive expression in cancer and thus give them priority in clinical trials.

1.3.1 General Objective

The main objective of this thesis was to *investigate to what extent the non-canonical proteome contributes to the immunopeptidome of cancer cells*.

1.3.2 Specific aims

The main objective of my thesis was addressed through the following 3 specific aims:

1. To identify MAPs translated from non-canonical proteome in cancer cells using a proteogenomic approach leveraging paired RNA and Ribo-seq data. (Chapter 2)

2. To compare the properties of detected non-canonical proteins with canonical proteins in cancer cells. (Chapters 2 and 3)
3. To develop a tool to facilitate the validation of non-canonical MAPs as tumor antigens. (Chapter 3)

1.3.3 Model cell lines

To study the contribution of non-canonical proteins to the immunopeptidome we used three human diffuse large B-cell lymphoma (DLBCL), HBL-1, DoHH2 and SUDHL-4. Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin's lymphoma (NHL), accounting for approximately 30-40% of new cases each year. Although approximately 60% of patients with DLBCL are cured with chemotherapy, 30-40% of them will develop relapse or other refractory disease that cannot be cured with standard procedures, highlighting the need for other, more effective therapies[188]. DLBCL is expected to frequently lose the ability to present MAPs, which is related to defects in β2-microglobulin, an essential subunit of the MHC I molecules[189]. This mechanism of immunoediting observed in DLBCL, however, does not preclude the value of CD8+ T-cell-based immunotherapies, as these can be highly sensitive and require few MAPs to trigger immune responses[190]. Thus, DLBCL may potentially be an effective target for vaccine treatments considering its predisposition to increase somatic mutations. This predisposition can result in mutated peptides capable of eliciting CD8+ T-cell recognition. In addition, DLBCL is associated with chromosomal alterations that significantly affect gene expression, leading to a likely increase in non-canonical translation[191].

1.4 References

1. Marshall, J.S., Warrington, R., Watson, W. & Kim, H.L. An introduction to immunology and immunopathology. *Allergy Asthma Clin Immunol* **14**, 49 (2018).
2. Nicholson, L.B. The immune system. *Essays Biochem* **60**, 275-301 (2016).
3. Chaplin, D.D. Overview of the immune response. *J Allergy Clin Immunol* **125**, S3-23 (2010).
4. Yewdell, J.W. Immunology. Hide and seek in the peptidome. *Science* **301**, 1334-1335 (2003).
5. Pishesha, N., Harmand, T.J. & Ploegh, H.L. A guide to antigen processing and presentation. *Nat Rev Immunol* (2022).
6. Blum, J.S., Wearsch, P.A. & Cresswell, P. Pathways of antigen processing. *Annu Rev Immunol* **31**, 443-473 (2013).
7. Neefjes, J., Jongsma, M.L., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* **11**, 823-836 (2011).
8. Walz, S. et al. The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for T-cell-based immunotherapy. *Blood* **126**, 1203-1213 (2015).
9. Schuster, H. et al. The immunopeptidomic landscape of ovarian carcinomas. *Proc Natl Acad Sci U S A* **114**, E9942-E9951 (2017).
10. Yewdell, J.W., Reits, E. & Neefjes, J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat Rev Immunol* **3**, 952-961 (2003).
11. Engelhard, V.H. Structure of peptides associated with class I and class II MHC molecules. *Annu Rev Immunol* **12**, 181-207 (1994).
12. Pearson, H. et al. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest* **126**, 4690-4701 (2016).
13. Caron, E. et al. The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol Syst Biol* **7**, 533 (2011).
14. Laumont, C.M. & Perreault, C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cell Mol Life Sci* **75**, 607-621 (2018).
15. Wahl, A., Schafer, F., Bardet, W. & Hildebrand, W.H. HLA class I molecules reflect an altered host proteome after influenza virus infection. *Hum Immunol* **71**, 14-22 (2010).

16. Esquivel, F., Yewdell, J. & Bennink, J. RMA/S cells present endogenously synthesized cytosolic proteins to class I-restricted cytotoxic T lymphocytes. *J Exp Med* **175**, 163-168 (1992).
17. Wooldridge, L. et al. A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem* **287**, 1168-1177 (2012).
18. Wherry, E.J. & Masopust, D. in Viral Pathogenesis (Third Edition). (eds. M.G. Katze, M.J. Korth, G.L. Law & N. Nathanson) 57-69 (Academic Press, Boston; 2016).
19. den Haan, J.M.M., Arens, R. & van Zelm, M.C. The activation of the adaptive immune system: Cross-talk between antigen-presenting cells, T cells and B cells. *Immunology Letters* **162**, 103-112 (2014).
20. Kaech, S.M. & Cui, W. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat Rev Immunol* **12**, 749-761 (2012).
21. Pollard, A.J. & Bijker, E.M. A guide to vaccinology: from basic principles to new developments. *Nat Rev Immunol* **21**, 83-100 (2021).
22. Vyas, J.M., Van der Veen, A.G. & Ploegh, H.L. The known unknowns of antigen processing and presentation. *Nat Rev Immunol* **8**, 607-618 (2008).
23. Haen, S.P., Loffler, M.W., Rammensee, H.G. & Brossart, P. Towards new horizons: characterization, classification and implications of the tumour antigenic repertoire. *Nat Rev Clin Oncol* **17**, 595-610 (2020).
24. Granados, D.P., Laumont, C.M., Thibault, P. & Perreault, C. The nature of self for T cells-a systems-level perspective. *Curr Opin Immunol* **34**, 1-8 (2015).
25. Alexandropoulos, K. & Danzl, N.M. Thymic epithelial cells: antigen presenting cells that regulate T cell repertoire and tolerance development. *Immunol Res* **54**, 177-190 (2012).
26. Hogquist, K.A., Baldwin, T.A. & Jameson, S.C. Central tolerance: learning self-control in the thymus. *Nat Rev Immunol* **5**, 772-782 (2005).
27. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat Rev Immunol* **9**, 833-844 (2009).

28. Kyewski, B. & Derbinski, J. Self-representation in the thymus: an extended view. *Nat Rev Immunol* **4**, 688-698 (2004).
29. Houghton, A.N. & Guevara-Patino, J.A. Immune recognition of self in immunity against cancer. *J Clin Invest* **114**, 468-471 (2004).
30. Dhatchinamoorthy, K., Colbert, J.D. & Rock, K.L. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. *Front Immunol* **12**, 636568 (2021).
31. Baghban, R. et al. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun Signal* **18**, 59 (2020).
32. Schreiber, R.D., Old, L.J. & Smyth, M.J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565-1570 (2011).
33. Berraondo, P. et al. Cytokines in clinical cancer immunotherapy. *Br J Cancer* **120**, 6-15 (2019).
34. Lian, J., Yue, Y., Yu, W. & Zhang, Y. Immunosenescence: a key player in cancer development. *J Hematol Oncol* **13**, 151 (2020).
35. Qian, S.B., Princiotta, M.F., Bennink, J.R. & Yewdell, J.W. Characterization of rapidly degraded polypeptides in mammalian cells reveals a novel layer of nascent protein quality control. *J Biol Chem* **281**, 392-400 (2006).
36. Chen, W., Smeekens, J.M. & Wu, R. Systematic study of the dynamics and half-lives of newly synthesized proteins in human cells. *Chem Sci* **7**, 1393-1400 (2016).
37. Schwanhausser, B. et al. Corrigendum: Global quantification of mammalian gene expression control. *Nature* **495**, 126-127 (2013).
38. Milner, E., Barnea, E., Beer, I. & Admon, A. The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol Cell Proteomics* **5**, 357-365 (2006).
39. Anton, L.C. & Yewdell, J.W. Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. *J Leukoc Biol* **95**, 551-562 (2014).
40. Princiotta, M.F. et al. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity* **18**, 343-354 (2003).

41. Schubert, U. et al. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **404**, 770-774 (2000).
42. Yewdell, J.W. DRiPs solidify: progress in understanding endogenous MHC class I antigen processing. *Trends Immunol* **32**, 548-558 (2011).
43. Yewdell, J.W., Anton, L.C. & Bennink, J.R. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *J Immunol* **157**, 1823-1826 (1996).
44. Croft, N.P. et al. Kinetics of antigen expression and epitope presentation during virus infection. *PLoS Pathog* **9**, e1003129 (2013).
45. Khan, S. et al. Cutting edge: neosynthesis is required for the presentation of a T cell epitope from a long-lived viral protein. *J Immunol* **167**, 4801-4804 (2001).
46. Bourdetsky, D., Schmelzer, C.E. & Admon, A. The nature and extent of contributions by defective ribosome products to the HLA peptidome. *Proc Natl Acad Sci U S A* **111**, E1591-1599 (2014).
47. Dersh, D., Holly, J. & Yewdell, J.W. A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion. *Nat Rev Immunol* (2020).
48. Granados, D.P. et al. Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat Commun* **5**, 3600 (2014).
49. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515 (2019).
50. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
51. Ingolia, N.T. et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**, 1365-1379 (2014).
52. Heinen, T.J., Staubach, F., Haming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr Biol* **19**, 1527-1531 (2009).
53. Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H. & Shiu, S.H. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res* **17**, 632-640 (2007).
54. Kapranov, P. et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919 (2002).

55. Clark, M.B. et al. The reality of pervasive transcription. *PLoS Biol* **9**, e1000625; discussion e1001102 (2011).
56. Bertone, P. et al. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246 (2004).
57. Samandi, S. et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife* **6**, e27860 (2017).
58. Zhao, J., Qin, B., Nikolay, R., Spahn, C.M.T. & Zhang, G. Translatomics: The Global View of Translation. *Int J Mol Sci* **20** (2019).
59. Bazzini, A.A. et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981-993 (2014).
60. Lu, S. et al. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res*, 8111-8125 (2019).
61. Kim, M.S. et al. A draft map of the human proteome. *Nature* **509**, 575-581 (2014).
62. van Heesch, S. et al. The Translational Landscape of the Human Heart. *Cell* **178**, 242-260 e229 (2019).
63. Slavoff, S.A. et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**, 59-64 (2013).
64. Bergeron, D. et al. An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J Biol Chem* **288**, 21824-21835 (2013).
65. Vanderperre, B. et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8**, e70698 (2013).
66. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140-1146 (2020).
67. Ruiz-Orera, J., Messeguer, X., Subirana, J.A. & Alba, M.M. Long non-coding RNAs as a source of new peptides. *Elife* **3**, e03523 (2014).
68. Jackson, R. et al. The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434-438 (2018).

69. Nelson, B.R. et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271-275 (2016).
70. Matsumoto, A. et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**, 228-232 (2017).
71. Prensner, J.R. et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat Biotechnol* **39**, 697-704 (2021).
72. Laumont, C.M. et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun* **7**, 10238 (2016).
73. Ouspenskaia, T. et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol* (2021).
74. Chong, C. et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* **11**, 1293 (2020).
75. Erhard, F., Dolken, L., Schilling, B. & Schlosser, A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunol Res* **8**, 1018-1026 (2020).
76. Erhard, F. et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**, 363–366 (2018).
77. Laumont, C.M. et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med* **10**, eaau5516 (2018).
78. Purcell, A.W., Ramarathinam, S.H. & Ternette, N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc* **14**, 1687-1707 (2019).
79. Illing, P.T., Ramarathinam, S.H. & Purcell, A.W. New insights and approaches for analyses of immunopeptidomes. *Curr Opin Immunol* **77**, 102216 (2022).
80. Nesvizhskii, A.I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114-1125 (2014).
81. Lanoix, J. et al. Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. *Proteomics* **18**, e1700251 (2018).
82. Kuznetsov, A., Voronina, A., Govorun, V. & Arapidi, G. Critical Review of Existing MHC I Immunopeptidome Isolation Methods. *Molecules* **25** (2020).

83. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699-711 (2004).

84. Hu, A., Noble, W.S. & Wolf-Yadlin, A. Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res* **5** (2016).

85. Tian, X., Permentier, H.P. & Bischoff, R. Chemical isotope labeling for quantitative proteomics. *Mass Spectrom Rev* **42**, 546-576 (2023).

86. Schubert, O.T., Rost, H.L., Collins, B.C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat Protoc* **12**, 1289-1294 (2017).

87. Ong, S.E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386 (2002).

88. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904 (2003).

89. Wiese, S., Reidegeld, K.A., Meyer, H.E. & Warscheid, B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7**, 340-350 (2007).

90. Uchida, Y. et al. A study protocol for quantitative targeted absolute proteomics (QTAP) by LC-MS/MS: application for inter-strain differences in protein expression levels of transporters, receptors, claudin-5, and marker proteins at the blood-brain barrier in ddY, FVB, and C57BL/6J mice. *Fluids Barriers CNS* **10**, 21 (2013).

91. Kiyonami, R. & Domon, B. Selected reaction monitoring applied to quantitative proteomics. *Methods Mol Biol* **658**, 155-166 (2010).

92. Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S. & Coon, J.J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* **11**, 1475-1488 (2012).

93. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. *Anal Chem* **92**, 8072-8080 (2020).

94. Bilbao, A. et al. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15**, 964-980 (2015).

95. Gillet, L.C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**, O111 016717 (2012).

96. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **77**, 964-973 (2005).

97. Wan, Y., Yang, A. & Chen, T. PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. *Anal Chem* **78**, 432-437 (2006).

98. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **17**, 2337-2342 (2003).

99. Tran, N.H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* **114**, 8247-8252 (2017).

100. Muth, T., Hartkopf, F., Vaudel, M. & Renard, B.Y. A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *Proteomics* **18**, e1700150 (2018).

101. Muth, T. & Renard, B.Y. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* **19**, 954-970 (2018).

102. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).

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

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

105. Kall, L., Storey, J.D., MacCoss, M.J. & Noble, W.S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* **7**, 29-34 (2008).

106. Nesvizhskii, A.I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**, 2092-2123 (2010).

107. Aggarwal, S. & Yadav, A.K. False Discovery Rate Estimation in Proteomics. *Methods Mol Biol* **1362**, 119-128 (2016).

108. Wang, P. & Wilson, S.R. Mass spectrometry-based protein identification by integrating de novo sequencing with database searching. *BMC Bioinformatics* **14 Suppl 2**, S24 (2013).

109. Wang, X. et al. JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. *Mol Cell Proteomics* **13**, 3663-3673 (2014).

110. Na, S. & Paek, E. Software eyes for protein post-translational modifications. *Mass Spectrom Rev* **34**, 133-147 (2015).

111. Mann, M., Kumar, C., Zeng, W.F. & Strauss, M.T. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst* **12**, 759-770 (2021).

112. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* **20**, e1900334 (2020).

113. Engelhard, V.H., Altrich-Vanlith, M., Ostankovitch, M. & Zarling, A.L. Post-translational modifications of naturally processed MHC-binding epitopes. *Curr Opin Immunol* **18**, 92-97 (2006).

114. Solleder, M. et al. Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands. *Mol Cell Proteomics* **19**, 390-404 (2020).

115. Zarling, A.L. et al. Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *Proc Natl Acad Sci U S A* **103**, 14889-14894 (2006).

116. Mohammed, F. et al. The antigenic identity of human class I MHC phosphopeptides is critically dependent upon phosphorylation status. *Oncotarget* **8**, 54160-54172 (2017).

117. Cobbold, M. et al. MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia. *Sci Transl Med* **5**, 203ra125 (2013).

118. Penny, S.A. et al. Tumor Infiltrating Lymphocytes Target HLA-I Phosphopeptides Derived From Cancer Signaling in Colorectal Cancer. *Front Immunol* **12**, 723566 (2021).

119. Lin, M.H. et al. Immunological evaluation of a novel HLA-A2 restricted phosphopeptide of tumor associated Antigen, TRAP1, on cancer therapy. *Vaccine X* **1**, 100017 (2019).
120. Mester, G., Hoffmann, V. & Stevanovic, S. Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands. *Cell Mol Life Sci* **68**, 1521-1532 (2011).
121. Abelin, J.G. et al. Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry. *Nat Protoc* **10**, 1308-1318 (2015).
122. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
123. Dinger, M.E., Pang, K.C., Mercer, T.R. & Mattick, J.S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4**, e1000176 (2008).
124. Blakeley, P., Overton, I.M. & Hubbard, S.J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**, 5221-5234 (2012).
125. Li, H. et al. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **17**, 1031 (2016).
126. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
127. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).
128. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).
129. Evans, V.C. et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods* **9**, 1207-1211 (2012).
130. Rivero-Hinojosa, S. et al. Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat Commun* **12**, 6689 (2021).
131. Hirama, T. et al. Proteogenomic identification of an immunogenic HLA class I neoantigen in mismatch repair-deficient colorectal cancer tissue. *JCI Insight* **6** (2021).

132. Scull, K.E., Pandey, K., Ramarathinam, S.H. & Purcell, A.W. Immunopeptidogenomics: Harnessing RNA-Seq to Illuminate the Dark Immunopeptidome. *Mol Cell Proteomics* **20**, 100143 (2021).

133. Ehx, G. et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity* **54**, 737-752 e710 (2021).

134. Zhao, Q. et al. Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer. *Cancer Immunol Res* **8**, 544-555 (2020).

135. Cleyle, J. et al. Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability. *Mol Cell Proteomics* **21**, 100228 (2022).

136. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223 (2009).

137. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. & Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**, 1534-1550 (2012).

138. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).

139. Ingolia, N.T., Hussmann, J.A. & Weissman, J.S. Ribosome Profiling: Global Views of Translation. *Cold Spring Harb Perspect Biol* **11** (2019).

140. Raj, A. et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5** (2016).

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

142. Calviello, L. & Ohler, U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet* **33**, 728-744 (2017).

143. Mudge, J.M. et al. Standardized annotation of translated open reading frames. *Nat Biotechnol* **40**, 994-999 (2022).

144. Ruiz Cuevas, M.V. et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep* **34**, 108815 (2021).
145. Lubec, G. & Afjehi-Sadat, L. Limitations and pitfalls in protein identification by mass spectrometry. *Chem Rev* **107**, 3568-3584 (2007).
146. Hinrichs, C.S. & Restifo, N.P. Reassessing target antigens for adoptive T-cell therapy. *Nat Biotechnol* **31**, 999-1008 (2013).
147. Coulie, P.G., Van den Eynde, B.J., van der Bruggen, P. & Boon, T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat Rev Cancer* **14**, 135-146 (2014).
148. Chen, J. et al. Enhancing the Efficacy of Tumor Vaccines Based on Immune Evasion Mechanisms. *Front Oncol* **10**, 584367 (2020).
149. Topalian, S.L., Drake, C.G. & Pardoll, D.M. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer Cell* **27**, 450-461 (2015).
150. Massarelli, E. et al. Combining Immune Checkpoint Blockade and Tumor-Specific Vaccine for Patients With Incurable Human Papillomavirus 16-Related Cancer: A Phase 2 Clinical Trial. *JAMA Oncol* **5**, 67-73 (2019).
151. Esprit, A. et al. Neo-Antigen mRNA Vaccines. *Vaccines (Basel)* **8** (2020).
152. Miao, L., Zhang, Y. & Huang, L. mRNA vaccine for cancer immunotherapy. *Mol Cancer* **20**, 41 (2021).
153. Tran, E. et al. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N Engl J Med* **375**, 2255-2262 (2016).
154. Rizvi, N.A. et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124-128 (2015).
155. Yadav, M. et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572-576 (2014).
156. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

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

158. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463-1469 (2016).

159. Schumacher, T.N. & Schreiber, R.D. Neoantigens in cancer immunotherapy. *Science* **348**, 69-74 (2015).

160. Ott, P.A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217-221 (2017).

161. Ott, P.A. et al. A Phase Ib Trial of Personalized Neoantigen Therapy Plus Anti-PD-1 in Patients with Advanced Melanoma, Non-small Cell Lung Cancer, or Bladder Cancer. *Cell* **183**, 347-362 e324 (2020).

162. Martin, S.D. et al. Low Mutation Burden in Ovarian Cancer May Limit the Utility of Neoantigen-Targeted Vaccines. *PLoS One* **11**, e0155189 (2016).

163. Sriram, A., Bohlen, J. & Teleman, A.A. Translation acrobatics: how cancer cells exploit alternate modes of translational initiation. *EMBO Rep* **19** (2018).

164. Robichaud, N., Sonenberg, N., Ruggero, D. & Schneider, R.J. Translational Control in Cancer. *Cold Spring Harb Perspect Biol* **11** (2019).

165. van der Bruggen, P. et al. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* **254**, 1643-1647 (1991).

166. Scanlan, M.J., Simpson, A.J. & Old, L.J. The cancer/testis genes: review, standardization, and commentary. *Cancer Immun* **4**, 1 (2004).

167. Hofmann, O. et al. Genome-wide analysis of cancer/testis gene expression. *Proc Natl Acad Sci U S A* **105**, 20422-20427 (2008).

168. Morgan, R.A. et al. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J Immunother* **36**, 133-151 (2013).

169. Michaels, Y.S. et al. Precise tuning of gene expression levels in mammalian cells. *Nat Commun* **10**, 818 (2019).

170. 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* **180**, 35-42 (1994).
171. Nordlund, J.J. et al. Vitiligo in patients with metastatic melanoma: a good prognostic sign. *J Am Acad Dermatol* **9**, 689-696 (1983).
172. Quaglino, P. et al. Vitiligo is an independent favourable prognostic factor in stage III and IV metastatic melanoma patients: results from a single-institution hospital-based observational cohort study. *Ann Oncol* **21**, 409-414 (2010).
173. Van Tine, B. et al. ADP-A2M4 (MAGE-A4) in patients with synovial sarcoma. *Annals of Oncology* **30**, v684-v685 (2019).
174. Parkhurst, M.R. et al. T cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. *Mol Ther* **19**, 620-626 (2011).
175. Zhang, B. et al. Clinical potential of mass spectrometry-based proteogenomics. *Nat Rev Clin Oncol* **16**, 256-268 (2019).
176. Smart, A.C. et al. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol* **36**, 1056-1058 (2018).
177. Chong, C., Coukos, G. & Bassani-Sternberg, M. Identification of tumor antigens with immunopeptidomics. *Nat Biotechnol* **40**, 175-188 (2022).
178. Linette, G.P. et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122**, 863-871 (2013).
179. Jurtz, V. et al. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* **199**, 3360-3368 (2017).
180. Ogishi, M. & Yotsuyanagi, H. Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space. *Front Immunol* **10**, 827 (2019).
181. Vitiello, A. & Zanetti, M. Neoantigen prediction and the need for validation. *Nat Biotechnol* **35**, 815-817 (2017).
182. Marcu, A. et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer* **9** (2021).

183. Apavaloaei, A., Hardy, M.P., Thibault, P. & Perreault, C. The Origin and Immune Recognition of Tumor-Specific Antigens. *Cancers (Basel)* **12** (2020).

184. Vizcaino, J.A. et al. The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. *Mol Cell Proteomics* **19**, 31-49 (2020).

185. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).

186. Romero, P. et al. The Human Vaccines Project: A roadmap for cancer vaccine development. *Sci Transl Med* **8**, 334ps339 (2016).

187. Cuzzubbo, S. et al. Cancer Vaccines: Adjuvant Potency, Importance of Age, Lifestyle, and Treatments. *Front Immunol* **11**, 615240 (2020).

188. Zhang, J., Medeiros, L.J. & Young, K.H. Cancer Immunotherapy in Diffuse Large B-Cell Lymphoma. *Front Oncol* **8**, 351 (2018).

189. Challa-Malladi, M. et al. Combined genetic inactivation of beta2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* **20**, 728-740 (2011).

190. Sykulev, Y., Joo, M., Vturina, I., Tsomides, T.J. & Eisen, H.N. Evidence that a single peptide-MHC complex on a target cell can elicit a cytolytic T cell response. *Immunity* **4**, 565-571 (1996).

191. Schmitz, R. et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med* **378**, 1396-1407 (2018).

Chapter 2 – Most Non-canonical Proteins Uniquely Populate the Proteome or Immunopeptidome

Maria Virginia Ruiz Cuevas[1,2,7], Marie-Pierre Hardy[1,7], Jaroslav Holly[3,7], Éric Bonneil[1], Chantal Durette[1], Mathieu Courcelles[1], Joël Lanoix[1], Caroline Côté[1], Louis M. Staudt[4], Sébastien Lemieux[1,2,8], Pierre Thibault[1,5,8], Claude Perreault[1,6,8,9,*], Jonathan W. Yewdell[3,8,*]

[1] Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[2] Department of Biochemistry and Molecular Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[3] Cellular Biology Section, Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

[4] Department of Chemistry, Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[5] Department of Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[6] Lymphoid Malignancies Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[7] These authors contributed equally

[8] Senior author

[9] Lead Contact

***Correspondence**: claude.perreault@umontreal.ca (C.P.), jyewdell@nih.gov (J.W.Y)

Cell Report, volume 34, 10.1016/j.celrep.2021.108815 (Mars 9, 2021)

1.5 Context

The immunopeptidome is the MHC-I-associated peptides (MAPs) repertoire presented on the surface of cells. MAPs play a key role in the immune system's recognition and response to abnormal substances, such as those derived from viruses or cancer cells. MAPs have previously been reported to be derived from a variety of genomic regions, including both protein-coding and non-protein-coding sequences, such as endogenous retroelements (EREs). Indeed, ribosome profiling (Ribo-seq) experiments have revealed that translation occurs largely outside of annotated protein-coding genes, resulting in non-canonical proteins. With this in mind, we sought to combine RNA-seq with Ribo-seq to facilitate peptide identification using MS. The goal was to assess the contribution of such non-canonical translations in the DLBCL cell lines immunopeptidome and proteome.

In this article, we present Ribo-db, a proteogenomic tool developed to create custom databases for MS peptide identification. Using Ribo-db, the immunopeptidome of DLBCL was analyzed finding that non-canonical proteins made up a significant proportion (10%) of the MAP repertoire. Non-canonical proteins were also found to have lower transcription and translation rates and were predicted to be less stable in vivo compared to canonical proteins. Upon further analysis of the whole proteome, low overlap between the non-canonical MAPs source proteome and that identified in the whole proteome was observed, suggesting the presence of two distinct non-canonical proteomes.

As anticipated, our analysis also found that genomic abnormalities related to the oncogenic program of DLBCL may be responsible for its non-canonical translation landscape. Based on these findings and fueled by previous reports showing the potential of non-canonical MAPs to serve as actionable targets, we launched the project described hereunder **Chapter 3.** In this project the goal was of systematically evaluate MAPs as potential targets for safe immunotherapy.

1.6 Authors' contributions

Maria Virginia Ruiz Cuevas: conceptualization of the study, software development, made formal analysis, wrote the first draft of the manuscript.

Marie-Pierre Hardy: conceptualization of the study, made formal analysis, wrote the first draft of the manuscript.

Jaroslav Holly: made formal analysis and investigation (processing cell samples, performing RNA-seq and Ribo-seq).

Éric Bonneil: participated in the MS analysis of the whole proteome extracts.

Chantal Durette: participated in the MS analysis of the immunopeptidome performing database searches for the MAPs identification.

Mathieu Courcelles: participated in the MS analysis of the immunopeptidome performing database searches for the MAPs identification.

Joël Lanoix: participated in the MS analysis of the immunopeptidome performing the immunoprecipitation.

Caroline Côté: participated in the cultured of cells for the MS whole proteome extracts analysis.

Louis M. Staudt: Supervised the study.

Sébastien Lemieux: Supervised the study.

Pierre Thibault: Supervised the study.

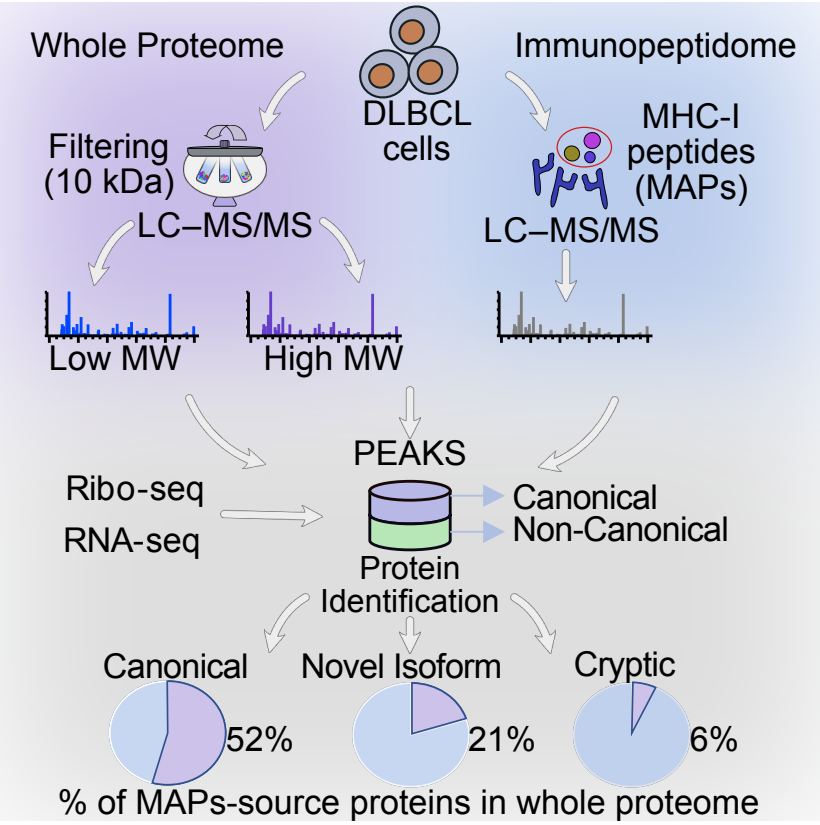
Claude Perreault: Supervised the study.

Jonathan W. Yewdell: Supervised the study.

All authors reviewed, edited, and approved the final version of the manuscript

1.7 Abstract

Combining RNA sequencing, ribosome profiling, and mass spectrometry, we elucidate the contribution of non-canonical translation to the proteome and major histocompatibility complex (MHC) class I immunopeptidome. Remarkably, of 14,498 proteins identified in three human B cell lymphomas, 2,503 are non-canonical proteins. Of these, 28% are novel isoforms and 72% are cryptic proteins encoded by ostensibly non-coding regions (60%) or frameshifted canonical genes (12%). Cryptic proteins are translated as efficiently as canonical proteins, have more predicted disordered residues and lower stability, and critically generate MHC-I peptides 5-fold more efficiently per translation event. Translating 5' ''untranslated'' regions hinders downstream translation of genes involved in transcription, translation, and antiviral responses. Novel protein isoforms show strong enrichment for signaling pathways deregulated in cancer. Only a small fraction of cryptic proteins detected in the proteome contribute to the MHC-I immunopeptidome, demonstrating the high preferential access of cryptic defective ribosomal products to the class I pathway.



1.8 Introduction

Ribosome profiling (Ribo-seq) and mass spectrometry (MS) analyses reveal that many proteins are encoded by non-canonical open reading frames (ORFs)[1-4]. Non-canonical proteins are encoded by both ostensibly noncoding ORFs and canonical ORFs in +1 or +2 reading frames. Accumulating evidence suggests that, far from representing translational noise, non-canonical proteins often exhibit critical and diverse cellular functions[5, 6]. Notably, when compared to classic ORFs, non-canonical ORFs present several distinctive features: they are shorter, have lower transcription and translation rates, commonly initiate translation on near-cognate codons (i.e., differ from AUG by a single nucleotide) and are predicted to be less stable in-vivo[3, 6-11].

Due to their short length and low abundance, non-canonical proteins are challenging to detect in whole-cell extracts by shotgun MS analyses. However, in the cells of jawed vertebrates, major histocompatibility complex class I molecules (MHC-I) have the remarkable ability to non-covalently bind and protect peptides, many of which derive from defective ribosomal products (DRiPs) and short-lived proteins (SLiPs)[12]. DRiPs are translation products that do not achieve functional integration to the proteome and are degraded with an average half-life on the order of 8 minutes[13-17]. MHC-I-peptide complexes are transported to the cell surface to enable T cell immunosurveillance of infected and neoplastic cells. Cell surface MHC-I-associated peptides (MAPs) exhibit half-lives on the order of 12 hours[18, 19], far longer than their source polypeptides in the case of SLiPs and DRiPs[20]. Thus, MHC-I serves as a sink for peptides whose source protein translation would otherwise be invisible to MS due to their rapid degradation.

Indeed, accumulating evidence indicates that a sizeable fraction of MAPs is encoded by non-canonical ORFs[21-24], which provide most tumor-specific antigens[22, 25]. Due to its tight linkage to translation, the class I immunopeptidome is highly dynamic and sensitive to metabolic perturbation, infection, and neoplastic transformation[22, 26, 27]. By contrast, the MHC class II immunopeptidome largely derives from large and stable proteins, with a trace contribution of non-canonical ORFs[23], due to the predominant loading of class II molecules in the lysosomal/endosomal compartment.

MS analysis provides concrete evidence for the translation of a given polypeptide. Large-scale MS analyses of proteins and MAPs have been considerably refined over the last few years, with considerable increases in sensitivity and accuracy[23, 24, 28-31]. However, shotgun MS still requires creating a reference database to identify peptides present in a given sample. This becomes limiting when searching for non-canonical peptides that potentially originate from any genomic sequence. All-frame in-silico translation of entire transcriptomes creates enormous databases, and searching MS data against such inflated reference databases generates false positives at an unacceptable rate[32-34]. Various approaches have been employed to optimize the reference database size based on the in-silico translation of transcriptomic data.

One reductionist approach to identify unique tumor-specific MAPs rests on purging the reference database of sequences present in non-tumor cells[22, 25]. More recently, two proof-of-principle studies established that cancer MAPs can be identified using reference databases built from Ribo-seq[23, 24]. Here, we describe a proteogenomic approach to identify non-canonical translation products present in whole-cell extracts and the immunopeptidome. Our findings demonstrate distinct features of the non-canonical translatome and their critical contribution to tumor immunosurveillance.

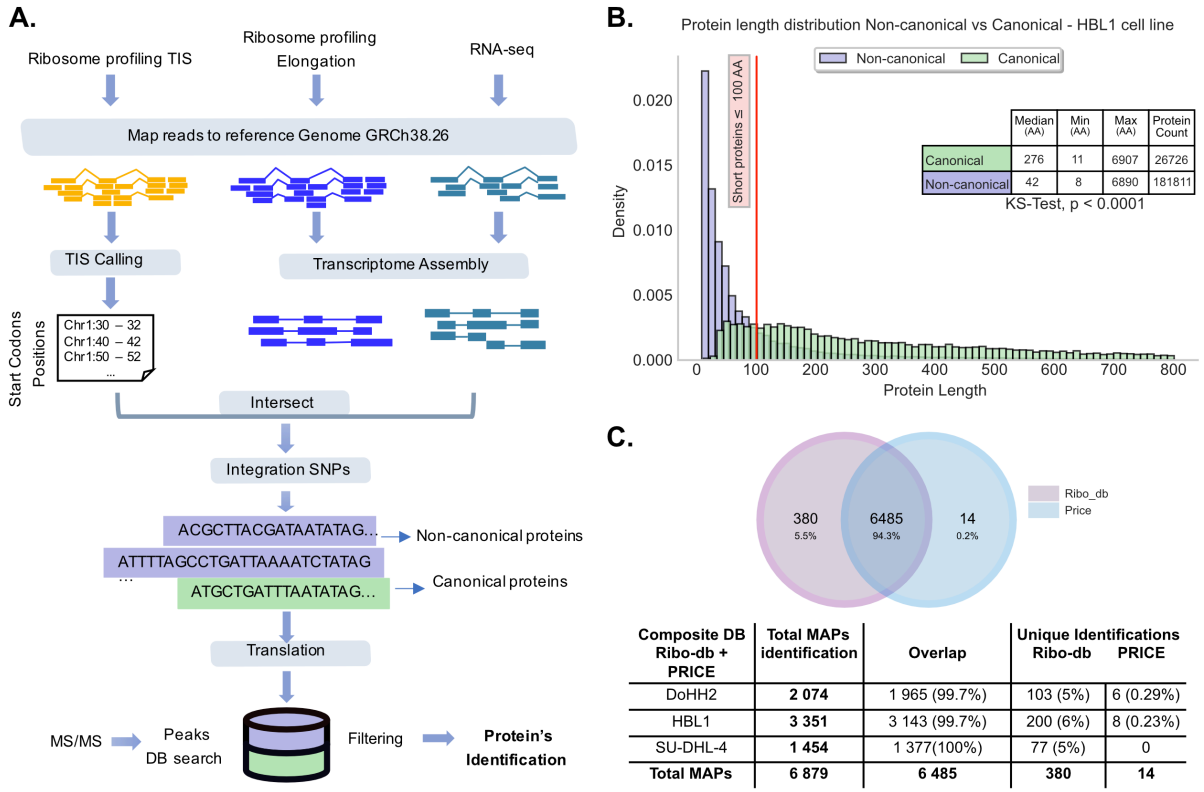
1.9 Results

1.9.1 A Proteogenomic Strategy for Identification of Non-canonical Translation Products

To identify non-canonical proteins, we developed an approach that combines Ribo-Seq and RNA sequencing (RNA-seq) data to create non-redundant sample-specific protein databases (Ribo-db) containing only actively translated sequences. Indeed, after retrieving and sequencing ribosome-protected RNA fragments, Ribo-seq produces a detailed map of active cell translation events[35]. Here, we collected Ribo-seq translation initiation sites (TISs), elongation, and RNA-seq data from three human diffuse large B cell lymphomas (DLBCLs), HBL-1, DoHH2, and SU-DHL-4. We intersected genomic positions of the start codons to the genomic positions of the assembled transcripts (Ribo-seq elongation and RNA-seq) to generate the set of ORFs (coupled start codon with an assembled transcript) for *in-silico* translation (see Methods and Figure 1A). From this set of ORFs, we define canonical proteins as those translated from an annotated start codon coupled to the corresponding transcript according to genome version GRCh38.p10 (GENCODE version 26). We define non-canonical translation products as those originating from a non-annotated initiation site, a new transcript, or both. We combined translation products into a sample-specific database for MS analysis (Figure 1A).

We first analyzed the general features of Ribo-db predicted canonical and non-canonical translation products. As reported[7], non-canonical proteins were more numerous but shorter than canonical proteins (Figure 1B). Indeed, ~70% of non-canonical proteins in the three cell lines were ≤100 amino acids (Figure 1B and Supplementary Figure 7A-7B). Next, we assessed the sensitivity and specificity of Ribo-db by comparison to PRICE as a benchmark[8]. PRICE was developed to identify non-canonical translation events that generate MAPs. Because the calculation of the False Discovery Rate (FDR) is directly related to the size of the database under target-decoy approaches[32, 33, 36], it is difficult to make a valid comparison between databases in which their size differs significantly (Table 1). To mitigate this, for each DLBCL, we generated a composite database combining Ribo-db and PRICE sequences to identify MAPs detected by tandem MS. We based MAP identification on three criteria: a peptide length between 8 and 11 amino acids, a

predicted MHC binding affinity in the top 2% for the corresponding HLA class I molecules expressed by each tumor, and a sample-specific FDR (see Methods and Supplementary Figure 7E). We recognize that peptides with lower predicted MHC binding affinity can represent genuine MAPs[37]. However, given the very high number of predicted non-canonical proteins (Figure 1B), we deemed it preferable, at this stage, to employ stringent selection criteria that may underestimate the number of non-canonical MAPs. Our Ribo-db approach identified 99.7% of MAPs identified with PRICE and 5% to 6% of MAPs missed by PRICE (Figure 1C). The number of MAPs identified per cell line positively correlated with the total class I cell surface expression determined by the binding of the W6/32 pan HLA class I monoclonal antibody (mAb) (Supplementary Figure 7C and 7D). We conclude that Ribo-db is well suited to discovering non-canonical translation products, outperforming PRICE, the previous best-in-class method for probing peptides arising from non-canonical translation.



Composite DB Ribo-db + PRICE	Total MAPs identification	Overlap	Unique Identifications Ribo-db	PRICE
DoHH2	**2 074**	1 965 (99.7%)	103 (5%)	6 (0.29%)
HBL1	**3 351**	3 143 (99.7%)	200 (6%)	8 (0.23%)
SU-DHL-4	**1 454**	1 377(100%)	77 (5%)	0
Total MAPs	**6 879**	**6 485**	**380**	**14**

Figure 1. – Ribo-seq-based proteogenomic approach for MS identification of non-canonical translation products.

(A) General overview of the workflow used to generate sample-specific databases containing active canonical and non-canonical translations based on Ribo-Seq data.

(B) Length distribution of canonical vs. non-canonical proteins from HBL-1 cells. ****p<0.0001, Kolmogorov-Smirnov Test. Proteins with a length >800 amino acids are not displayed.

(C) Venn diagram and table showing MAPs identified with the Ribo-db approach and the PRICE method.

1.9.2 The Global Landscape of Non-canonical MAPs

To optimize MAP identification and evaluate the contribution of non-canonical translation products, we performed MS searches using the Ribo-db customized databases. Because this database is smaller than the composite (Ribo-db+PRICE) database (Table 1), we discarded fewer identified MAPs because of the FDR. Despite the smaller size of the Ribo-db database, we identified 166 more MAPs than if we had used the composite database (7,045 versus 6,879 total MAPs, respectively) (Supplementary Figure 7C). To identify MAP source proteins, we considered that any MAP sequence might be redundant in the database. Therefore, we used a strategy to

assign the most likely origin for individuals MAPs, based on 1) the start codon score issued from the TIS-calling method, 2) the presence of an optimal or strong Kozak motif embedding the start codon[38], and 3) the expression level of the source transcript as determined by read numbers (Supplementary Figure 7E).

Out of the 7,045 identified MAPs, 6,520 source ORFs were canonical and 525 were non-canonical (Figure 2A). Key features of canonical and non-canonical MAPs were highly similar: length distribution (mostly nonamers), PEAKS peptide confidence score (20.92 canonical versus 20.15 non-canonical median scores), NetMHC-pan predicted MHC binding affinity in the top 2% for the corresponding HLA allotype (median binding rank % of 0.16 for canonical and 0.15 for non-canonical MAPs).

We then assessed the accuracy of non-canonical identifications using three validation methods. First, we compared the observed retention times of liquid chromatography-tandem mass spectrometry (LC-MS/MS)-sequenced peptides[39] to the DeepLC algorithm predicted-retention times[40]. Both canonical and non-canonical peptides showed an excellent correlation between experimental and predicted retention times (Figure 2B). Second, we evaluated the relative mass error between the measured experimental values and the expected mass for all peptides. No significant difference was found in the distribution of mass errors of canonical versus non-canonical peptides (Figure 2C). Lastly, we repeated all peptide searches using Comet[41]. The average percentage of PEAKs to Comet peptides re-identification was similar for canonical and non-canonical peptides (85% for canonical and 83% for non-canonical peptides) (Figure 2D). Together, these validations further reinforce the authenticity of our non-canonical identifications.

The 6,520 canonical MAPs derive from 4,493 canonical proteins (91%) and the 525 non-canonical MAPs from 451 non-canonical proteins (9%) (Supplementary Figure 8A). Consistent with the differential length of canonical and non-canonical proteins (Figure 1B, Supplementary Figure 7A, and 7B), non-canonical MAPs derived from shorter proteins than canonical MAPs (Figure 2E). Non-canonical MAP source proteins were classified according to their gene biotype (transcript classification) using GENCODE annotation[42]. The majority (79%) derive from sequences within protein-coding transcripts (including novel isoforms, UTRs, and frameshifts); 12% from

transcripts assumed to be non-coding, such as pseudogenes, non-coding RNAs, or processed transcripts; 7% from intergenic regions; and 2% from introns (Figure 2F). This is consistent with evidence for peptides generated from these ostensibly non-coding regions of the genome[3, 5, 21, 43, 44], though it does not support a major role for introns in generating the immunopeptidome in these cells.

As previously shown[6], among the non-canonical proteins derived from protein-coding transcripts, MAP source ORFs attributed to 5'UTR were 4-fold more frequent than 3'UTR (13% versus 3% of total non-canonical proteins) (Figure 2F and Supplementary Figure 8B). MAPs resulting from canonical gene frameshifting (13%) confirmed the proteome's malleability since a canonical protein may not be the transcript's sole translation product. Such translation can occur from ribosomes bypassing a start codon or shifting frames during translation due to mRNA structure[45].

Half (50%; n = 225) of the non-canonical proteins originated from novel isoforms (Figure 6F and Supplementary Figure 8B). This group corresponds to proteins in frame with a canonical protein for which we either found few initiation events at the annotated start codon or the absence of an annotated start codon. Because their sequence overlaps with canonical proteins and their large size, these proteins were considered hereafter as novel isoforms. Consequently, for subsequent analyses, we analyzed novel isoforms separately from the rest of the non-canonical proteins. The remaining non-canonical proteins were further qualified as cryptic proteins.

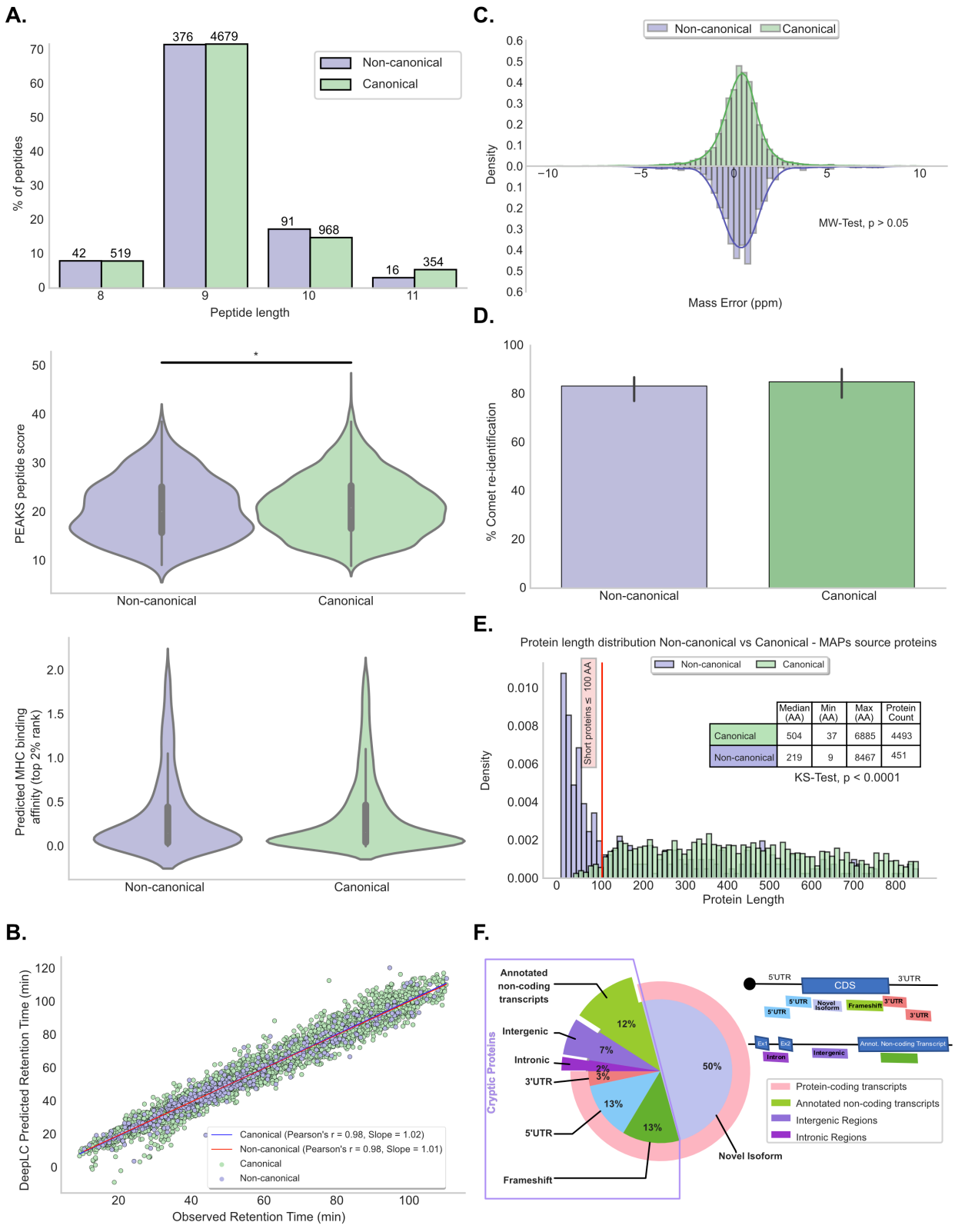


Figure 2. – Features of MAPs derived from canonical and non-canonical proteins.

(A-C) Displayed data refer to all canonical (n=6,520) and non-canonical (n=525) MAPs (total from 3 cell lines, 2 replicates each).
(A) Length, spectrum score (*p<0.05, T-test), MHC binding (p>0.05, Kolmogorov-Smirnov Test).
(B) Pearson correlations between observed and DeepLC-predicted retention times of MAPs derived from canonical and non-canonical proteins.
(C) Relative mass error of MAPs derived from canonical and non-canonical proteins. p> 0.05, two-sided Mann-Whitney U Test.
(D) Percentage of successful MAPs re-identification with Comet. p>0.05, two-sided Mann-Whitney U Test. Bar plot shows the median with error bars: 95% CI (n=3 cell lines).
(E) Length distribution of canonical (n=4,493) and non-canonical (n=451) MAPs source proteins. ****p<0.0001, Kolmogorov-Smirnov Test. Proteins with a length >800 amino acids are not displayed.
(F) Non-canonical MAPs source proteins derive from coding and noncoding transcripts. Pie chart showing the percentages of non-canonical proteins for each biotype and diagram illustrating how various types of transcripts were designated as a function of their genomic location.

1.9.3 Divergent Properties of Cryptic and Canonical MAP Source Proteins

Next, we elucidated the features of cryptic proteins, novel isoforms, and canonical MAP source proteins. By definition, canonical (annotated) proteins initiated almost exclusively (99.9%) on an AUG codon. Importantly, Ribo-seq TIS revealed that, first, 40% of newly identified proteins initiated on unannotated AUG initiation sites and, second, more than half of the cryptic and novel isoform MAP source proteins (53% and 67%, respectively) initiated from a non-AUG near-cognate codon (Figure 3A and Supplementary Figure 8C). As previously reported[10, 46-48], CUG was the most efficient codon at initiating unannotated proteins, though AAG was also frequently used, and others near-cognate codons were well represented.

In line with previous reports[49, 50], canonical MAPs derive from transcripts with higher expression than transcripts that do not generate MAPs (non-source transcripts; Figure 3B). Similarly, for cryptic MAPs and MAPs from novel isoforms, transcripts that generate MAPs are more abundant than non-source transcripts. Hence, for any genomic region, transcript levels positively correlated with MAP generation. Among MAP source transcripts, we found small but significant differences in abundance according to the following hierarchy: canonical proteins > novel isoforms > cryptic protein (median = 4.51 transcripts per million [TPM], 3.24 TPM, and 2.15 TPM, respectively; note that each cell has 500,000 mRNAs) (Figure 3B). Cryptic transcripts

contained significantly fewer exons, with a median of 2 exons compared to a median of 11 exons for transcripts coding for canonical proteins and novel isoforms (Figure 3C). Indeed, 73% of cryptic MAP source proteins contained only one or two exons.

Next, using Ribo-seq and RNA-seq data, we compared the translation efficiency of each MAP source transcript (translation events per mRNA) (Figure 3D). We observed that the translation efficiency of novel isoforms was only marginally inferior to that of canonical proteins, which in turn was similar to cryptic MAP source proteins. Among MAP source cryptic proteins, those deriving from an intergenic region showed the highest translation efficiency (Supplementary Figure 8D). We further examined how the subcellular localization of MAP source proteins influences translation efficiency (see STAR methods). We compared the translation efficiency of MAP source proteins from 6 subcellular localizations: cytosol, membrane, nucleus, extracellular, mitochondrion and secretory pathway. As a negative control, we computed the translation efficiency of the canonical proteins non-source of MAPs (background), independently of their localization. Two points can be made from these analyses. First, the translation efficiency of canonical proteins generating no MAPs was lower than that of MAP source proteins from any localization, except for proteins located in the nucleus (Supplementary Figure 8E and 8F). Second, proteins targeted to membranes or mitochondria were the most efficiently translated, followed by the secretory pathway and extracellular proteins.

Cryptic MAP source proteins had a mean length of only 49 amino acids compared to 504 and 582 residues for canonical proteins and novel isoforms, respectively (Figure 3E). For canonical proteins, the number of MAPs presented is related to protein length[50]. If this applies to all translation products, the short size of cryptic proteins should significantly decrease their chance of generating MAPs. In accordance with this, we validated that the number of identified MAPs increased linearly with source protein length (Supplementary Figure 9A). Then, for each protein, we calculated the number of amino acids detected in the immunopeptidome versus the number of amino acids in the source protein. This ratio was much higher for cryptic proteins versus canonical proteins (~5-fold) and novel isoforms (~7-fold) (Figure 3F). We conclude that, relative to canonical transcripts, cryptic transcripts are shorter, less abundant, and translated at similar efficiency but are ~5-fold more efficient at generating MAPs.

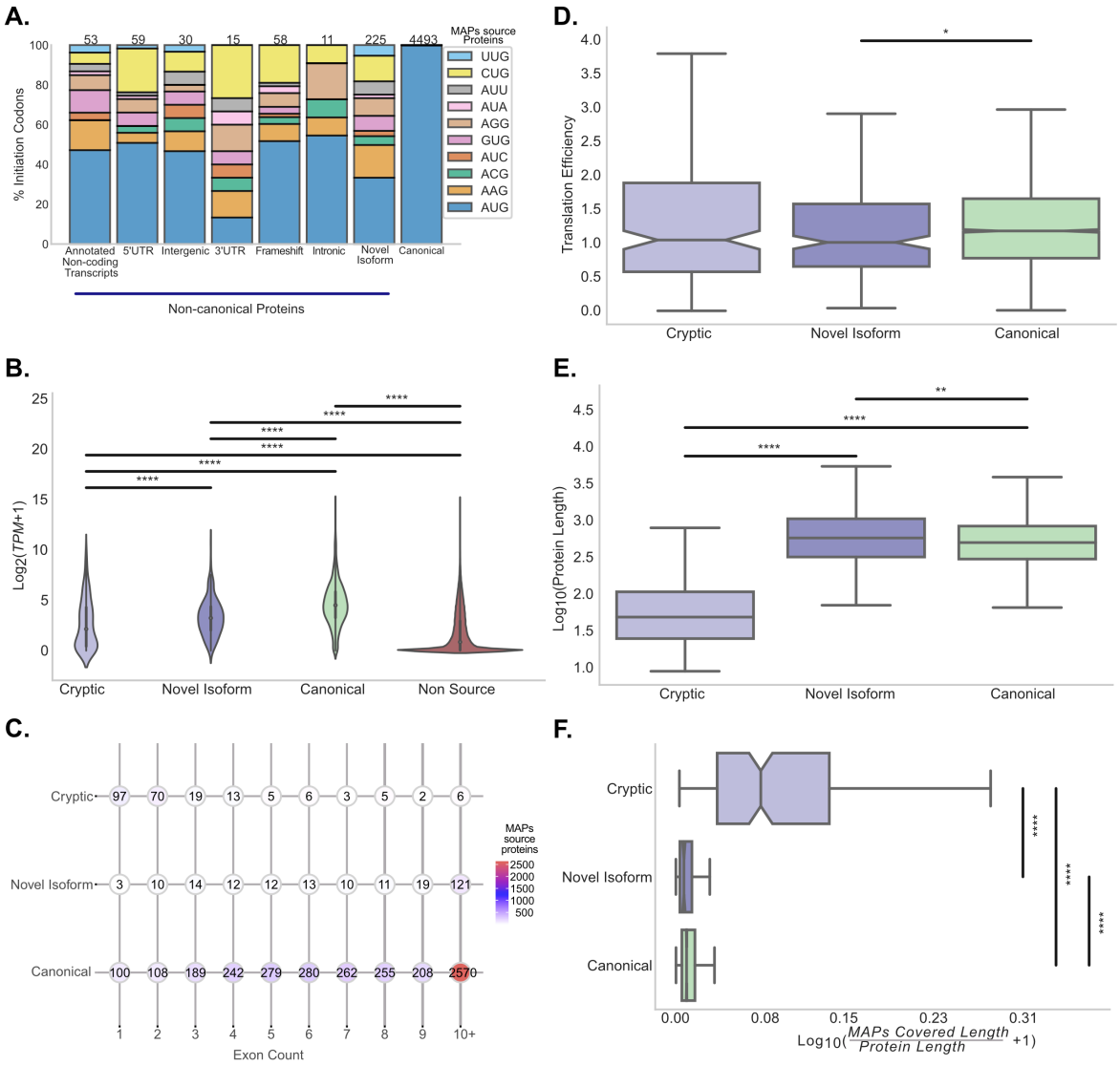


Figure 3. – Properties of MAP source proteins.

(A) More than a half of the non-canonical MAPs source proteins (60%) initiated at a near-cognate codon. Stacked bar-plot showing the percentage of proteins deriving from AUG and near-cognate codons for canonical proteins and various sub-group of non-canonical MAP source proteins.

(B) Transcript expression level distribution of canonical (n=4,493), novel isoforms (n=225) and cryptic (n=226) MAPs source transcripts vs. non-source proteins (n=647,686). ****p<0.0001, Kolmogorov-Smirnov test.

(C) Dot charts displaying the exons count for each category of MAP source proteins; each dot corresponds to the number of proteins bearing a given number of exons. Cryptic proteins have lower number of exons compared to novel isoform and canonical proteins (median = 2 exons for cryptic, 11 exons for novel isoform and canonical proteins).

(D) Translation efficiency of MAP source proteins. Boxplots show the translation efficiency distribution for each category of MAP source proteins. *p<0.05, two-sided Mann-Whitney U Test.

(E) Boxplots indicate the length distribution of MAPs source proteins for each category: cryptic, novel isoform and canonical. Median length in cryptic (49 amino acids), canonical (504 amino acids) and novel isoform (582 amino acids) is shown. **p<0.01, ****p<0.0001, two-sided Mann-Whitney U test.
(F) Cryptic proteins are proficient to generate MAPs. Boxplots show the ratio of the length covered by MAPs to the protein's length in number of amino acids. ****p<0.0001, two-sided Mann-Whitney U tests.

1.9.4 The Global Landscape of Cryptic Proteins in the Whole Cell Proteome

MS protein detection is proportional to protein abundance and length[51]. To enhance cryptic protein detection in whole-cell extracts of the three DLBCL lines, we performed tandem analyses on fractions separated by molecular weight before trypsin digestion. Low-molecular-weight fractions (≤10 kDa) contained proteins bearing less than ~100 amino acids, whereas high molecular weight (>10 kDa) contained longer proteins. We used PEAKs software to identify tryptic peptides of 7 and 25 amino acids and used the same strategy to assign the most likely source protein as for MAPs (FDR < 1%) (Figure 4A).

We identified 1,505 low- and 10,463 high-molecular-weight proteins. The vast majority of low-molecular-weight proteins were cryptic (81%), with canonical proteins (91%) dominating the high-molecular-weight fraction (Figure 4B). Interestingly, intergenic regions are the principal source of high-molecular-weight cryptic proteins (33%), although most (55%) low-weight cryptic proteins derive from protein-coding transcripts, with significant enrichment for 5' UTR-encoded proteins (34%; Figure 4C). Similar to MAP source proteins (Figure 3E), cryptic proteins identified in whole-proteome analyses were significantly shorter than canonical proteins and novel isoforms (median size of 387 amino acids for canonical proteins; 372 for novel isoforms versus 67 for cryptic proteins) (Figure 4D).

Cryptic proteins from whole-proteome extracts initiated less frequently at an AUG codon (23%; Figure 4E) than cryptic proteins detected in the immunopeptidome (40%; Figure 3A). Indeed, CUG (21%) was nearly as likely as AUG (23%) to initiate translation of cryptic proteome proteins. As with the immunopeptidome, transcripts coding MS-identified proteins were more abundant than transcripts coding for undetected proteins (Figure 4F). And, as with MAP source

proteins, the translation efficiency of cryptic proteins detected in the whole proteome was similar to that of canonical proteins and slightly superior to that of novel isoforms (Figure 4G).

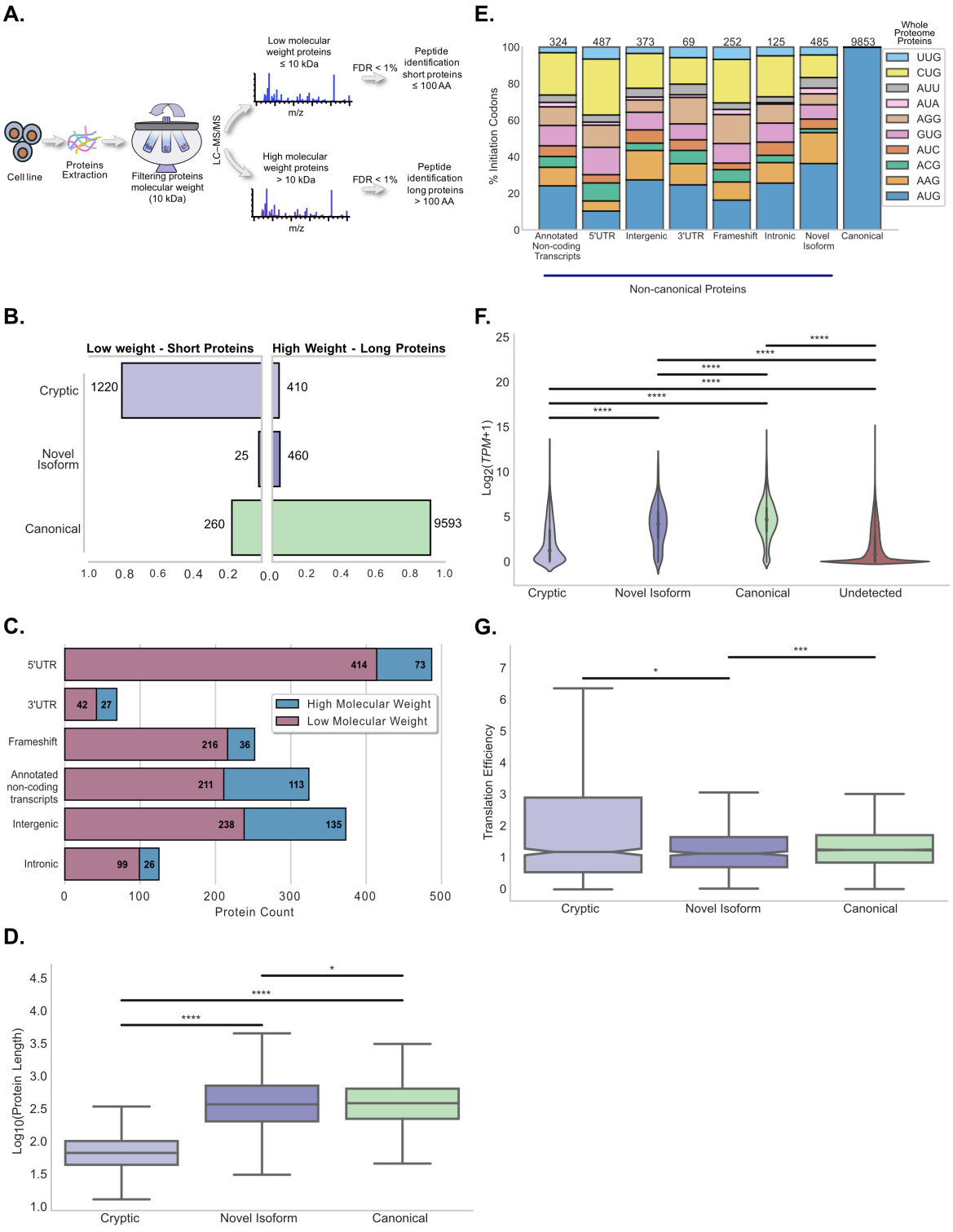


Figure 4. – Features of canonical and cryptic proteins detected in tryptic digests of whole cell extracts.

(A) Schematic overview of the method used for whole-proteome analyses. Proteins were filtered according to their molecular weight to maximize the detection of short proteins, which are a rich source of cryptic proteins.

(B-D) Displayed data refer to 3 cell lines, 1 replicate each.

(B) Proportion of each protein category detected in low- versus high-molecular-weight fractions. Low-weight fraction is enriched in cryptic proteins, whereas high-weight fraction is enriched in canonical proteins.

(C) Genomic origin of cryptic proteins identified in the whole-proteome extracts.

(D) Boxplots indicating the length distribution of proteins for each category: cryptic, novel isoform, and canonical. Median length of cryptic (67 amino acids), canonical (387 amino acids), and novel isoform (372 amino acids) proteins is shown. *p<0.05, ****p<0.0001, two-sided Mann–Whitney U test.

(E) Stacked bar plot showing the percentage of proteins deriving from AUG and near-cognate codons for canonical proteins along with each subgroup of the unannotated proteins from whole-proteome extracts.

(F) RNA expression level of transcripts coding for detected (n=11,968) proteins compared to transcripts coding for undetected proteins (n=640,662). ****p<0.0001, Kolmogorov-Smirnov test**.**

(G) Boxplots showing the translation efficiency of various categories of proteins identified from whole proteome extracts. *p<0.05, **p<0.01, two-sided Mann-Whitney U Test.

1.9.5 Disorder and Instability of Cryptic MAP Source Proteins

Even for conventional proteins, the whole-cell proteome only partially overlaps with the immunopeptidome[50, 52-54]. Thus, we detected only 52% (2,351 out of 4,493) of conventional MAP source proteins in whole proteomes (Figure 5A). Notably, this ratio decreased to 6% (14/226) in the case of cryptic MAP source proteins: why such a dramatic discrepancy?

First, consistent with the idea that MS favors detecting abundant proteins, the low expression of cryptic MAPs source transcripts (relative to canonical MAP source transcripts) hampers their detection in the whole proteome (Figure 3B). Accordingly, transcript expression correlates with detecting MAP source proteins in the whole-cell proteome (Figures 5B and 5C, left panels). Leveraging our Ribo-seq data, we determined that translation level (ribosome occupancy) was higher in proteome-detected vs. non-detected MAP source proteins, confirming that protein abundance impacts MS detection (Figures 5B and 5C, right panels). Second, detecting

cryptic proteins in whole proteomes is hampered by their brevity, which alone results in zero to few (median=3) predicted tryptic peptides per protein compared to 23 for conventional proteins (Figure 5D). Third, we considered the contribution of rapid degradation. Proteasomal digestion is the main route for protein degradation and MAP generation[55]. Proteasomes initiate degradation at disordered substrate regions; most, but not all, substrates need to be ubiquitylated, particularly for MAP generation[56]. We found a lower density of degradation signals (ubiquitination sites, D box, and KEN box motifs)[57-59] in cryptic relative to canonical proteins (Figure 5E). However, protein disorder analysis revealed that disordered regions occurred at twice the frequency in cryptic (31% of amino acids) versus conventional MAP source proteins (15% of amino acids) (Figure 5F). The instability index[60] also predicts the decreased stability of cryptic proteins (Figure 5G).

Finally, we analyzed the correlation between ribosome stalling in up-, mid-, and downstream coding regions of MAP-source transcripts and their detection in the proteome (Supplementary Figure 9B). We found a small but significant decrease in ribosome coverage in the upstream coding region of proteome-detected proteins, consistent with diminished stalling relative to non-detected proteins. These data collectively indicate that MAP source cryptic proteins contain zero to very few tryptic peptides, are low-abundance proteins generated with fewer stalling events, and are highly disordered and unstable. These factors likely account for their over-representation in the immunopeptidome and under-representation in the proteome.

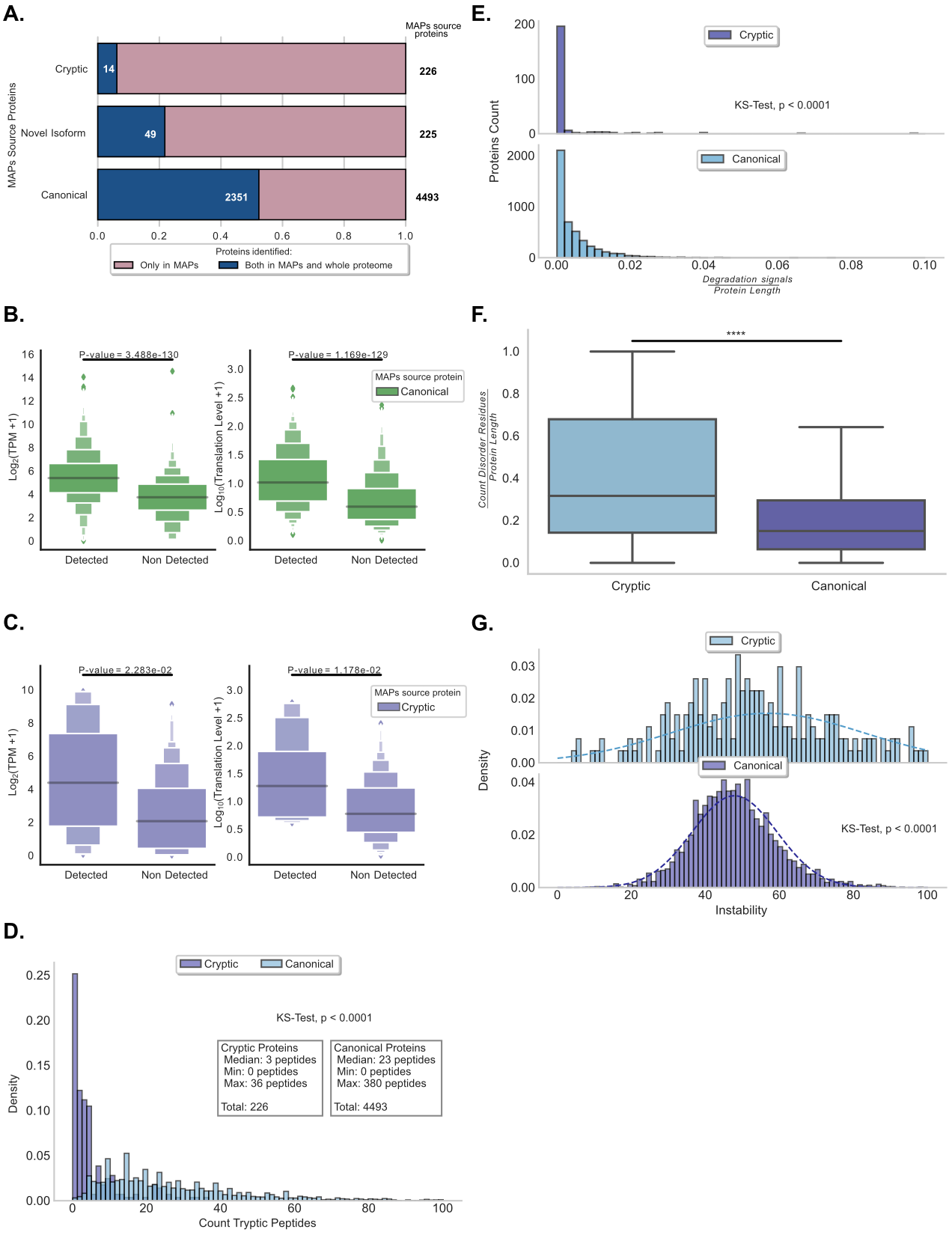

A.
MAPs source proteins
Cryptic
Novel Isoform
Canonical
14
49
2351
226
225
4493
MAPs Source Proteins
0.0
0.2
0.4
0.6
0.8
1.0
Proteins identified:
Only in MAPs
Both in MAPs and whole proteome
B.
P-value = 3.488e-130
P-value = 1.169e-129
MAPs source protein
Canonical
Log2(TPM +1)
Log10(Translation Level +1)
Detected
Non Detected
C.
P-value = 2.283e-02
P-value = 1.178e-02
MAPs source protein
Cryptic
D.
Cryptic
Canonical
KS-Test, p < 0.0001
Cryptic Proteins
Median: 3 peptides
Min: 0 peptides
Max: 36 peptides
Total: 226
Canonical Proteins
Median: 23 peptides
Min: 0 peptides
Max: 380 peptides
Total: 4493
Density
Count Tryptic Peptides
E.
Cryptic
Canonical
KS-Test, p < 0.0001
Proteins Count
Degradation signals / Protein Length
F.
****
Count Disorder Residues / Protein Length
Cryptic
Canonical
G.
Cryptic
Canonical
KS-Test, p < 0.0001
Density
Instability

Figure 5. – Cryptic proteins are disordered and unstable.
(A) MAP source proteins are underrepresented in the whole-proteome analysis. Bar plot depicting the total number of proteins identified in the immunopeptidome (pink bars) and the overlap with proteins detected in the whole proteome (blue bars) is shown. Cryptic proteins showed a low overlap (6%) compared to novel isoforms (21%) and canonical proteins (52%).
(B) Transcription- and translation-level abundance of canonical MAP source proteins. Left panel: box plots show the transcription expression level of transcripts at the origin of canonical MAP source proteins detected and non-detected in the whole-proteome analysis. Right panel: box plots show the translation level of transcripts at the origin of canonical MAP source proteins detected and non-detected in the whole-cell proteome analysis. Statistical difference was assessed by Mann-Whitney U test.
(C) Transcription- and translation-level abundance of cryptic MAP source proteins. Left panel: box plots show the transcription expression level of transcripts at the origin of cryptic MAP source proteins detected and non-detected in the whole-proteome analysis. Right panel: box plots show the translation level of transcripts at the origin of cryptic MAP source proteins detected and non-detected in the whole-cell proteome analysis. Statistical difference was assessed by Mann-Whitney U test.
(D) Distribution of the number of predicted tryptic peptides per MAP source protein (median = 3 peptides for cryptic proteins and 23 peptides for canonical proteins). Statistical significance was assessed by Kolmogorov-Smirnov test.
(E) Cryptic proteins present fewer degradation signals compared to canonical proteins. Histogram plots in the top and bottom panels depict the number of predicted degradation signal (canonical ubiquitination sites, D box, and KEN box motifs) relative to the protein size for cryptic and canonical proteins, respectively. Statistical significance was assessed by Kolmogorov-Smirnov test.
(F) Cryptic proteins contain significantly more disordered residues than canonical proteins. Boxplots depicting the number of disordered residues predicted per protein relative to the protein's length for cryptic and canonical proteins source of MAPs are shown. ****$p < 0.0001$; two-sided Wilcoxon rank-sum test.
(G) Cryptic proteins are less stable in vivo. Histogram plot showing the distribution of the instability index predicted for cryptic and canonical proteins. Statistical significance was assessed by Kolmogorov-Smirnov test.

1.9.6 Features of Non-canonical Proteins

We next evaluated several features of non-canonical proteins identified in the immunopeptidome and/or the whole proteome of the DLBCL lines (Supplementary Figure 10A). Non-canonical proteins demonstrate little bias in chromosomal origin (Figure 6A). However, chromosome 16 derived proteins exhibited an increased proportion of novel isoforms. This may result from cytogenetic abnormalities involving chromosome 16 in DLBCL[61]. Notably, an unexpectedly high

proportion of MAP source proteins derived from chromosome 12 (Supplementary Figure 10B), consistent with the shared DLBCL abnormalities (e.g., polysomy) involving chromosome 12[62, 63]. Overall, these findings indicate that, although all chromosomes generate numerous non-canonical proteins, their expression can be enhanced by cancer-associated genetic alterations.

Novel isoforms constitute a major fraction of unconventional proteins (28%) (Figure 6B and Supplementary Figure 10C). Alternative start codon initiation resulting in alternative protein isoforms translation is a common event in cancer[64]. It affects the balance between multiple forms of a protein, which can have distinct and even opposite functions. We interrogated our dataset to identify signaling pathways enriched among the canonical genes generating these novel isoforms (n = 403) (Figure 6C). Interestingly, these genes were mostly involved in signaling pathways often deregulated in cancer, including AXIN, mitogen-activated protein kinase 4 (MAPK4), MAPK6, NOTCH1, NOTCH4, PTEN, RUNX3, and transforming growth factor b (TGF-b). NOTCH signaling, which is commonly perturbed in DLBCL and other cancers[65, 66], was the most overrepresented in our analysis.

5’ UTRs represented the second most important cryptic protein source (21%;) (Figure 6B and Supplementary Figure 10C). Because upstream ORFs can modulate translation of main-ORFs[67], we examined how canonical protein translation is altered by upstream 5’ UTR translation of a cryptic protein. We found that the canonical ORF of transcripts encoding 5’ UTR cryptic proteins had significantly lower ribosome occupancy than those encoding 3’ UTR and frameshift proteins (Figure 6D). This observation suggests that translating cryptic 5’ UTR proteins hijacks ribosomes to hamper translation of the corresponding main ORF.

Finally, to evaluate the potential impact of 5’ UTR cryptic proteins on cell function, we used the reactome pathways annotation to analyse pathways associated with the genes encoding these proteins (n = 501). We found a conspicuous enrichment in genes involved in transcription, translation, and antiviral responses (Figure 6E), consistent with a functional role for 5’ UTR cryptic proteins in regulating various cellular processes.

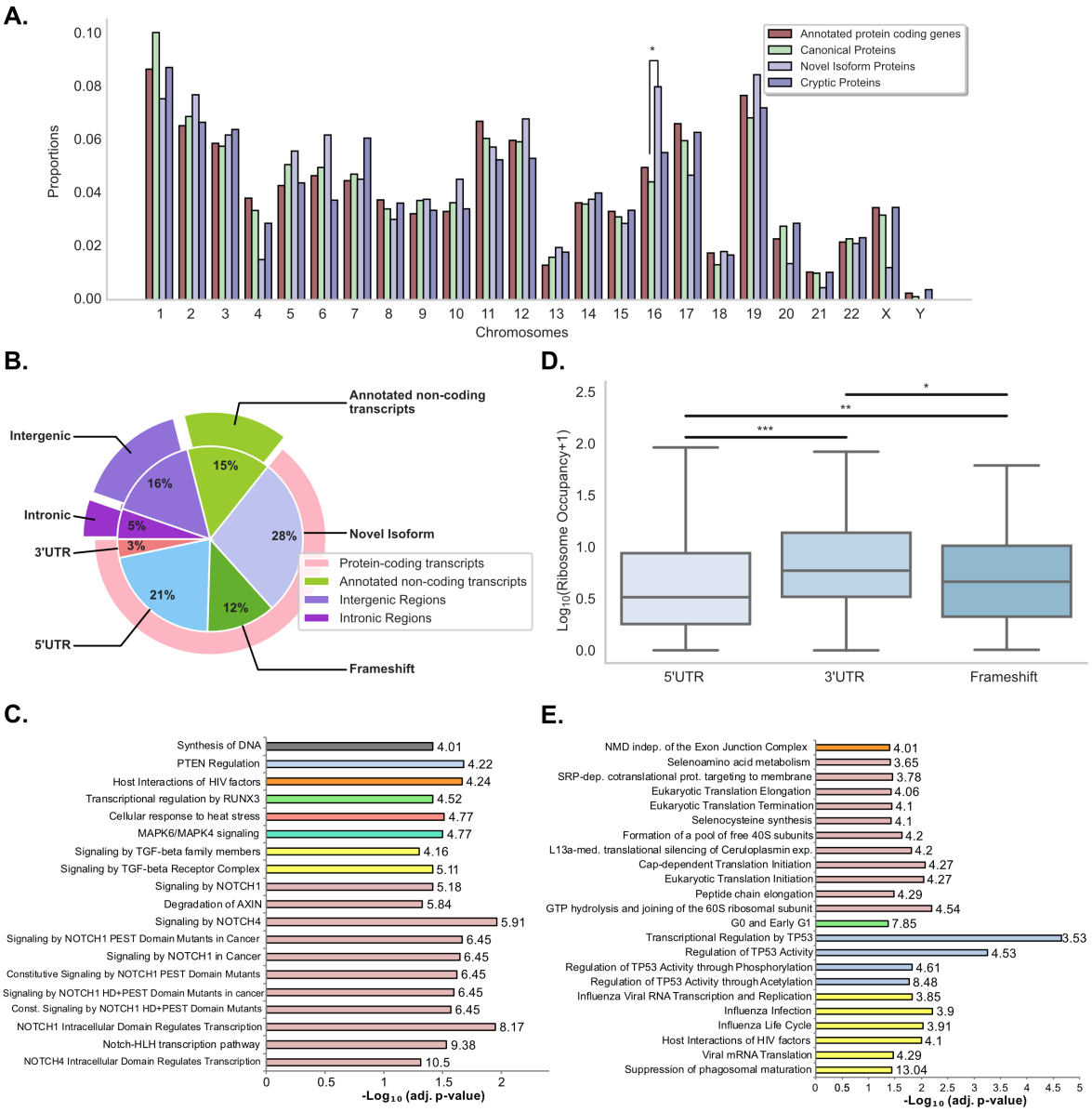


Figure 6. – Chromosomal origin and function of non-canonical proteins.

(A) Non-canonical identified proteins derive from all chromosomes. Bar graph shows the chromosomal origin of each category of proteins. *p<0.05, two-sided Fisher's exact test.

(B) Genomic origins of the whole set of non-canonical identified proteins. Pie chart shows the percentages of unannotated proteins derived from different genomic regions.

(C) Novel isoforms derive from genes that regulate pathways commonly perturbed in DLBCL and other cancers. Reactome pathways enriched in the list of genes corresponding to proteins for which a novel isoform was identified (n=403 unique genes). Panther over-representation test; numbers in the bar graph correspond to fold enrichment of each pathway. Fisher's exact test with FDR correction, adj. p-value<0.05, fold enrichment >4.

(D) 5’UTR cryptic proteins hinder translation of main-ORFs. Ribosome occupancy of the canonical coding sequence (CDS) of genes producing a cryptic protein via frameshift, 5’UTR or 3’UTR translation. *p<0.05, **p<0.01, ***p<0.001, two-sided Mann-Whitney U test.
(E) 5’UTR cryptic proteins regulate the translation of canonical proteins involved in transcription, translation and antiviral responses (n=501 unique genes). Panther over-representation test; numbers on the bar graph correspond to fold enrichment of each pathway. Fisher’s exact test with FDR correction, adj. p-value<0.05, fold enrichment >3.

1.10Discussion

We have developed a proteogenomic method to identify unannotated proteins whose peptides are detected in the whole-cell proteome and immunopeptidome. Our approach, which integrates RNA-seq, Ribo-seq, and MS data, identified 2,503 new non-canonical human proteins expressed from all chromosomes: 1,842 cryptic proteins (72%) and 661 novel isoforms (28%) (Figure 6B). As expected, a majority (85%) of translation events detected by Ribo-seq was not identified by MS (Tables 2 and 3). This was remarkably conspicuous for non-canonical proteins, as only 0.44% could be found by MS. Two facts can explain this. First, Ribo-seq-built databases must, to some extent, overestimate real translation products, especially non-canonical ones, due to imperfect sequence matching with genomic information. Second, and more importantly, MS captures only a small fraction of what is translated. Despite these caveats, our findings clearly demonstrate that ribosome profiling is a powerful tool to detect the translation of non-canonical transcripts, which are generally absent from MS databases because of their unannotated status.

Cryptic proteins are particularly interesting: 83% derived from ostensibly non-coding ORFs and 17% from alternative frame translation of canonical ORFs. Cryptic transcripts were slightly less abundant than canonical transcripts. Integrating Ribo-seq and RNA-seq data reveals that cryptic and canonical proteins are, surprisingly, translated with similar efficiency. Extending previous findings[21-24], cryptic proteins are coded by relatively short ORFs and frequently initiate with non-AUG near-cognate codons (which except for CUG[10], are typically decoded as Met[68]). Cryptic proteins were far more likely than canonical proteins to be only detected in the immunopeptidome.

Critically, cryptic transcripts generated MAPs ~5-fold more efficiently than canonical transcripts (Figure 3F). The most plausible explanation is that cryptic proteins are rapidly degraded because they are disordered and unstable (Figures 5F and 5G), rendering them prototypical DRiPs. As a corollary, the global proteome, mainly consisting of stable proteins, has limited overlap with the immunopeptidome. Remarkably, only 6% of cryptic MAP source proteins were detected in tryptic digests of whole cell extracts (Figure 5A). Such selective antigenicity is a critical feature of class I antigen presentation, which cannot function as a mirror of the proteome,

which is dominated by a relatively small number of gene products (just 250 housekeeping proteins comprising ~50% of the proteome). This could also be explained by the few predicted tryptic sites in cryptic proteins (Figure 5D), consistent with a negative bias in detecting short proteins due to the standard enzyme used in proteomic analysis.

Stable isotope labeling with amino acids in cell culture (SILAC) mass spectrometry kinetic studies in tumor cell lines also point to a limited correlation between the proteome and immunopeptidome and suggest a substantial contribution of DRiPs/SLiPs as a source of MAPs[16, 17]. Most short-lived MAP source proteins identified by SILAC MS kinetic analyses are subunits of multiprotein complexes. These likely become SLiPs due to stoichiometric subunits imbalances or other difficulties in becoming incorporated into their intended complex. A large fraction of MAPs identified in the present study would be missed entirely in such SILAC MS kinetic analyses due to the method-inherent shortcomings (e.g., search database limited by annotated proteins; failure to detect [tryptic] peptide in multiple time points and samples in SILAC MS analysis to determine MAP source;[16, 17]). This would bias the identification of DRiP-derived MAPs to longer and more-abundant source proteins.

Cryptic proteins detected in the cell proteome were longer (median of 67 amino acids) than those found in the immunopeptidome (median of 49 amino acids) (Figures 3E and 4D), likely a reflection of the likelihood that longer peptides can achieve a more-stable structure. Cryptic proteins detected in the immunopeptidome were initiated more frequently at an AUG codon than those found in the whole proteome (Figures 3A and 4E). This suggests that a subset of proteins initiated in AUG codons may have preferred access to the MHC-I presentation pathway, extending findings that CUG and other near-cognate-based initiation favor peptide generation under stress conditions[69].

Whereas intergenic regions are the primary source of longer cryptic proteins found in the whole proteome, translation of 5’ UTRs was particularly common for shorter cryptic proteins found both in the immunopeptidome and the whole-cell proteome. Notably, translation of 5’ UTR cryptic proteins correlated with decreased ribosome occupancy of the main ORF, which was not seen with cryptic proteins derived from other regions in protein- coding transcripts (3’ UTR and

frameshift) (Figure 6D). The main ORFs whose translation was hindered by 5' UTR cryptic proteins mainly regulate transcription, translation, and antiviral responses (Figure 6E). Translation of 5' UTRs is known to negatively regulate translation of downstream ORF in cell stress[67, 70, 71]. Our findings suggest that this extends to cryptic proteins. Additional studies are needed to generalize these findings from DLBCLs to other cancer cells and normal cells.

The 661 novel isoforms reported herein further illustrate the polycistronic nature of human genes[2]. Arguably, their most intriguing feature was that they showed a strong enrichment for signaling pathways deregulated in cancer, NOTCH being the most striking example (Figure 6C). Chromosome 16 was a particularly rich source of novel isoforms (Figure 6A). Accordingly, in DLBCLs, this chromosome commonly presents aberrations (e.g., duplications and trisomies), whose frequency increases with patient age[61]. We also observed that chromosome 12, which is also commonly rearranged in DLBCLs, was a particularly rich source of cryptic MAPs. Together, these data suggest that underlying genomic aberrations may impact the non-canonical translation landscape by increasing the production of novel isoforms or cryptic proteins. How this affects the presentation of tumor-specific antigens that can be targeted for immunotherapy will be explored in further studies.

We detected only a small number of peptides from introns (135/7,045) (Supplementary Figure 10A). Based on studies that peptides are efficiently derived from introns via translation of pre-spliced mRNA in the nucleus[43, 72], this is surprising, particularly given the fact that introns encode up to 10-fold more amino acids than exons [73]. However, we note that, by performing Ribo-seq on cytoplasmic RNA, we may have missed a large pool of intron-encoded peptides translated in the nucleus.

Finally, it is worth considering the biological relevance of non-canonical translation of unstable proteins. This might result from the high entropy of cancer cells, which evolve to maximally proliferate at the organism's cost, with little or no selection for the economical use of available resources. It is likely, however, that at least some of the gene products have functions, particularly if their degradation is conditionally regulated, for example, by the cell cycle or stress.

A more-general function of this class of proteins would be to enhance tumor immunosurveillance. The cancer-specific nature of such translation is an obvious starting point for future studies.

1.11Acknowledgments

We thank Pierre-Henri Wuillemin and Natalia Sokolovska for invaluable support in model design. We thank Qingchuan Zhao, Anca Apavaloaei and Assya Trofimov for sound biological insights and all other members of our laboratories for their thoughtful suggestions. We also thank Patrick Gendron and Jean-Philippe Laverdure for assistance with bioinformatics tools. This work was supported by grants from IVADO and the Canada First Research Excellence Fund (Apogée/CFREF), the Canadian Cancer Society (#705604), the Canadian Institutes of Health Research (FDN 148400) and The Oncopole (EMC^2 Grant). Jaroslav Holly and Jonathan Yewdell are supported by the Division of Intramural Research, NIAID, and Louis Staudt by the Division of Intramural Research, NCI. This work was additionally generously supported by a FLEX grant from the Division of Intramural Research, NCI. We would also like to thank the CCR Sequencing Facility at Frederick National Laboratory for Cancer Research for performing the sequencing.

1.12Declaration of interest

The authors declare no competing interests.

Received: August 3, 2020

Revised: January 29, 2021

Accepted: February 10, 2021

Published: March 9, 2021

1.13References

1. Ingolia, N.T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**, 22-33 (2016).
2. Brunet, M.A., Levesque, S.A., Hunting, D.J., Cohen, A.A. & Roucou, X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res* **28**, 609-624 (2018).
3. Lu, S. et al. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res*, 8111-8125 (2019).
4. Brunet, M.A. et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res* **47**, D403–D410 (2018).
5. van Heesch, S. et al. The Translational Landscape of the Human Heart. *Cell* **178**, 242-260 e229 (2019).
6. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140-1146 (2020).
7. Samandi, S. et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife* **6**, e27860 (2017).
8. Erhard, F. et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**, 363–366 (2018).
9. Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F. & Baranov, P.V. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res* **39**, 4220-4234 (2011).
10. Starck, S.R. et al. Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. *Science* **336**, 1719-1723 (2012).
11. Fields, A.P. et al. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* **60**, 816-827 (2015).
12. Yewdell, J.W. Immunology. Hide and seek in the peptidome. *Science* **301**, 1334-1335 (2003).
13. Schubert, U. et al. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **404**, 770-774 (2000).

14. Qian, S.B., Princiotta, M.F., Bennink, J.R. & Yewdell, J.W. Characterization of rapidly degraded polypeptides in mammalian cells reveals a novel layer of nascent protein quality control. *J Biol Chem* **281**, 392-400 (2006).
15. Reits, E.A., Vos, J.C., Gromme, M. & Neefjes, J. The major substrates for TAP in vivo are derived from newly synthesized proteins. *Nature* **404**, 774-778 (2000).
16. Bourdetsky, D., Schmelzer, C.E. & Admon, A. The nature and extent of contributions by defective ribosome products to the HLA peptidome. *Proc Natl Acad Sci U S A* **111**, E1591-1599 (2014).
17. Milner, E., Barnea, E., Beer, I. & Admon, A. The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol Cell Proteomics* **5**, 357-365 (2006).
18. Prevosto, C. et al. Allele-Independent Turnover of Human Leukocyte Antigen (HLA) Class Ia Molecules. *PLoS One* **11**, e0161011 (2016).
19. Blaha, D.T. et al. High-Throughput Stability Screening of Neoantigen/HLA Complexes Improves Immunogenicity Predictions. *Cancer Immunol Res* **7**, 50-61 (2019).
20. Dersh, D., Holly, J. & Yewdell, J.W. A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion. *Nat Rev Immunol* (2020).
21. Laumont, C.M. et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun* **7**, 10238 (2016).
22. Laumont, C.M. et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med* **10**, eaau5516 (2018).
23. Chong, C. et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* **11**, 1293 (2020).
24. Ouspenskaia, T. et al. Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. *bioRxiv*, 2020.2002.2012.945840 (2020).
25. Zhao, Q. et al. Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer. *Cancer Immunol Res* **8**, 544-555 (2020).

26. Caron, E. et al. The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol Syst Biol* **7**, 533 (2011).
27. Wei, J. et al. Ribosomal Proteins Regulate MHC Class I Peptide Generation for Immunosurveillance. *Mol Cell*, 1162-1173 (2019).
28. Purcell, A.W., Ramarathinam, S.H. & Ternette, N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc* **14**, 1687-1707 (2019).
29. Ghosh, M. et al. Guidance Document: Validation of a High-Performance Liquid Chromatography-Tandem Mass Spectrometry Immunopeptidomics Assay for the Identification of HLA Class I Ligands Suitable for Pharmaceutical Therapies. *Mol Cell Proteomics* **19**, 432-443 (2020).
30. Courcelles, M. et al. MAPDP: A Cloud-Based Computational Platform for Immunopeptidomics Analyses. *J Proteome Res* **19**, 1873-1881 (2020).
31. Vizcaino, J.A. et al. The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. *Mol Cell Proteomics* **19**, 31-49 (2020).
32. Nesvizhskii, A.I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**, 2092-2123 (2010).
33. Nesvizhskii, A.I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114-1125 (2014).
34. Finotello, F., Rieder, D., Hackl, H. & Trajanoski, Z. Next-generation computational tools for interrogating cancer immunity. *Nat Rev Genet* **20**, 724-746 (2019).
35. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223 (2009).
36. Blakeley, P., Overton, I.M. & Hubbard, S.J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**, 5221-5234 (2012).
37. Capietto, A.H. et al. Mutation position is an important determinant for predicting cancer neoantigens. *J Exp Med* **217** (2020).

38. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15**, 8125-8148 (1987).
39. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature Communications* **11**, 1759 (2020).
40. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *bioRxiv*, 2020.2003.2028.013003 (2020).
41. Eng, J.K. et al. A Deeper Look into Comet—Implementation and Features. *Journal of The American Society for Mass Spectrometry* **26**, 1865-1874 (2015).
42. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
43. 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* **110**, 17951-17956 (2013).
44. 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* **92**, 7976-7980 (1995).
45. Bullock, T.N. & Eisenlohr, L.C. Ribosomal scanning past the primary initiation codon as a mechanism for expression of CTL epitopes encoded in alternative reading frames. *J Exp Med* **184**, 1319-1329 (1996).
46. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
47. Ivanov, I.P., Loughran, G., Sachs, M.S. & Atkins, J.F. Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc Natl Acad Sci U S A* **107**, 18056-18060 (2010).
48. Lee, S. et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **109**, E2424-2432 (2012).

49. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics* **14**, 658-673 (2015).
50. Pearson, H. et al. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest* **126**, 4690-4701 (2016).
51. Lubec, G. & Afjehi-Sadat, L. Limitations and pitfalls in protein identification by mass spectrometry. *Chem Rev* **107**, 3568-3584 (2007).
52. Granados, D.P., Laumont, C.M., Thibault, P. & Perreault, C. The nature of self for T cells-a systems-level perspective. *Curr Opin Immunol* **34**, 1-8 (2015).
53. Yewdell, J.W., Dersh, D. & Fahraeus, R. Peptide Channeling: The Key to MHC Class I Immunosurveillance? *Trends Cell Biol* **29**, 929-939 (2019).
54. Shraibman, B. et al. Identification of Tumor Antigens Among the HLA Peptidomes of Glioblastoma Tumors and Plasma. *Mol Cell Proteomics* **18**, 1255-1268 (2019).
55. Myers, N. et al. The Disordered Landscape of the 20S Proteasome Substrates Reveals Tight Association with Phase Separated Granules. *Proteomics* **18**, e1800076 (2018).
56. Wei, J. et al. Varied Role of Ubiquitylation in Generating MHC Class I Peptide Ligands. *J Immunol* **198**, 3835-3845 (2017).
57. Liu, Z. et al. GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. *PLoS One* **7**, e34370 (2012).
58. Radivojac, P. et al. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* **78**, 365-380 (2010).
59. Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, W329-W337 (2018).
60. Guruprasad, K., Reddy, B.V. & Pandit, M.W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* **4**, 155-161 (1990).
61. Vick, E.J. et al. Age-Related Chromosomal Aberrations in Patients with Diffuse Large B-Cell Lymphoma: An In Silico Approach. *World J Oncol* **9**, 97-103 (2018).

62. Chan, W.Y., Wong, N., Chan, A.B., Chow, J.H. & Lee, J.C. Consistent copy number gain in chromosome 12 in primary diffuse large cell lymphomas of the stomach. *Am J Pathol* **152**, 11-16 (1998).
63. Younes, A. et al. Polysomy of chromosome 12 in 60 patients with non-Hodgkin's lymphoma assessed by fluorescence in situ hybridization: differences between follicular and diffuse large cell lymphoma. *Genes Chromosomes Cancer* **9**, 161-167 (1994).
64. Xu, Y. & Ruggero, D. The Role of Translation Control in Tumorigenesis and Its Therapeutic Implications. *Annual Review of Cancer Biology* **4**, 437-457 (2020).
65. Karube, K. et al. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia* **32**, 675-684 (2018).
66. Aster, J.C., Pear, W.S. & Blacklow, S.C. The Varied Roles of Notch in Cancer. *Annu Rev Pathol* **12**, 245-275 (2017).
67. Young, S.K. & Wek, R.C. Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *J Biol Chem* **291**, 16927-16935 (2016).
68. Na, C.H. et al. Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. *Genome Res* **28**, 25-36 (2018).
69. Starck, S.R. & Shastri, N. Nowhere to hide: unconventional translation yields cryptic peptides for immune surveillance. *Immunol Rev* **272**, 8-16 (2016).
70. Jiang, Z. et al. Ribosome profiling reveals translational regulation of mammalian cells in response to hypoxic stress. *BMC genomics* **18**, 638 (2017).
71. Reverendo, M., Mendes, A., Arguello, R.J., Gatti, E. & Pierre, P. At the crossway of ER-stress and proinflammatory responses. *FEBS J* **286**, 297-310 (2019).
72. Martins, R.P. et al. Nuclear processing of nascent transcripts determines synthesis of full-length proteins and antigenic peptides. *Nucleic Acids Res* **47**, 3086-3100 (2019).
73. Francis, W.R. & Wörheide, G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome Biology and Evolution* **9**, 1582-1598 (2017).
74. McGlincy, N.J. & Ingolia, N.T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112-129 (2017).

75. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. & Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**, 1534-1550 (2012).
76. Lauria, F. et al. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLoS Comput Biol* **14**, e1006169 (2018).
77. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).
78. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
79. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
80. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
81. Lanoix, J. et al. Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. *Proteomics* **18**, e1700251 (2018).
82. Jurtz, V. et al. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* **199**, 3360-3368 (2017).
83. Kote, S., Pirog, A., Bedran, G., Alfaro, J. & Dapic, I. Mass Spectrometry-Based Identification of MHC-Associated Peptides. *Cancers (Basel)* **12**, 535 (2020).
84. 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* **111**, 4507-4512 (2014).
85. Veres, D.V. et al. ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Res* **43**, D485-493 (2015).

1.14 STAR☆Methods

1.14.1.1 Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
InVivoMAb anti-human MHC Class I (W6/32)	BioXcell	Cat# BE0079; RRID: AB_1107730
Anti-human HLA-ABC (W6/32)	Biolegend	Cat# 311402; RRID: AB_314871
Mouse IgG2a, κ Isotype Ctrl Antibody	Biolegend	Cat# 400201
Chemicals, peptides, and recombinant proteins		
Advanced RPMI 1640 Medium	Thermo Fisher	Cat# 12633012
Fetal Bovine Serum	Seradigm	Cat# 1500-500
AIM V medium	Thermo Fisher	Cat# 12055091
Penicillin-Streptomycin (10,000 U/mL)	Thermo Fisher	Cat# 15140122
GlutaMAX Supplement	Thermo Fisher	Cat# 35050061
Gentamycin	Thermo Fisher	Cat# 15750060
Harringtonine	LKT Laboratories	Cat# H0169
DPBS, calcium, magnesium	GIBCO	Cat# 14040141
UltraPure Sucrose	Invitrogen	Cat# 15503022
TRI Reagent Solution	Invitrogen	Cat# AM9738
Cycloheximide, High Purity - CAS 66-81-9 - Calbiochem	Sigma-Aldrich	Cat# 239764

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RNase I, *E. coli*	Lucigen	Cat# N6901K
DNase I	Zymo Research	Cat# E1009-A
Nuclease-Free Water (not DEPC-Treated)	Invitrogen	Cat# AM9932
SDS, 20% Solution, RNase-free	Invitrogen	Cat# AM9820
Formic acid	Sigma-Aldrich	Cat#FX0440-7
C18 Jupiter Phenomenex	Phenomenex	Cat# 04A-4263
Acetonitrile	Thermo Fisher	Cat# A996SK-4
Ammonium bicarbonate	Sigma-Aldrich	Cat# A6141
TCEP [Tris(2-carboxyethyl) phosphine hydrochloride	Thermo Fisher	Cat# 20490
Chloroacetamide	Sigma-Aldrich	Cat# C0267
Trypsin	Promega	Cat# V511A
Critical commercial assays		
Universal Mycoplasma Detection Kit	ATCC	Cat# 30-1012K
Qubit RNA BR Assay Kit	Invitrogen	Cat# Q10211
Ribo-Zero Gold rRNA Removal Kit (Human, Mouse, Rat)	Illumina	Cat# MRZG12324
RNA Clean & Concentrator-5	Zymo Research	Cat# R1013
TruSeq Stranded mRNA Library Prep kit	Illumina	Cat# 20020594
PureProteome protein A magnetic beads	Millipore	Cat# LSKMAGA10
QIFIKIT	Agilent	Cat# K0078

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
DLBCL cell line samples: RNaseq data	This study	NCBI SRA: PRJNA647736
DLBCL cell line samples: Ribo-seq data	This study	NCBI SRA: PRJNA647736
DLBCL samples immunopeptidomic and whole proteome tryptic data	This study	PRIDE: PXD020620
Experimental models: cell lines		
HBL-1 cell line	Lab of Martin Dyer	RRID: CVCL_4213
SU-DHL-4 cell line	Lab of Mark Raffeld	RRID: CVCL_0539
DoHH2 cell line	DSMZ	Cat# ACC-47; RRID: CVCL_1179
Software and algorithms		
Ribo-db Pipeline	This study	https://github.com/lemieux-lab/Ribo-db
STAR	(Dobin et al., 2013)	https://github.com/alexdobin/STAR
SAMtools	(Li et al., 2009)	http://www.htslib.org/doc/
StringTie	(Pertea et al., 2015)	https://ccb.jhu.edu/software/stringtie/
PEAKS X	Bioinformatics Solutions	https://www.bioinfor.com/
Comet	(Eng et al., 2015)	http://comet-ms.sourceforge.net/
NetMHCpan 4.0	(Jurtz et al., 2017)	http://www.cbs.dtu.dk/services/NetMHCpan-4.0/

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Freebayes	(Garrison and Marth, 2012)	https://github.com/freebayes/freebayes
Proteowizard	Proteowizard Software	http://proteowizard.sourceforge.net
BEDtools	(Quinlan and Hall, 2010)	https://bedtools.readthedocs.io/en/latest/
MAPDP	(Courcelles et al., 2020)	https://gitlab.com/iric-proteo/mapdp
DeepLC 0.1.14	(Bouwmeester et al., 2020)	https://github.com/compomics/DeepLC
PRICE v.1.0.3	(Erhard et al., 2018)	https://github.com/erhard-lab/gedi/wiki/Price
GPS-ARM version 1.0	(Liu et al., 2012)	http://arm.biocuckoo.org/
UbPred	(Radivojac et al., 2010)	http://www.ubpred.org/
IUPred2	(Mészáros et al., 2018)	https://iupred2a.elte.hu/
Biopython module SeqUtils	Biopython module	https://biopython.org/docs/1.75/api/Bio.SeqUtils.html
Panther classification system	Panther algorithm	http://www.pantherdb.org/
riboWaltz	(Lauria et al., 2018)	https://github.com/LabTranslationalArchitectomics/riboWaltz

1.15 Resource availability

1.15.1 Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Claude Perreault (claude.perreault@umontreal.ca).

1.15.2 Materials Availability

This study did not generate new unique reagents.

1.15.3 Data and Code Availability

The Python, bash scripts and Jupyter notebooks generated during this study are available at GitHub, https://github.com/lemieux-lab/Ribo-db.

The accession number for MS raw data and associated databases reported in this paper is PRIDE: PXD020620.

The accession number for RNA-seq and ribosomal profiling raw sequencing data reported in this paper is NCBI SRA: PRJNA647736.

1.16 Experimental model and subject details

1.16.1 Cell lines

DLBCL lines HBL-1, DoHH2, and SU-DHL-4 bearing HLA A02:06, B51:01, C14:02; A01:01, B08:01, B44:02, C07:01, C07:04 and A02:01, A31:01, B15:01, C03:04, respectively, were cultured in complete medium consisting of Advanced RPMI medium (Gibco) supplemented with 5% heat-inactivated fetal bovine serum (Seradigm), 1% Penicillin/Streptomycin (Gibco), and Glutamax (Gibco). Cells were grown in humidified atmosphere at 37°C with 5% CO2 and routinely tested for mycoplasma contamination using Universal Mycoplasma Detection Kit (ATCC). Cell line identity was confirmed by copy number variant fingerprinting of 16 loci of each cell line genomic DNA (Jonathan Keats, personal communication). Sex of cells used: Male, HBL-1, DoHH2, SU-DHL-4.

1.17 Method details

1.17.1 Ribosomal profiling, RNA-seq sample preparation and sequencing

Ribosomal profiling was performed as previously described[74] with modifications as follows: DLBCL cell cultures were seeded at $2x10^5$ cells/ml in 50 mL of complete medium in duplicates for each cell line and condition. Enrichment for initiating ribosomes was done by treating the cell cultures with harringtonine (LKT Laboratories) at 5 µg/ml for 30 minutes at 37°C before harvesting. Thirty-six hours after seeding the cells were pelleted by centrifugation (300g, 5 min., RT), cell pellets were immediately put on ice and washed with ice-cold DPBS (GIBCO), centrifuged (300g, 5 min., 4°C) and cell pellets flash-frozen in liquid nitrogen. Samples processing proceeded without delay until sucrose cushion purified ribosomes were resuspended in TRI. Reagent Solution (Ambion) and stored at -80°C. Cycloheximide was included only in lysis buffer at 100 mg/ml. RNA concentration in cell lysates was quantitated by Qubit RNA BR Assay Kit (Invitrogen) using Qubit 4 fluorometer. The lysates containing 30 µg of RNA were diluted to the final volume of 200 µl with polysome buffer and treated with 15 U of RNase I (10 U/µl, Lucigen) at room temperature (24°C) for 45 min on tube rotator. The ribosomal RNA depletion was done in two steps: First, size-selected ribosome protected fragments were depleted by Ribo-Zero Gold rRNA Removal Kit (Human, Mouse, Rat) (Illumina). Second, circularized cDNA was depleted using biotinylated complementary oligonucleotides as previously described[75]. Ribosomal profiling libraries were sequenced on Illumina HiSeq 4000 to achieve 350-400 million raw reads per sample (~100 million for harringtonine treated samples). Ribosome profiling footprint library quality was assessed using riboWaltz[76] via trinucleotide codon periodicity plotting against annotated protein-coding ORFs. Ribosome profiling samples exhibiting clear trinucleotide periodicity were retained for subsequence ORF detection. RNAseq libraries were prepared from the same cell lysates as the ribosome profiling sequencing libraries. Five micrograms of RNA per sample lysate was diluted with nuclease-free water to the final volume of 40 µl, treated with DNase I (Zymo Research) at RT for 15 min, and diluted with sodium dodecyl sulfate solution to the final concentration of 1%. Total RNA was purified using RNA Clean & Concentrator-5 (Zymo Research). RNAseq libraries

were prepared using TruSeq Stranded mRNA Library Prep kit (Illumina) and sequenced as PE 75 cycles on Illumina NextSeq 550 to high depth.

1.17.2 Quantification MHC-I molecules per cell

MHC-I's absolute membrane density was evaluated on 3 DLBCL cell lines by indirect labeling with a purified anti-human HLA-ABC (clone W6/32) or a mouse IgG2a isotype control, using commercially available QIFIKIT (Dako) according to the manufacturer's instructions.

1.17.3 Ribosome Profiling data pre-processing

Illumina adapters from the 3' end of the Ribosome Profiling TIS and Elongation Sequencing Fragments (RPSF) were removed using fastx_clipper (http://hannonlab.cshl.edu/fastx_toolkit/). UMI detection and extraction were performed using UMI_tools[77]. Next, only relevant RPSFs (i.e., reads with a length between 26 and 34 nucleotides) were retained for further human genomic coordinate mapping (reference genome version GRCh38.p10/hg38) using STAR v.2.6.1.d[78]. We ran STAR with default settings except for the following modified parameters: --outSAMtype BAM SortedByCoordinate, --alignEndsType EndToEnd, --seedSearchStartLmax 15, --outFilterMismatchNoverLmax 0.05, --outFilterMatchNmin 25. Finally, the BAM files were deduplicated using UMI_Tools.

1.17.4 RNA-sequencing data pre-processing

Illumina adapters from the 3' end of the RNA-sequencing reads were removed using Trimmomatic version 0.35 and then mapped to the reference genome version GRCh38.p10/hg38 using STAR v.2.6.1.d[78].We ran STAR with default settings except for the following modified parameters: --outSAMtype BAM SortedByCoordinate, --outFilterMismatchNoverLmax 0.05, --outFilterMatchNmin 40.

1.17.5 Ribo-db approach: detection of active translation sequences

To generate a complete and noiseless sample-specific database suitable for MS searches, we translated in-silico the actively translated sequences (canonical and non-canonical ORF) assessed by combining Ribo and RNA-seq data as follows:

A) TIS calling: to detect sample-specific Translation Initiation Sites (TIS) from the aligned Ribo-seq TIS reads, we developed a probabilistic approach to estimate a confidence score used to identify the genomic positions of putative start codons that differentiates start codon positions from background, considering all near-cognate start codons.

To achieve this, we assumed that all annotated start codons aligning with Ribo-TIS reads were true start codons, from this we propose to estimate the probability of each position (pos) into each read length $l = (26, \ldots, 34)$, to act as the first nt of the ribosomal p-site therefore being the first nt of a start codon (sc), as follows:

Let $r = \{$reads being at the first nt of a start codon$|\ len = l, then\ pos = p\}$

Let $R = \{$total reads being at the first nt of a start codon$|len = l\}$;

$$P(sc \mid len = l,\ pos = p) = \frac{|r|}{|R|}$$

where $P(sc \mid len = l,\ pos = p)$ is the probability of a sc at the read position pos in the read of length $l = (26, \ldots, 34)$.

Then, we computed two heuristics to evaluate the certainty of the ribosomal P-site location into each read length l, and the relevance of the read-alignment regarding its multimapping.

The first heuristic $H_1(l)$ assigned a normalized weight to each read length (26-34 nt), computed through the standard deviation of the read positions acting as start codons, as follows:

Let $\sigma = \{\sigma_l |\ stdev\ of\ read\ positions\ acting\ as\ start\ codons\ for\ l = (26, \ldots, 34)\}$;

$$H_1(\mathrm{l}) = 1 - \left(\frac{\sigma_l - \min(\sigma)}{\max(\sigma) - \min(\sigma)} * 0.99\right)$$

The second heuristic $H_2(R_r)$ assigned a weight to each Ribo-Tis read according to its rank (R_r) in which STAR has reported such alignments, as follows:

$$H_2(R_r) = 1 - \left(\frac{R_r - 1}{max_R - 1} * 0.99\right)$$

where max_R is the max number of hits reported by STAR (default = 10). Thus, a fragment that has been mapped several times will have decreasing weight per alignment. For instance, a Ribo-Tis read that has 3 alignments in the genomes would have for R_1 a weight equal to 1, R_2 a weight equal to 0.89 and for R_3 a weight equal to 0.78.

The combination of these three criteria allowed us to weight reads mapped to the genome for the identification of the start codons, using the following probability model:

$$\mathrm{P}(c|\, Ribo-Tis\ reads\ mapped\ to\ x) = \frac{\sum_{r\ read}^{Ribo-Tis} P(sc \mid len\ =\ l,\ pos\ =\ p\ ) \cdot H_1(l) \cdot H_2(R_r)}{\sum_{r\ read}^{Ribo-Tis} H_1(l) \cdot H_2(R_r)}$$

where x is the genomic position of the first nucleotide of a candidate start codon and c is the event that indicates that the position x is a start codon sc.

Finally, to establish a threshold on $P(pos|c)$ to retain only the start codons candidates with high confidence, we ranked the computed confidence results to plot a receiver operating characteristic curve (ROC curve). This curve was plotted using the known start codons as positives and any other start codon candidates as negatives. For each point on the curve, we computed the Euclidean distance to a perfect classifier (0,1) and then reported the threshold corresponding to the shortest distance to that point. Thus, any start codon candidate whose computed confidence was above the threshold was considered as a positive start codon position and was retained for further analysis.

B) Assembly of reads into transcripts: to capture the complete transcriptome including both annotated and unannotated transcripts, we generated sample-specific transcriptomes assemblies from Ribo-seq elongation data collected from actively translating cells and RNA-Seq data. To this end, we used StringTie v1.3.6[79] guided by a reference annotation (Ensembl release 88) in RNA-seq and Ribosome Profiling Elongation BAM files.

C) Intersect: to detect the set of actively translated ORFs, we use the intersection function of the BEDTools[80] suite in the BED file with the genomic positions of the positives start codons as well as each of the gtf files reported by StringTie either transcriptome assemblies based on Ribosome profiling Elongation and RNA-seq. Therefore, start codons intersecting assembled transcripts (i.e.,

pairs (starts codons, transcripts)) were collected as they represent the active ORFs that will be translated in-silico. From this set of ORFs, we define canonical proteins as those translated from an annotated start codon coupled to the corresponding transcript (known couplings) according to genome version GRCh38.p10 (GENCODE version 26). We define non-canonical translation products as those originating from unknown couplings.

D) SNPs integration: to generate sample-specific transcription information, we integrated high-quality single-nucleotide polymorphisms (SNPs) identified from RNA-seq data to the assembled transcripts. Single-nucleotide variants 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. The high-quality sample-specific SNPs identified (freeBayes quality > 20), were then inserted at their correct position into the intersected transcripts. When there was ambiguity for a given position, the integration was done through the corresponded IUPAC symbol.

E) In-silico translation: to generate a sample-specific database, each transcript (from RNA-seq or Ribosome Profile Elongation) was translated from the frame dictated by the coupled start codon until the first in-frame stop codon. Any protein sequence longer or equal to 8 AA was retained. Any protein sequence nested in a larger sequence was not added to the database. However, we keep track of all information about proteins (i.e., which proteins were added to the database and which were not), as we use it to assign the most likely origin of each peptide. To avoid combinatorial explosion, we translated the transcripts containing the IUPAC symbols, the complete protein sequence once, and translated small sequences around the locations of the IUPAC symbols (20 ntd in the flanking regions of the SNPs).

We used Ribo-seq data from translation initiation site (TIS), elongation and RNA-seq data from three human diffuse large B-cell lymphomas (DLBCL), HBL-1, DoHH2 and SU-DHL-4 to generate sample-specific databases using the Ribo-db approach. These databases were used to perform mass spectrometry analysis of the immunopeptidome and the whole proteome. The number of proteins identified in these analyses is shown in Table S2. The percentage of proteins detected by MS among the proteins identified by Ribo-seq is shown in Table S3.

1.17.6 Immunopeptidome sample preparation

Cells for immunopeptidome analysis were grown and harvested the same way and in parallel with ribosome profiling cell cultures. The cells were counted during the washing step with ice-cold DPBS and aliquots of 200 million cells were centrifuged and pellets flash-frozen in liquid nitrogen, stored at -80°C.

1.17.7 Mass spectrometry analysis: immunoprecipitation and sequencing by LC-MS/MS

For MHC-I peptides isolation, we performed immunoprecipitation on 2 replicates per cell line using W6/32 antibody (BioXCell, 1mg per 10^8 cells) as previously described[81]. Replicates were composed of $2x10^8$ cells for HBL-1 and $4x10^8$ cells for SU-DHL-4 and DoHH2. Dried peptide extracts were resuspended in 4% formic acid and loaded on a homemade C18 analytical column (15 cm x 150 μm i.d. packed with C18 Jupiter Phenomenex of particle size 5 μmn and pore size 300 Å) with a 56-min gradient (DoHH2 and SU-DHL-4) or 106-minute gradient (HBL-1) from 0% to 30% ACN (0.2% formic acid) and a 600 nL/min flow rate on an nEasyLC II system. Samples were analyzed with a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) in positive ion mode with Nanospray 2 source at 1.6 kV. 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 30,000, an automatic gain control target of $2x10^4$, an injection time of 800 ms and collisional energy of 25%.

1.17.8 MAP identification

MAPs were eluted from three DLBCL cell lines and analyzed by liquid chromatography-MS/MS (LC-MS/MS). MS/MS spectra were searched against sample-specific customized databases using Peaks X (Bioinformatics Solution Inc.). 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 variable post-translational modifications.

Following peptide identification, a list of unique peptides was obtained for each sample. Binding affinities to the sample's HLA alleles were predicted with NetMHCpan 4.0[82] and only peptides

with length between 8 and 11 amino-acid and with a NetMHC percentile rank ≤ 2% were retained for further annotation. Finally, a false discovery rate (FDR) of 1% was applied on the remaining peptide scores, corresponding to sample-specific FDRs in the range of 1.4 to 2,9% if applied on total PSMs (DoHH2=1.6%, SU-DHL-4=2.9%, HBL-1=1.4%). These filtering steps were made with the use of MAPDP[30]. For each identified peptide, we interrogated all protein sequences to identify those that could be at the source of the peptide. We sequentially applied the following rules to assign to the peptide the most likely source protein based on (i) the highest starting codon confidence score, (ii) the presence of an optimal (GCC[R]CCstartG[V]) or strong ([R]NNstartG[V]) kozak motif[38] around the start codon, (iii) the level of expression of the source transcript through the StringTie computed TPM measurements.

Comet v2019.01.5, a different MS search engine, was used to perform PEAKs re-identification. The raw files were converted to mzXML format with the MsConvert tool of ProteoWizard and searched against the relevant sample-specific customized databases. Comet was used with the same parameters as for PEAKS. Following peptide identification, a list of unique peptides was obtained for each sample and a false discovery rate (FDR) of 1% was applied on the peptide scores. All canonical and non-canonical MAPs identified by PEAKS for each sample were queried in such peptide list and only perfect matches were considered as successful re-identifications.

To ensure that our cryptic peptides did not correspond to improperly assigned post-translationally modified canonical peptides, PEAKS searches were performed using the standard reference protein database (Ensembl GRCh38.88 annotations), including six most frequent PTMs reported for HLA class-I associated peptides[83, 84]. In addition to oxidation (M) and deamidation (NQ), we searched for peptides bearing either phosphorylation, cysteinylation, N-cyclisation (pyroQ) or N-terminal acetylation. Out of the 243 spectrum IDs assigned to cryptic peptides in our study, only 4 were re-assigned to a canonical sequence harboring a PTM, indicating that post-translational modifications might be a confounding factor for at most 1.6% of cryptic peptides.

1.17.9 Retention time prediction and relative mass error

As validation criteria of the MAPs identification robustness, we assessed the Pearson's correlation between the retention time observed and the predicted retention time for each MAPs category

(canonical and non-canonical). Peptide retention times were predicted using DeepLC 0.1.14[40], with default parameters. The model was calibrated using retention time of 250 peptides (top 10 PEAKS scoring peptides from 25 equal-sized retention time bins). In addition, we evaluated the relative mass error for each MAP and compared the distributions for the two MAPs category (canonical and non-canonical). Peptide relative mass error is presented in parts per million mass errors (ppm) unit and was assessed through the MAPDP platform[30].

1.17.10 Composite DB: Ribo-db + PRICE

To validate the relevance of the Ribo-db approach, we ran the PRICE v.1.0.3 method[8] on the BAM files containing mapped reads of ribosome Profiling TIS and Elongation of each cell line with default parameters besides the -novelTranscripts parameter. The predicted ORFs were translated following the same rules as for Ribo-db (i.e., SNP integration and in-silico translation) and were added to the sample-specific Ribo-db database. Next, for each cell line, MS/MS spectra were searched against each sample composite database. The lists of unique identifications obtained from PEAKS were filtered based on 1) length between 8 and 11 amino acids, 2) percentile rank ≤ 2% for at least one on the relevant MHC-I molecules as predicted by NetMHCpan 4.0, 3) FDR ≤ 1% estimation. Each sample-specific database (i.e., Ribo-db and PRICE-db) was independently queried for each peptide identified to count the number of unique and shared peptides found in the databases.

1.17.11 Biotype screening

Non-canonical proteins were desginated as a function of their transcript genomic location: 5’ or 3’UTR proteins are in 5’/3’UTR or overlapping CDS and 5’/3’UTR; frameshift proteins are in coding transcripts but out-of-frame of canonical translations; intronic proteins are in intronic regions or in exon-intron junction; annotated noncoding transcripts proteins are in transcripts annotated as pseudogenes, noncoding RNA and processed transcripts; intergenic proteins are in novel transcripts. We set out to determine the category associated to each non-canonical protein through two validation steps. First, as we used StringTie in the reference-guided manner, we used the reference_transcript (field returned by StringTie) of the transcript from which the protein originated. Therefore, if the non-canonical protein was derived from a protein-coding transcript,

depending on the location of the protein within the transcript relative to the canonical protein, the non-canonical protein was assigned to the categories: '5'UTR' and '3'UTR', 'Novel Isoform' (proteins that share the same reading frame of the canonical protein but originate from an alternative starting codon), 'Frameshift' (proteins in a different reading frame than the canonical one), or 'Intronic' (proteins derived from transcripts containing intronic regions of a canonical protein). If the non-canonical protein derived from an annotated noncoding transcript, then it was directly assigned to the category of 'annotated noncoding transcript'. Finally, if the non-canonical protein whose genomic location was not part of any annotated transcript was assigned to the Intergenic category. The second step was designed to find a consensus category for each protein. Since we knew that some assembled transcripts were not associated with a reference_transcript, we chose to interrogate the annotations (Ensembl gtf file) to find all possible categories associated with the location of the protein. Therefore, for each protein we had the possible categories that could be associated with the protein, and we assigned to the protein the category that was most represented.

1.17.12 Translation efficiency and ribosome occupancy

Translational efficiency of each MAPs source protein was calculated as the ratio between translation (derived from counts of ribosome profiling reads) over transcription (derived from RNA-Seq reads). These measurements were computed as described by Ingolia[46]. First, the ribosome occupancy (translation level) was computed as the number of Ribosome Profiling Elongation fragments aligned to the coding protein sequence divided by the length of the protein sequence. Second, such measurements were normalized by dividing by the total number of Ribosome Profiling Elongation fragments that aligned to any coding transcript sequence. Finally, as the same measurements were computed for RNA-seq reads, the translation efficiency of a gene was computed as the ratio of the normalized Ribosome Profiling Elongation to the normalized RNA-seq.

1.17.13 Translation efficiency analysis of canonical protein location

We extracted the information of the canonical protein subcellular compartments from ComPPI db[85] which provides confidence scores (0-1) for protein subcellular localizations. For each

canonical MAP source protein, we took into account all its major locations, i.e., those having confidence scores above 0.8. For proteins without a localization score above this threshold, we used the subcellular compartment with the highest score. Moreover, for the few proteins underrepresented in the database, we assessed manually their more likely localization in Uniprot. Therefore, we compared the translation efficiency of MAP source proteins from 6 subcellular localizations: cytosol, membrane, nucleus, extracellular, mitochondrion, secretory pathway, and we used as a negative control, the translation efficiency of the canonical proteins non-source of MAPs (background).

1.17.14 Stalling ribosomes

We examined the ribosome profiling elongation coverage of each transcript source of MAPs proteins and compared it to the coverage of transcripts non-source of MAPs (we chose proteins uniquely detected in the whole proteome as non-source of MAPs). We defined the upstream and downstream of a transcript as the first and last 33% of the length of the transcript (33% up and the last 33% down). The 34% of the length of the transcript at the middle was defined as Midstream. We assessed for each nucleotide of the transcript the number of ribosome profiling elongation reads and then computed for each whole transcript its median coverage. Next, we computed the median at the Upstream, Midstream and Downstream sections of the transcript in order to compute the fold change relative to the median coverage of the whole transcript.

1.17.15 Whole proteome analysis

Protein pellets were resuspended in 50 mM ammonium bicarbonate and separated with an Amicon Ultra-15 10K centrifugal filter device. Proteins staying on the filter were resuspended in 50mM ammonium bicarbonate. 10 mM TCEP [Tris(2-carboxyethyl) phosphine hydrochloride Thermo Fisher Scientific] was added to the samples and samples were vortexed for 1 h at 37°C. Chloroacetamide (Sigma-Aldrich) was added for alkylation to a final concentration of 55 mM. Samples were vortexed for another hour at 37°C. One microgram of trypsin was added, and digestion was performed for 8 h at 37°C. Samples were dried and solubilized in 5% ACN-0.2% formic acid (FA). Peptides were separated on a home-made reversed-phase column (150-µm i.d. by 200 mm) with a 216-min gradient from 10 to 30% ACN-0.2% FA and a 600-nl/min flow rate on

an Easy nLC-1000 connected to an Orbitrap Fusion (Thermo Fisher Scientific, San Jose, CA). Each full MS spectrum acquired at a resolution of 120,000 was followed by tandem-MS (MS-MS) spectra acquisition on the most abundant multiply charged precursor ions for a maximum of 3s. Tandem-MS experiments were performed using collision-induced dissociation (CID) at a collision energy of 30%. The data were processed using PEAKS X (Bioinformatics Solutions, Waterloo, ON) and the sample-specific databases and a false discovery rate (FDR) of 1% was applied on the peptide scores. Mass tolerances on precursor and fragment ions were 10 ppm and 0.3 Da, respectively. Fixed modification was carbamidomethyl (C). Selected variable posttranslational modifications were oxidation (M), deamidation (NQ), phosphorylation (STY), acetylation (N-ter). To assign protein origin of tryptic peptides we used the same rules as for immunopeptidomics experience. For each identified peptide, we interrogated all protein sequences to identify those that could be at the source of the tryptic peptide. We sequentially applied the following rules to assign to the peptide the most likely protein origin based on (i) the highest starting codon confidence score, (ii) the presence of an optimal (GCC[R]CCstartG[V]) or strong ([R]NNstartG[V]) kozak motif[38] around the start codon, (iii) the level of expression of the source transcript through the StringTie computed TPM measurements.

1.17.16 Theoretical trypsin digestion, UB sites, disordered regions and instability index prediction

For each cryptic and canonical protein, we counted the theoretical number of tryptic peptides generated after *in-silico* digestion and preserved those with a length ranging between 7 and 25 aa. Degradation signals were predicted based on i) GPS-ARM version 1.0[57] to predict D -box and KEN -box motifs with high confidence (target sequence of anaphase-promoting complex), ii) UbPred[58] to predict canonical ubiquitination sites with high confidence (≥0.84). To identify Intrinsically Disordered Protein Regions (IDPRs), we used the biophysics-based approach IUPred2[59], with a disorder value cut-off set at 0.5. The instability index of each protein, which predicts protein stability based on the order and frequency of certain dipeptides, was computed using the function ProteinAnalysis from the module ProtParam of the Biopython module SeqUtils. This function implements the method described by Guruprasad *et al.*[60]. Proteins with instability indexes over 40 are predicted to be less stable.

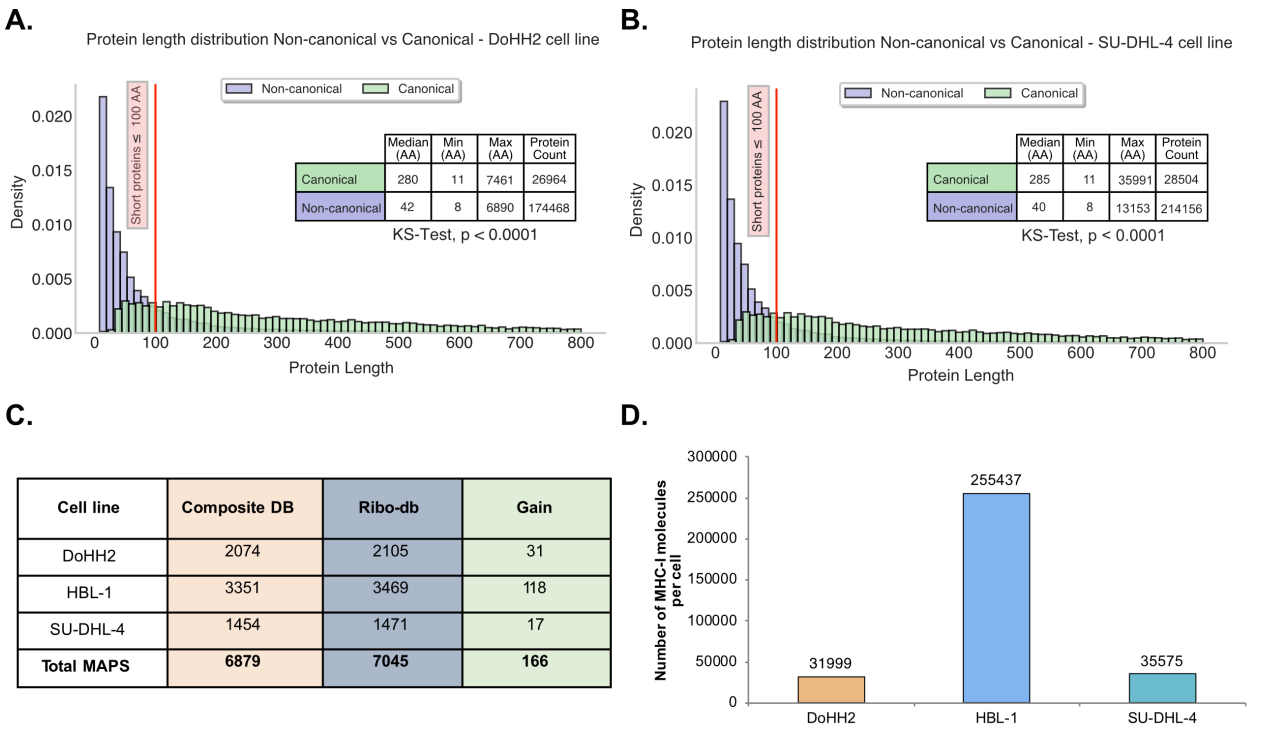
1.17.17 Reactome pathway overrepresentation test

Genes corresponding to annotated canonical proteins encoded within novel isoforms sequences or downstream of 5'UTR-initiated cryptic proteins were submitted to Panther's "Statistical overrepresentation test" (http://www.pantherdb.org/) using reactome pathways as the annotation set. The whole list of homo sapiens genes was used as a reference. Statistical significance of the enrichment of each pathway was assessed using Fisher's exact test, with the Benjamini–Hochberg false discovery rate (FDR) correction for multiple comparisons (adjusted p-value< 0.05). To limit the number of pathways displayed in graphs, we applied, in addition to p< 0.05, a threshold on the level of enrichment of each pathway. Therefore, for novel isoforms, statistically overrepresented pathways with enrichment > 4 were displayed. For 5'UTR cryptic proteins, statistically overrepresented pathways with enrichment > 3 were displayed.

1.17.18 Quantification and Statistical Analysis

Analyses and figures were performed using Python v2.7.6 or R v3.5.1. Correlation test was done using the python function scipy.stats.linregress. All statistical tests used are mentioned in the respective figure legends. Significant level (*p< 0.05, **p< 0.01, ***p< 0.001 and ****p< 0.0001) are reported in the figures. Kolmogorov-Smirnov, Fisher's exact, Mann-Whitney U, T-Test, Wilcoxon rank-sum tests were performed using ks_2samp, fisher_exact, mannwhitneyu, ttest_ind, wilcoxon functions from scipy.stats python module, respectively. Unless mentioned otherwise, all boxes in box plots show the third (75th) and first quartiles (25th) and the box band show the median (second quartile) of the distribution; whiskers extend to 1.5 times the interquartile distance from the box.

1.18 Supplemental Information



Cell line	Composite DB	Ribo-db	Gain
DoHH2	2074	2105	31
HBL-1	3351	3469	118
SU-DHL-4	1454	1471	17
Total MAPS	**6879**	**7045**	**166**

Figure 7. – Related to Figure 5. Sample-specific database composition.

(A) Length distribution of canonical and non-canonical proteins from DoHH2 database showed significant differences. P< 0.0001, Kolmogorov-Smirnov Test. Total proteins for both categories are indicated on the legend, besides their median, minimum (Min) and maximum (Max) observed lengths. Proteins with a length >800 AA are not displayed on the graph.

(B) Length distribution of canonical and non-canonical proteins from SU-DHL-4 database showed significant differences. P< 0.0001, Kolmogorov-Smirnov Test. Total proteins for both categories

are indicated on the table legend, besides their median, minimum (Min) and maximum (Max) observed lengths. Proteins with a length >800 AA are not displayed on the graph.
(C) MAPs identified through composite databases Ribo-db + PRICE method vs MAPs identified solely on Ribo-db-derived databases. MS-Peaks database searches performed solely on Ribo-db allowed to gain (2%) more MAPs dentifications.
(D) Absolute number of MHC-I molecules per cell, in 3 cell lines, measured by flow cytometry using QIFIKIT (see Methods).
(E) MAPs identification process. Diagram that details, from the intersect step of the general workflow overview (Figure 1A), the strategy used for database generation and the filtering for the MAPs identification. The size (Mb) and the total number of amino acids (AA) of each database are shown at the bottom of the figure, along with the FDR used and the number of expected erroneous identifications (False Positive, FP).

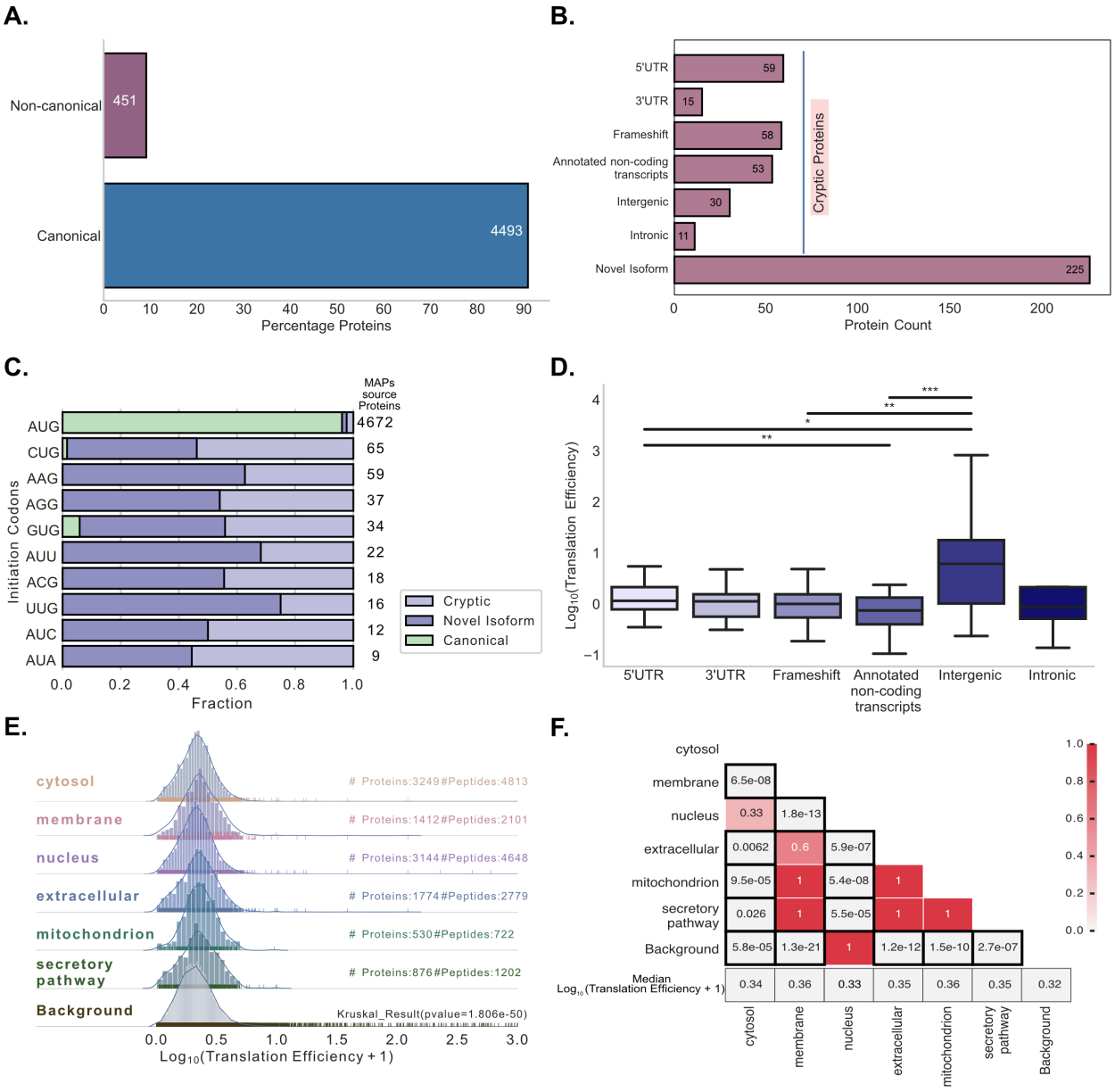


Figure 8. – Related to Figures 6,7. Properties of the novel proteins identified in the immunopeptidome analysis.

(A) At least 10% of the MAPs source proteins derived from non-canonical proteins. Bar plot depicting the percentage of proteins source of MAPs. The purple bar shows the percentage of cryptic and novel isoform proteins (10%), the blue bar shows the percentage of canonical proteins (90%).

(B) Protein count of the MAPs cryptic and novel isoform proteins. The bar plot depicts the total number for each category of the Cryptic along with the total number of Novel Isoform proteins.

(C) Most of the new proteins initiated at near cognate codons. Bar plots showing the fraction of cryptic, novel isoform and canonical proteins initiated through each initiation codon.

(D) Translation efficiency of the MAPs source cryptic proteins. Boxplots showing the translation efficiency distribution for each one of the categories into the MAPs source cryptic proteins. Translational efficiency of each MAPs source protein was calculated as the ratio of translation (derived from counts of ribosome profiling reads) to transcription (derived from RNA-seq reads). *P< 0.05, **P< 0.01, ***P< 0.001, ****P< 0.0001, two-side Mann-Whitney U Test corrected with Bonferroni correction.
(E) Translation efficiency distributions of MAPs source proteins according to their subcellular localization. Background proteins are canonical proteins non-source of MAPs. Number of proteins and number of peptides are presented for each localization. Statistical difference was assessed by Kruskal-Wallis.
(F) Heatmap presenting the adjusted p-values according to the post-hoc comparison for the translation efficiency of the canonical MAPs source proteins. Statistical differences were assessed by Mann–Whitney U tests with Bonferroni correction.

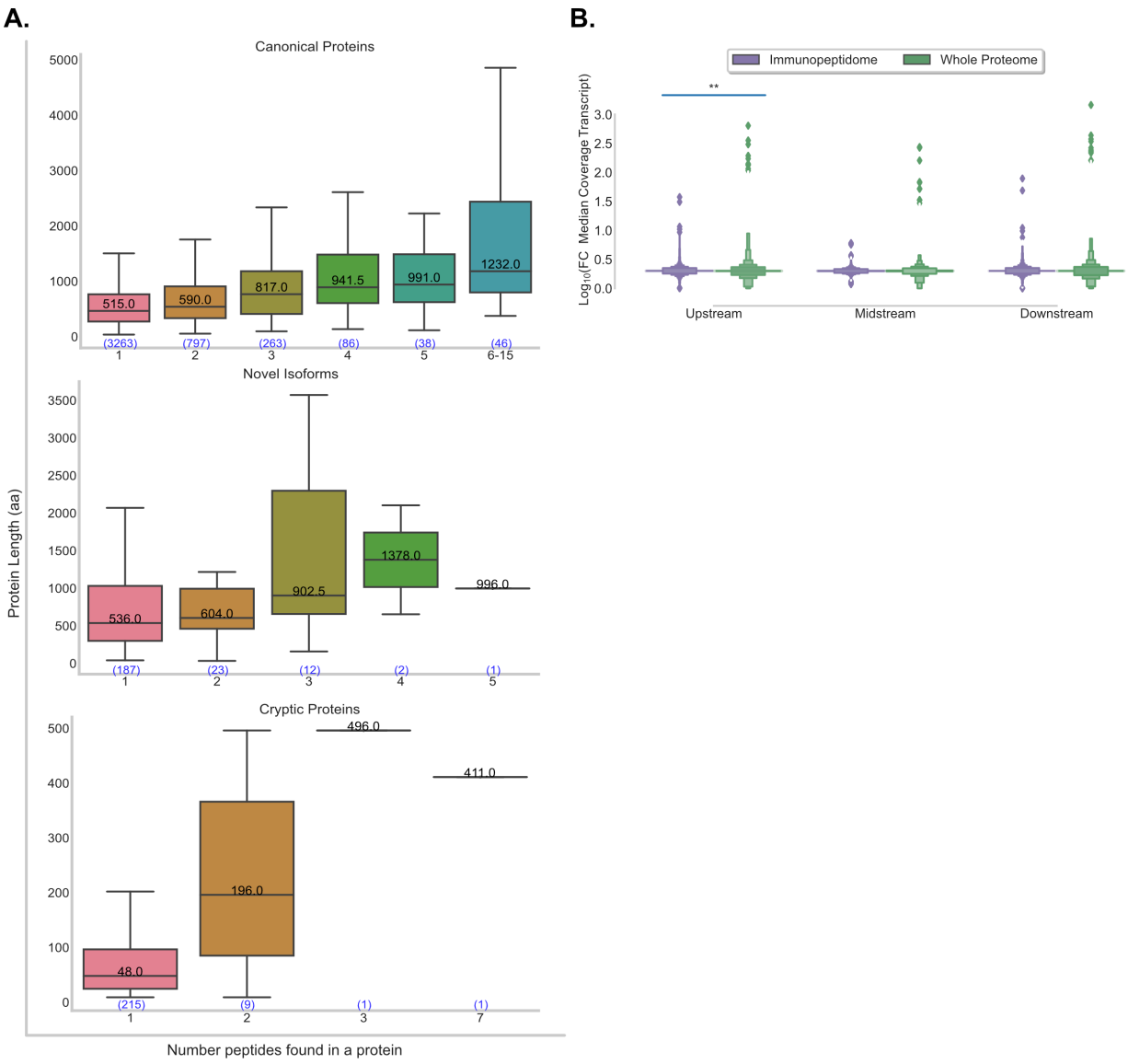


Figure 9. – Related to Figures 7 and 9. Properties of the novel proteins identified in the immunopeptidome analysis.

(A) Boxplot graphs for canonical proteins, novel isoforms and cryptic proteins, showing the number of identified MAPs vs. length of the source protein. The median length of the proteins is shown into each boxplot. The number of proteins that have the number of peptides specified on the X-axis is shown at the bottom of the box chart (blue numbers).

(B) Boxplots showing the fold change of the median coverage for Up-Mid-Downstream relative to the median coverage of the whole transcript, for the MAPs source proteins vs the non-source proteins detected in the whole proteome analysis. The resulting distributions are plotted in log10 of the fold change. Statistical difference was assessed by Wilcoxon signed-rank test.

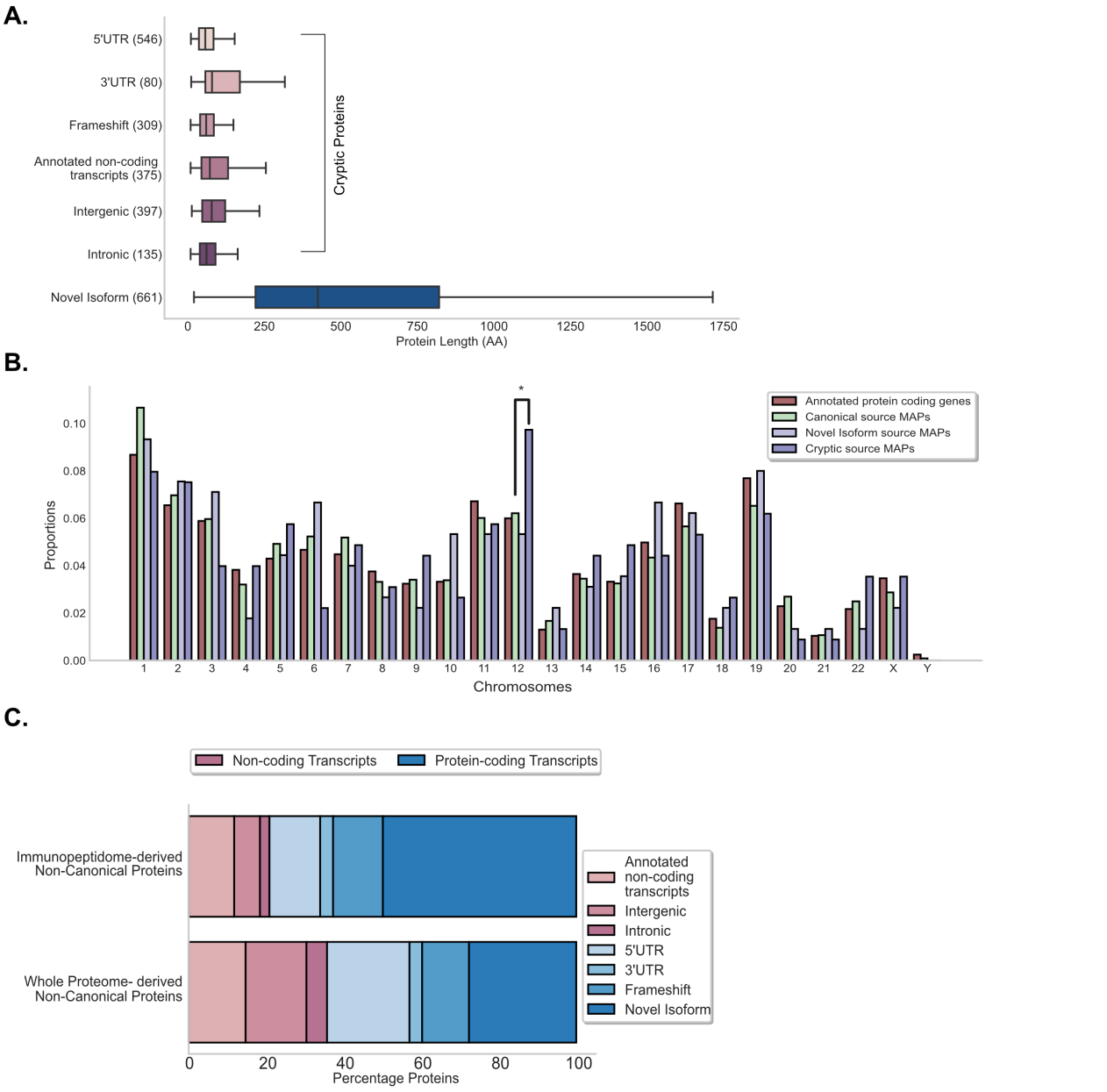


Figure 10. – Related to Figure 10. Features of the newly elucidated proteins.

(A) Cryptic proteins are significantly shorter than novel isoform proteins. Boxplots indicating the length distribution of the newly identified proteins for each category: cryptic proteins (5'UTR, 3'UTR, frameshift, annotated noncoding transcripts, intergenic, intronic) and novel isoforms. Median length in novel isoform (448 aa) and cryptic proteins (65 aa) differed significantly according to two-side Mann-Whitney U test, ****P< 0.0001.

(B) MAPs source newly identified proteins derive from all chromosomes. Bar graph showing, in proportion, the chromosomal origin of each category of proteins, compared to canonical protein coding genes. Chromosome 12 appeared to be rich in cryptic proteins, *P< 0.05, two-side Fisher's exact test.

(C) MAPs source novel proteins derived preferentially from protein-coding transcripts (79% protein-coding vs 21% noncoding transcripts) compared to the percentages on the whole non-canonical proteome (whole proteome analysis-derived proteins). Stacked bar plot showing the percentage of novel identified proteins deriving either noncoding (red bars) and protein-coding transcripts (blue bars) for MAPs non-canonical source proteins vs whole non-canonical proteome.

	Database size (Mb)		
	PRICE	**Ribo-db**	**Composite db Ribo-db +PRICE**
DoHH2	*10.4*	*31.5*	*41.9*
SUD-HL-4	*10.5*	*35.2*	*45.7*
HBL-1	*11.8*	*31.4*	*43.2*

Table 1. – Related to Figure 6C. Size of protein databases used in this study.

			MS-Identified Proteins		
		DB	**Immunopeptidome**	**Whole Proteome***	**Total**
DoHH2	*Canonical*	26964	1366	2478	4612
	Non-canonical	174468	141	627	
SUD-HL-4	*Canonical*	28504	1017	2798	4712
	Non-canonical	214156	95	802	
HBL-1	*Canonical*	26726	2110	2226	5174
	Non-canonical	181811	215	623	
		Total	4944	9554	14498

*(without overlapped proteins in Immunopeptidome)

Table 2. – Related to Methods: Ribo-db approach: detection of active translation sequences. Number of MS identified proteins in the 3 cell lines.

	% (MS / total db)	
	Immunopeptidome	**Proteome**
Canonical	5.51	9.11
Non-canonical	0.08	0.36

Table 3. – Related to Methods: Ribo-db approach: detection of active translation sequences. Percentages of MS-detected proteins among Ribo-seq identified proteins.

Chapter 3 – BamQuery: a proteogenomic tool for the genome-wide exploration of the immunopeptidome

Maria Virginia Ruiz Cuevas[1,2], Marie-Pierre Hardy[1], Jean-David Larouche[1,3], Anca Apavaloaei[1,3], Eralda Kina[1,3], Krystel Vincent[1], Patrick Gendron[1], Jean-Philippe Laverdure[1], Chantal Durette[1], Pierre Thibault[1,4,6], Sébastien Lemieux[1,2,6], Claude Perreault[1,3,6,7] and Grégory Ehx[1,5,6,7]

[1] Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[2] Department of Biochemistry and Molecular Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[3] Department of Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada.

[4] Department of Chemistry, Université de Montréal, Montreal, QC H3C 3J7, Canada

[5] Laboratory of Hematology, GIGA-I3, University of Liege and CHU of Liege, Liege, Belgium

[6] Senior authors

[7] Lead Contact

***Correspondence**: g.ehx@uliege.be (GE)

1.19Context

MHC-I-associated peptides (MAPs) are small molecules that are presented on the surface of cells. MAPs play a key role in the immune system's recognition and response to abnormal substances, such as those derived from viruses or cancer cells. MAPs can come from a variety of genomic regions, including both protein-coding and non-protein-coding sequences, such as endogenous retroelements (EREs). Identifying and quantifying the expression of MAPs in both healthy and cancerous cells is important for understanding the immune response to tumors and for identifying potential tumor antigens (TAs) that could be targeted by immune therapies. However, this can be a challenge for immunologists.

To address this challenge, we have developed a computational tool called BamQuery that allows for the comprehensive mapping of MAPs to trace their expression in healthy and cancer tissues. BamQuery can analyze bulk and single-cell RNA-sequencing data to attribute expression to MAPs of any origin, including exons, introns, untranslated regions, and intergenic regions. We show that non-canonical MAPs, including TAs, can come from multiple different genomic regions, and that they can be abundant in normal tissues. Also, we demonstrate that supposedly tumor-specific mutated MAPs, viral MAPs, and MAPs derived from proteasomal splicing can arise from different unmutated non-canonical genomic regions.

Overall, the genome-wide approach of BamQuery allows for a more complete understanding of MAP expression in both healthy and cancerous tissues and help to predict MAP immunogenicity and identify potential TAs for immune therapies.

1.20Authors' contributions

Maria Virginia Ruiz Cuevas: coded the final version of BamQuery, designed the study, performed the analyses, interpreted the data, and wrote the article.

Marie-Pierre Hardy: helped with bioinformatic analyses and data interpretation.

Jean-David Larouche: performed the single-cell analyses.

Anca Apavaloaei: helped with bioinformatic analyses and data interpretation.

Eralda Kina: helped with bioinformatic analyses and data interpretation.

Krystel Vincent: helped with bioinformatic analyses and data interpretation.

Patrick Gendron: deployed the web portal and prepared BamQuery docker.

Jean-Philippe Laverdure: generated DLBCL cell lines MS databases.

Chantal Durette: performed MS searches for DLBCL samples.

Pierre Thibault: designed the study, interpreted the data, and wrote the manuscript.

Sébastien Lemieux: designed the study, interpreted the data, and wrote the manuscript.

Claude Perreault: designed the study, interpreted the data, and wrote the manuscript.

Grégory Ehx: designed the study, interpreted the data, and wrote the manuscript.

All authors edited and approved the manuscript.

1.21 Abstract

MHC-I-associated peptides (MAPs) derive from selective yet highly diverse genomic regions, including allegedly non-protein-coding sequences, such as endogenous retroelements (EREs). Quantifying canonical (exonic) and non-canonical MAPs-encoding RNA expression in malignant and benign cells is critical for identifying tumor antigens (TAs) but represents a challenge for immunologists. We present BamQuery, a computational tool attributing an exhaustive RNA expression to MAPs of any origin (exon, intron, UTR, intergenic) from bulk and single-cell RNA-sequencing data. We show that non-canonical MAPs (including TAs) can derive from multiple different genomic regions (up to 35,343 for EREs), abundantly expressed in normal tissues. We also show that supposedly tumor-specific mutated MAPs, viral MAPs, and MAPs derived from proteasomal splicing can arise from different unmutated non-canonical genomic regions. The genome-wide approach of BamQuery allows comprehensive mapping of all MAPs in healthy and cancer tissues. BamQuery can also help predict MAP immunogenicity and identify safe and actionable TAs.

KEYWORDS: immunopeptidome; computational biology; major histocompatibility complex; tumor antigens

Abbreviations: MAP: MHC-I associated peptide; TA: Tumor antigen; ncMAP: non-canonical MAP; ncRNA: non-coding RNA; ERE: Endogenous retroelement; MCS: MAP coding sequence; ncMCS: non-canonical MCS; RPHM: Read-per-hundred-million; mTEC: Medullary thymic epithelial cell; DC: Dendritic cell; TSA: Tumor-specific antigen; CTA: Cancer-testis antigen; DLBCL: Diffuse large B-cell lymphoma.

1.22Introduction

The immunopeptidome is the repertoire of MHC-I-associated peptides (MAPs) that represents in real-time the landscape of the intracellular proteome as it is molded by protein translation and degradation[1]. In recent years, immunopeptidomic data has been harvested to identify relevant and targetable tumor antigens. Indeed, MAPs deriving from mutations characterizing the neoplastic transformation (mutated tumor antigens (TA), also known as neoantigens) can be recognized by cytotoxic T cells and used as anti-cancer therapeutic targets[2].

The immunopeptidome is typically assumed to result from the degradation of canonical proteins, coded by exons and translated from known open-reading frames. However, recent proteogenomics (proteomic informed by genomics such as RNA sequencing (RNA-seq)) findings evidenced that ~5-10% MAPs can also derive from non-canonical (nc) regions of the genome, such as introns, non-coding RNAs (ncRNA) or endogenous retroelements (EREs), as well as from out-of-frame translation of exons[3-6]. While 99% of somatic mutations are located in non-coding regions[7], the vast majority of the discovered ncMAPs are non-mutated[4, 8-11]. Many ncMAPs are found exclusively in cancer cells and attract attention as (1) they can be immunogenic *in vitro* as well as *in vivo*; (2) they are more numerous in the immunopeptidome of malignant cells than mutated TAs and (3) several non-coding TAs are widely-shared between cancer patients whereas mutations mainly generate private antigens[12, 13]. In the context of proteogenomics usage, ncMAPs discovery and actionable TAs identification have raised three challenges that are often addressed inconsistently by immunologists.

First, the attribution of an exact RNA expression to MAPs. Typically, proteogenomic pipelines quantify MAPs RNA expression through the estimation of their parental transcript expression by using conventional transcript abundance quantification tools. However, such tools cannot be used reliably for ncMAPs which often derive from unannotated genomic regions. Furthermore, such approaches do not consider that MAPs (8-11 residues) could derive from multiple regions of the genome due to the degeneracy of the genetic code. Therefore, studies failing to consider all genomic regions susceptible to generating a given MAP would underestimate its RNA expression.

Second, the attribution of a biotype to MAPs. Due to the multiplicity of genomic regions able to generate the same MAP, and possibly having different biotypes, a MAP could be mislabeled for example as ERE-derived while a canonical region could also generate it through out-of-frame translation.

The third challenge is to prioritize TAs. Ideally, TAs should be immunogenic and specifically expressed (or overexpressed) by malignant cells[14]. Because RNA expression is a reliable proxy of the MAP presentation probability[9, 15], RNA-seq data of tumor and normal samples are powerful tools to perform TA prioritization. While tumor specificity can be evaluated by comparing MAP RNA expression between tumor and normal samples, evaluating MAPs RNA expression in medullary thymic epithelial cells (mTECs) should be a good predictor of immunogenicity because mTEC MAPs induce central immune tolerance[16]. However, for the reasons mentioned above, comparing reliably MAPs RNA expression between tumors, their paired normal samples, and mTECs requires considering all their possible genomic regions of origin.

To address these challenges, we developed BamQuery, an annotation-independent tool that enables the attribution of an exhaustive RNA expression profile to any MAP of interest in any RNA-seq dataset of interest.

1.23Results

1.23.1Exhaustive capture of MAPs RNA expression

Because genomic annotations cover vast regions that are unlikely to represent accurately the local RNA expression of an 8-11 residues peptide (especially for ncMAPs deriving from introns, Extended Data Fig. 1a) and because no annotations are available for MAPs deriving from intergenic regions, we designed BamQuery to evaluate MAPs RNA expression independently of annotations. Due to the small size of MAP-coding sequences (MCS, 24-33 nucleotides), counting the RNA-seq reads containing each MCS able to code for a given peptide is the most thorough and less error-prone method to evaluate MAPs RNA expression. To make BamQuery readily available, it had to work on a broadly used data format. Given that querying MCS in fastq files is time-consuming (> 1 minute / MCS), we designed BamQuery to work on bam files in five steps

(Fig. 1a, and Methods): (1) reverse-translation of each MAP into all possible MCS; (2) mapping of MCS to the genome using STAR[17] to identify those having perfect matches with the reference and attribute them a genomic location. At this step, we also include to the reference genome the mutations from the dbSNP annotations[18] to enable the mapping of mutated sequences; (3) counting of the primary RNA-seq reads encompassing exactly the MCS at their respective location (~0.0005 minute / MCS / location) and sum read counts of each MAP across locations; (4) normalization of the read count of each MAP by the total primary alignment read count of the sample and multiplication by 1×10^8 to yield read-per-hundred-million (RPHM) numbers and (5) attribution of biotypes to MAPs based on the reference annotations overlapping the various expressed (RPHM>0) regions.

To test BamQuery, we collected robustly validated benign MAPs from the HLA Ligand Atlas[19] (1,702 canonical MAPs shared across at least 20 tissues, Extended Data Fig. 1b,c) and queried them in the transcriptome of eight mTEC samples sequenced previously[10, 20]. As a control, we used the primary reads contained in the mTEC bam files previously aligned with STAR to generate a database of 27-nucleotide-long k-mers (reads chunked into shorter sequences) using Jellyfish[21], a tool that counts k-mer occurrences in the primary read sequences (Methods). Importantly, we preferred designing BamQuery to work on bam files instead of Jellyfish k-mer files of original fastq files because of the elevated disk space that k-mer databases require (4 databases would be needed per sample to query MAPs of 8 to 11 amino acid length) and because such databases would not provide information about the genomic region of the queried MCS.

We queried this 27-nucleotide-long k-mer database for all possible 27-mer-MCSs encoding 9-amino acid-long MAPs (1,211/1,702). The comparison of total read counts between BamQuery and total k-mer occurrences for each MAP showed a correlation equal to 1, demonstrating the exhaustivity of BamQuery (Fig. 1b). Importantly, the main outlier in this correlation was the RVHPQVTVY peptide, deriving from the HLA-DRB3 gene. Previously, the STAR aligner was shown to have poor performance in hypervariable genomic regions such as HLA genes[22]. Consequently, this outlier results from the limited capacity of STAR to map MCS to the HLA-DRB3 gene when performing the BamQuery analysis. A more detailed comparison between MCS counts given by BamQuery and k-mer counts in the database also showed an excellent correlation, except for the

MCS coding for the RVHPQVTVY peptide (94% mean accuracy) (Extended Data Fig. 1d-e). Next, we compared BamQuery to Kallisto[23], a transcript abundance quantification tool (reference MAPs RNA quantification method) that was chosen because it provides results similar to other tools while having the fastest computing speed[24]. A poor correlation between Kallisto and BamQuery was found (Fig. 1c) as most Kallisto measurements were skewed toward lower values than BamQuery's. Specifically, Kallisto did not detect expression for 32 MAPs while BamQuery reported considerable RPHM values. In fact, BamQuery revealed that these MAPs are the result of multiple genomic locations (mean = 11) and are completely lost when only a single MAP source transcript is quantified (Extended Data Fig. 1f,g), as is typically done with transcript abundance quantification tools. Overall, these results evidence the accuracy and superiority of BamQuery over conventional approaches.

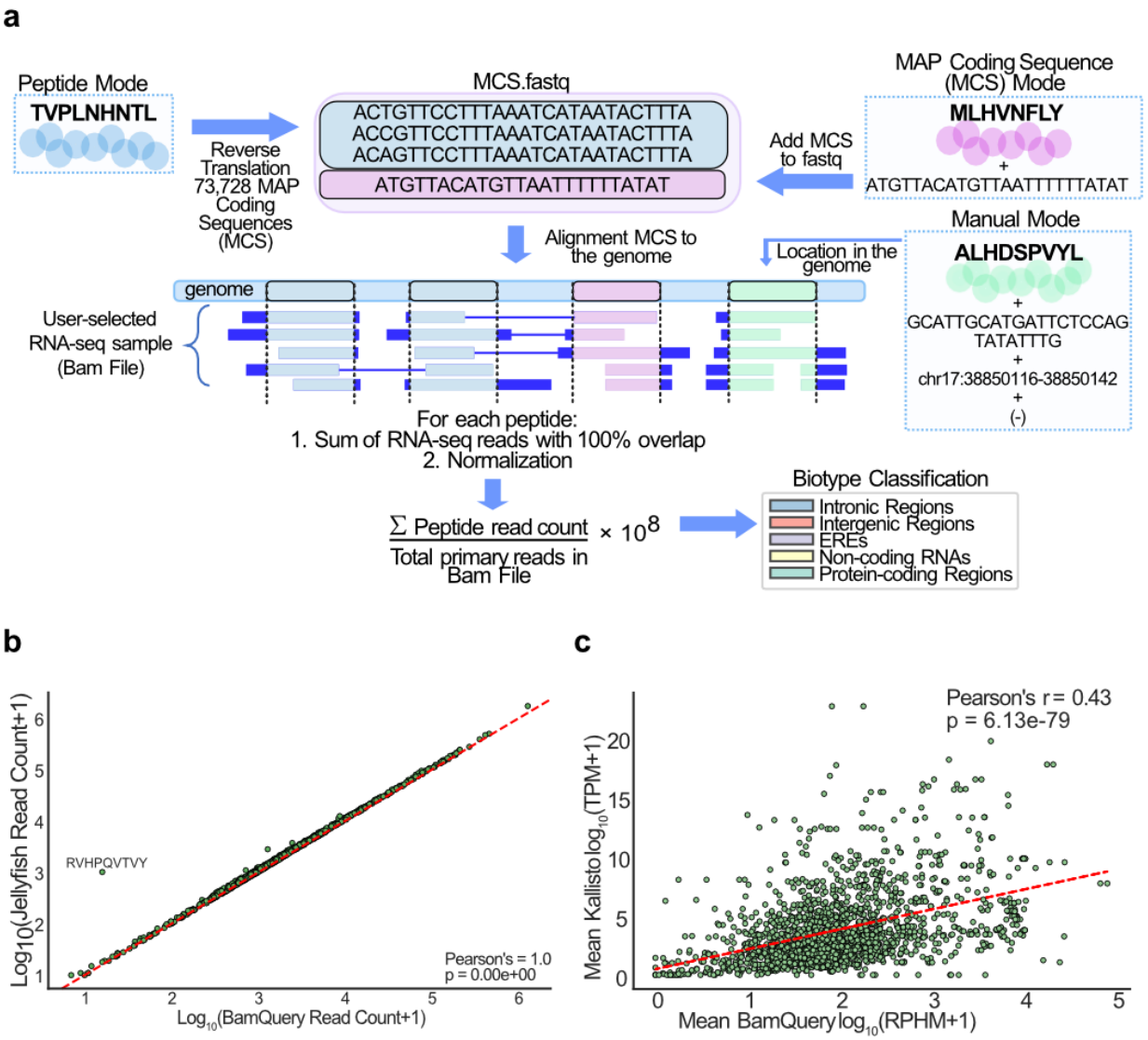


Figure 1. – Exhaustive capture of MAPs RNA expression.

a, Overview of the BamQuery approach to measuring MAPs RNA expression levels.

b, Pearson's correlation between BamQuery-acquired read counts and Jellyfish's K-mer counts for canonical nonamer MAPs (n=1,211) from the HLA Ligand Atlas (present in at least 20 different tissues) in eight mTEC samples.

c, Pearson's correlation between BamQuery RPHM quantification and Kallisto TPM quantification for canonical MAPs (n=1,702) from the HLA Ligand Atlas (present in at least 20 different tissues) and 8 mTEC samples. Red lines in (b) and (c) are linear regressions.

1.23.2New insights into the immunopeptidome biology

Next, we explored the biological features of the immunopeptidome by evaluating the expression of the 1,702 canonical MAPs from the HLA ligand atlas along with 724 MAPs previously reported as non-canonical (EREs, intronic, and ncRNAs-derived, Supplementary Table 1) in normal tissues, including mTEC samples[25], and tissues from GTEx[26] (Supplementary Table 2). BamQuery attributed a genomic location to 100% MAPs: among canonical MAPs, all originally annotated genes were attributed to their respective MAP by BamQuery and among a large list of well-annotated ncMAPs[9], the originally annotated genomic location was re-located by BamQuery with an accuracy of 100%.

Comparing all 9-mers together (to prevent biases due to differences of length proportions), a higher number of possible MCS (total number of MCS after reverse-translation) was found for non-canonical vs canonical MAPs, especially for those mapping to introns and EREs (Fig. 2a). To better understand this bias, we investigated whether this could be linked to the degeneracy of codons. We found that residues encoded by six synonymous codons (R/L/S) were enriched in intron- and ERE-derived MAPs, with leucine being the most enriched (Fig. 2b-c). Previously, we observed that MAP source transcripts use rare codons more frequently than transcripts that do not generate MAPs[4]. Therefore, we hypothesized that ncMAPs would use rare codons more frequently than canonical ones. Indeed, we found that the genomic codon frequency of residues encoded by 6 synonymous codons (R/L/S) was on average lower than those encoded by lower numbers of synonymous codons (Extended Data Fig. 2a) and that the codons of ncMCS presented a lower genomic frequency than canonical ones (Fig. 2d). As rare codons are rate limiting for protein synthesis[27-29] and as MAPs derive frequently from defective ribosomal products (DRiPs) generated by alterations of protein synthesis rate[30], our data suggest that DRiPs contribute more to the generation of ncMAPs than to canonical ones.

Next, we analyzed the relation between the number of possible MCS per MAP (i.e., diversity of synonymous codons) and the number of genomic regions able to code for a given MAP. Canonical MAPs essentially derived from a single genomic location (60%), while non-canonical MAPs could derive from multiple regions (Fig. 2e). ERE MAPs presented the greatest numbers of possible regions, in agreement with their repeated nature (between 1.536 and

2.9×10^6 possible regions). However, their number of possible MCS did not correlate with the number of possible locations, showing that amino acid residue composition cannot be used to predict the number of possible regions of origin (Fig. 2f).

Finally, given the multiplicity of possible regions of origin, we computed the most likely biotype of each MAP. For this, we used machine learning (expectation-maximization algorithm) to rank the biotypes (in-frame, intron, ERE, etc.) as a function of their likelihood of generating the reads covering them across the whole set of GTEx tissues. In general, canonical in-frame transcripts are more likely translated than non-canonical ones. For this reason, BamQuery's best guess automatically ranks as “in-frame” any MAP having at least one in-frame canonical origin, which was the case for all canonical MAPs from our dataset (Extended Data Fig. 2b). BamQuery can also attribute biotypes based only on the likelihood ranks (considering the number of reads overlapping each transcript). In this case, ~26% of canonical MAPs were assigned with a greater probability to ncRNAs (Fig. 2g). Furthermore, while ncRNA and intron MAPs were predicted to belong mainly from their identified biotype (73 and 81%) (Extended Data Fig. 2c), only 56% of ERE-derived MAPs were estimated to derive from EREs, and 6% of them could derive from canonical regions (5% in-frame) (Fig. 2h). Altogether, these data show that many published MAPs could be mislabeled, either as canonical or non-canonical.

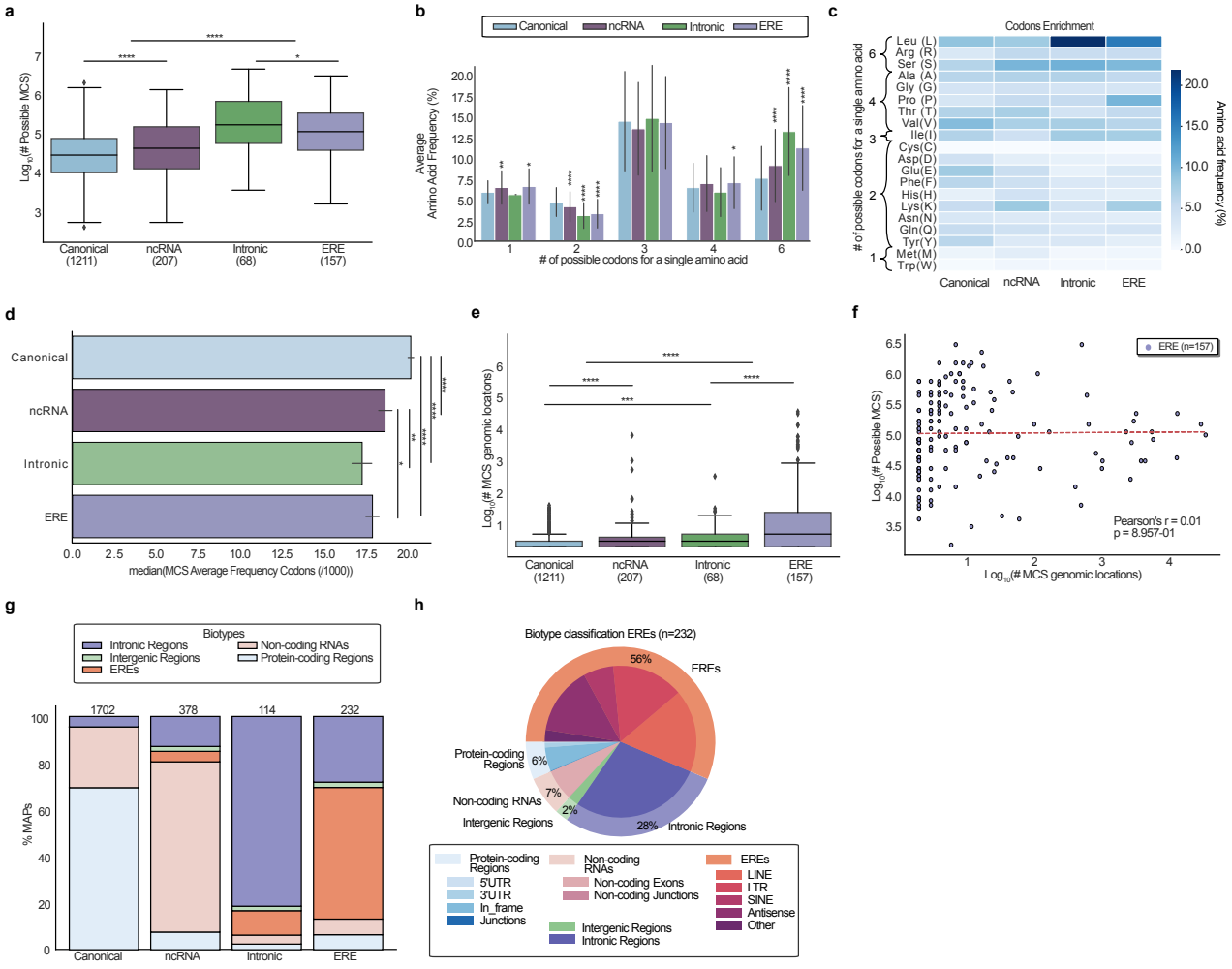


Figure 2. – New insights into the immunopeptidome biology.

a-h Published MAPs reported as canonical (n=1,702) and non-canonical (ncRNA (n=378), intronic (n=114) and EREs (n=232)) were searched with BamQuery in GTEx tissues and mTEC bam files in unstranded mode (GTEx data being unstranded) with genome version GRCh38.p13, gene set annotations release v38_104 and dbSNP release 151. Figures a,e,f,g were generated with the comparison of 9-mers only (n=1,211 canonical, n=207 ncRNA, n=68 intronic, n=157 EREs) to prevent possible biases introduced by variable frequencies of 8/10/11-mers among the compared groups. Figures b,c,h were generated with the complete MAP dataset (n=1,702 canonical, n=378 ncRNA, n=114 intronic, n=232 EREs). Mann-Whitney U test was used for indicated comparisons (*p<0.05, **p<0.01, ***p<0.001, ****p<0.0001).

a, Number of possible MCS after reverse-translation of indicated MAP groups.

b, Average frequency (%) of amino acids encoded by the indicated number of synonymous codons in indicated MAP groups.

c, Heat map of amino acid frequency in indicated MAP groups.

d, Mean of the MCS average usage frequency of codons (among 1000 codons located in human reference protein-coding sequences) encoding each of the 20 amino acids of indicated MAP groups.
e, Number of MCS genomic locations able to code for the indicated MAP groups.
f, Pearson's correlation between the number of possible MCS after reverse translation vs the number of MCS genomic locations able to code for the assessed ERE MAPs. The red line is a linear regression.
g, Percentage of MAPs attributed to indicated biotypes by BamQuery based on the EM-established biotype ranks and on the genomic regions expressed in GTEx tissues and mTECs. The X-axis indicates the biotype reported in the original study (groups). For clarity, BamQuery biotypes were summarized into five general categories: protein-coding regions, non-coding RNAs, EREs, intronic and intergenic.
h, Percentage of the most likely biotype attributed by BamQuery to EREs MAPs.

1.23.3Single-cell proteogenomic analyses

High-throughput single-cell RNA sequencing (scRNA-seq) enables the examination of individual cells' transcriptome[31, 32]. Therefore, we sought to perform single-cell analyses using BamQuery. Given the end-bias of the Chromium library design typically used in scRNA-seq, we evaluated whether read coverage would allow BamQuery analyses of canonical and non-canonical MAPs in cancerous[33] and normal[34] lung tissues scRNA-seq data. As expected, reads showed a bias toward the 3' end of the canonical genes (Extended Data Fig. 3a). However, the coverage extended far from the 3' end, in agreement with a report detecting mutations in various regions of the gene body[35]. We also found a surprisingly high (~50% of reads) and homogeneous read coverage in introns and ERE regions, in agreement with previous reports[36, 37], and suggesting that BamQuery would be able to detect expression for ncMAPs in scRNA-seq.

BamQuery detected expression for 50-60% of the canonical and non-canonical MAPs (Supplementary Table 1) in scRNA-seq, while 86% were found in bulk RNA-seq of GTEx lung samples (Fig. 3a). This lower number of MAPs expressed on single-cell data resulting from lower read coverage did not hamper the feasibility of scRNA-seq analyses. Indeed, canonical MAPs were uniformly expressed at the 5' and 3' ends of their transcripts (Fig. 3b). Also, the expressed rate of intronic and ERE MAPs in scRNA-seq data was more comparable to bulk RNA-seq data than canonical MAPs (Fig. 3c). This likely results from the more homogeneous read coverage observed in non-coding than in coding regions (Extended Data Fig. 3a).

Therefore, we explored the patterns of MAPs expression in normal and malignant lungs. Differential expression analysis showed that 12.86% (186/1446) and 16.46% (248/1506) of MAPs presented cell type-specific expression profiles in normal and malignant samples, respectively (Fig. 3d and Supplementary Tables 3-4). Several differentially expressed MAPs derived from genes having cell type-specific functions such as YTAVVPLVY in B cells (immunoglobulin J polypeptide), STFQQMWISK in muscle cells (Beta-actin-like protein 2), and FLLFPDMEA in macrophages (complement C1q B chain) (Extended Data Fig. 3b). To further assess the reliability of MAP expression, we re-clustered the normal lung dataset based uniquely on MAPs expression. This provided a clear separation of the hematopoietic and stromal compartments (Fig. 3e, Extended Data Fig. 3c) and allowed the clustering of specific cell populations such as alveolar cells or the monocytes and macrophages (Extended Data Fig. 3d,e). Strikingly, most MAPs identified as differentially expressed in the normal lung dataset had an expression restricted to either the hematopoietic or stromal lineages, showing a clear dichotomy between these two compartments in terms of MAP expression (Extended Data Fig. 3f).

Finally, given the growing interest in TAs shared between tumor cells, we assessed the clonality of 45 MAPs whose coding sequences were overexpressed by cancer cells through co-expression analyses. This highlighted two clusters of MAPs co-expressed in lung cancer cells (Fig. 3f) for which a distinct expression profile was observed in the lung (Fig. 3g). Indeed, MAPs of cluster 1 were expressed by a limited number of cancer cells, whereas MAPs of cluster 2 were ubiquitously expressed, making them more desirable immunotherapeutic targets. These data demonstrate the capacity of BamQuery to perform scRNA-seq analyses and evidence its potential to assess TAs intra-tumoral heterogeneity.

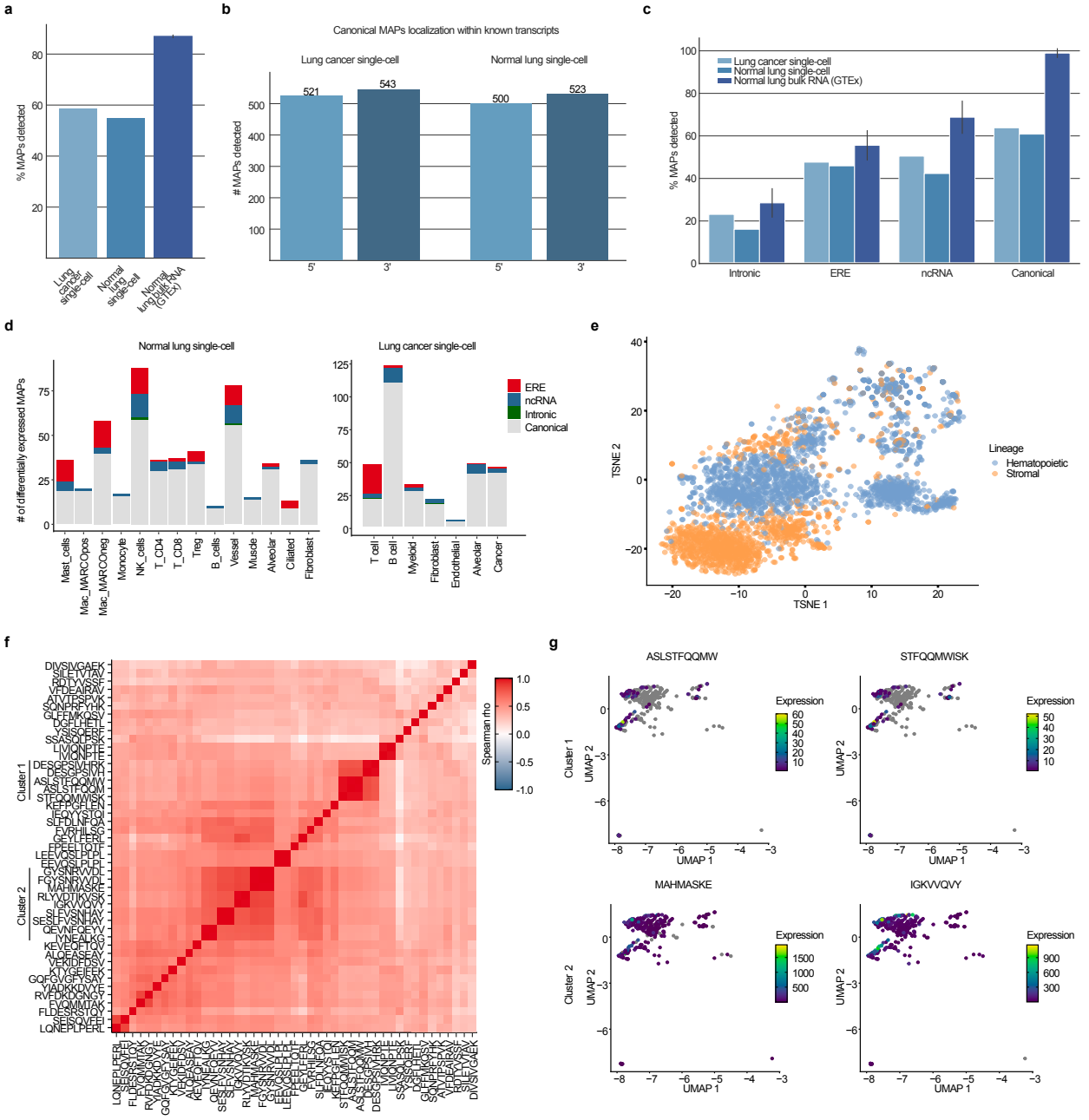


Figure 3. – Single cell proteogenomic analyses.

a-g Canonical (n=1,702) and non-canonical MAPs (ncRNA (378), intronic (114) and EREs (232)) were searched with BamQuery in bam files of scRNA-seq of normal and cancerous lung samples in single-cell in stranded mode with genome version GRCh38.p13, gene set annotations release v38_104 and dbSNP release 151.

a, Median percentage of MAPs detected in normal and cancerous lung scRNA-seq, as well as in bulk RNA-seq samples of normal lungs from GTEx (n=150).

b, Number of canonical MAPs located in the 5' (first half of the transcript) or 3' (second half of the transcript) region of the transcript detected in indicated scRNA-seq datasets.

c, Median percentage of indicated MAP groups detected in normal and cancerous lung scRNA-seq, as well as in bulk RNA-seq samples of normal lungs from GTEx.
d, Number of MAPs identified as differentially expressed by the different populations of cells in the normal lung (left panel) or cancerous lung (right panel). The originally reported biotype of the MAPs is indicated by the color code.
e, TSNE analysis of the hematopoietic (blue) and stromal (orange) cells from the normal lung based on their MAP expression.
f, Heatmap showing the co-expression (spearman rho, color bar) of MAPs overexpressed by lung cancer cells (rows vs columns). Two clusters of MAPs are highlighted on the left side of the heatmap (cluster 1 and cluster 2).
g, TSNE showing the expression of MAPs (color bar) from cluster 1 (higher panel) or from cluster 2 (lower panel). Grey color indicates the null expression of a MAP in a cell.

1.23.4MAP expression is underestimated in healthy tissues

Givn the ability of BamQuery to capture MAPs RNA expression exhaustively, we evaluated the genomic origin of previously reported MAPs. First, we examined 1,062 colorectal cancer (CRC) TAs identified by their presence and absence from the immunopeptidome of malignant and paired benign cells, respectively[38]. To evaluate their probability of being presented by normal cells, we queried them in 3 datasets: GTEx, mTECs, and sorted dendritic cells (DCs)[39, 40] (Supplementary Table 2). Four percent of TAs presented an expression <8.55 RPHM (minimum expression required to result in a probability >5% of generating a MAP[9]) in all normal tissues, except for testis, as these antigens would be classified as cancer-testis antigens (CTAs) (Fig. 4a). Strikingly, among the 7 TAs reported previously as being lowly expressed at RNA level in normal matched tissues, BamQuery revealed that only one (KYLEKYYNL) presented a low expression across all peripheral tissues. Finally, no expression was found for the RYLAVAAVF peptide (the only mutated TA reported in this study), while its wild-type counterpart was highly expressed, making it a promising target for CRC immunotherapies (Fig. 4b).

Second, we wondered whether mutated TAs would be as tumor-specific as expected. We analyzed 45 8-11 amino acid long mutated peptides (7 from gene fusions, 28 from aberrant splice junctions, and 10 from single nucleotide variations, SNV) reported as tumor-specific in medulloblastoma (no RNA expression in GTEx)[41]. BamQuery could attribute a genomic location to 39 of them and mapped 7/10 SNV peptides to their reported genes (Extended Data Fig. 4a).

Unexpectedly, BamQuery attributed non-discontinued ("unspliced") expressed genomic locations to 82% of fusion and spliced peptides, evidencing that non-mutated (and mostly non-canonical, Fig. 4c) genomic regions could also code for those peptides. Overall, only 26 of 45 TAs presented low expression in normal tissues (Extended Data Fig. 4b) including all detected SNV-derived peptides. Therefore, we wondered whether mutated MAPs reported as cancer-specific in previous publications and public databases[11, 42, 43] would be verified as such by BamQuery. From 323 mutated TAs (Supplementary Table 5), 23 (7%) were highly expressed in normal tissues where 25% of the peptides have more than 5 non-mutated genomic locations perfectly matching their MCS (Fig. 4d).

Third, we examined 6 ERE-derived MAPs reported as TAs (lowly expressed in normal tissues, including mTECs, and highly expressed in multiple cancer specimens) in triple-negative breast cancer[44]. While the original study identified an average of 8 locations for these peptides, BamQuery identified ~66 locations per MAP (Fig. 4e). Moreover, these MAPs showed higher expression in normal breast samples compared to cancer samples (Fig. 4f). These results highlight the importance of considering all genomic locations able to generate a given MAP when measuring RNA expression.

Fourth, we evaluated whether BamQuery would detect non-discontinued genomic locations and RNA expression for MAPs supposedly impossible to be expressed by the human genome. We first examined 99 MAPs deriving from proteasomal splicing (generated from post-translational recombination of protein fragments)[45]. Fifteen could be generated by expressed regions (Fig. 4g), suggesting a possible misclassification of these peptides. Finally, considering the tight link between Epstein–Barr virus (EBV) infection and autoimmune disorders such as multiple sclerosis[46], we examined the expression of 511 EBV-derived MAPs in the IEDB database. Four of them could be coded by the human genome and were expressed at high levels by normal tissues (Fig. 4h). Interestingly, one of them, CPLSKILL, can be presented by HLA-B8 molecules, an allele frequently associated with autoimmune disorders[47].

Altogether, these results demonstrate that BamQuery is crucial to attribute an exhaustive RNA expression to MAPs and suggest that it could help select safe-to-target MAPs.

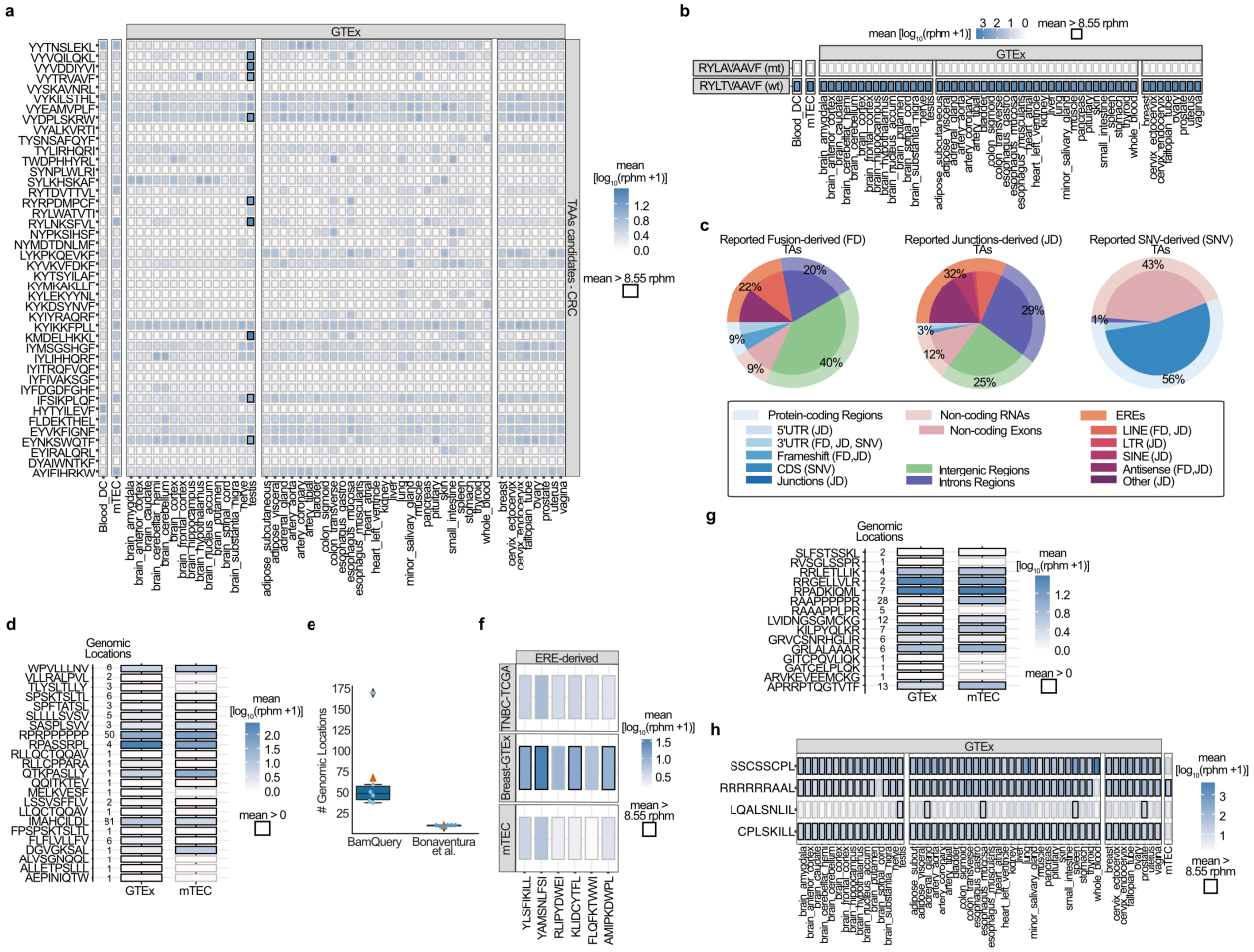


Figure 4. – Underestimated MAP expression in healthy tissues.

a-h Published human colorectal cancer (CRC) TAAs, mutated TAs, ERE-derived TSAs, proteasomal splicing peptides, and Epstein-Barr virus (EBV) MAPs were searched with BamQuery in the GTEx tissues (n=12–50 / tissue), mTECs (n=11) and/or DCs (n=19) bam files in unstranded mode with genome version GRCh38.p13, gene set annotations release v38_104 and dbSNP release 155 (except for the search for mutated TAs (d) where dbSNP was not considered, dbSNP=0).

a, Heatmap of average RNA expression of published CRC TAAs in indicated tissues. Boxes in which a peptide has an average rphm>8.55 are highlighted in black.

b, Heatmap of average RNA expression of the CRC mutated TA RYLAVAAVF and its wild type RYLTVAAVF in indicated tissues.

c, Percentage of the most likely biotype attributed by BamQuery to published fusions, junctions, and SNVs-derived TAs.

d, Heatmap of average RNA expression of published mutated TAs (n=23) in indicated tissues. The number of genomic locations expressed is presented on the left.

e, Number of genomic locations at which the expression of the EREs TSAs was assessed by BamQuery vs by the original study. Light blue dots represent each assessed MAP and the orange triangle represents the average.

f, Heatmap of average RNA expression of the EREs-derived TSAs in mTECs, normal breast tissues from GTEx (n=50), and triple-negative breast cancer samples from TCGA (n=158).
g, Heatmap of average RNA expression of published proteasomal splicing MAPs (n=99) in indicated tissues. The number of genomic locations expressed is presented on the left.
h, Heatmap of average RNA expression of EBV MAPs in indicated tissues.

1.23.5Discovery of tumor-specific antigens in diffuse large B-cell lymphoma

Given the capacity of BamQuery to prioritize TAs, we wondered whether it could help identify tumor-specific antigens (TSAs) from raw immunopeptidomic data. By using a proteogenomic approach enabling the identification of TSAs[10], we identified 6,869 MAPs from 3 published datasets of diffuse large B-cell lymphoma samples (DLBCL)[5].

We first quantified the expression of the 6,869 MAPs in mTECs with BamQuery. A genomic location was found for 6,833 of them and most of them (~86%) were highly expressed in mTECs (≥8.55 RPHM). To discriminate MAPs at risk of causing off-target toxicity when targeted, the remaining MAPs (14%) were queried in GTEx as well as in sorted benign B cells[39, 48], and 5% of them were retained as being lowly expressed (<8.55 RPHM). Finally, the retained MAPs being upregulated (fold change ≥5) by the DLBCL samples in TCGA vs benign B cells and having evidence of translation based on the presence of ribosomal profiling elongation reads (queried with BamQuery in matched RIBO-seq data[5], Extended Data Fig. 5a,b) were flagged as TSAs (67 MAPs, ~1%, Fig. 5a, Supplementary Tables 6-7). Among them, 11 were promising as they were highly shared between DLBCL patients (Fig. 5b).

BamQuery biotype classification showed that most TSAs derived from protein-coding regions of the genome, as only ~25% of them derived from non-coding RNA (20%), EREs (1%), and intronic (4%) regions (Fig. 5c). Furthermore, based on their high expression in testis, 29 TSAs were flagged as CTAs[49] (Supplementary Table 8) where most of them were known cancer biomarkers, supporting their relevance as immunotherapeutic targets. Additionally, upregulated TSAs in DLBCL samples compared to normal tissues (GTEx blood and benign B cells) had higher immunogenicity scores predicted by Repitope[50] compared to previously published non-immunogenic controls[51] (Fig. 5d). The expression of these TSAs correlated also with a greater expression of cytotoxic T cell markers (CD8A+CD8B), as well as with TCR signaling and other pro-

inflammatory responses in DLBCL patients (Fig. 5e-f, Supplementary Table 9), supporting the biological value of TSAs discovered with BamQuery.

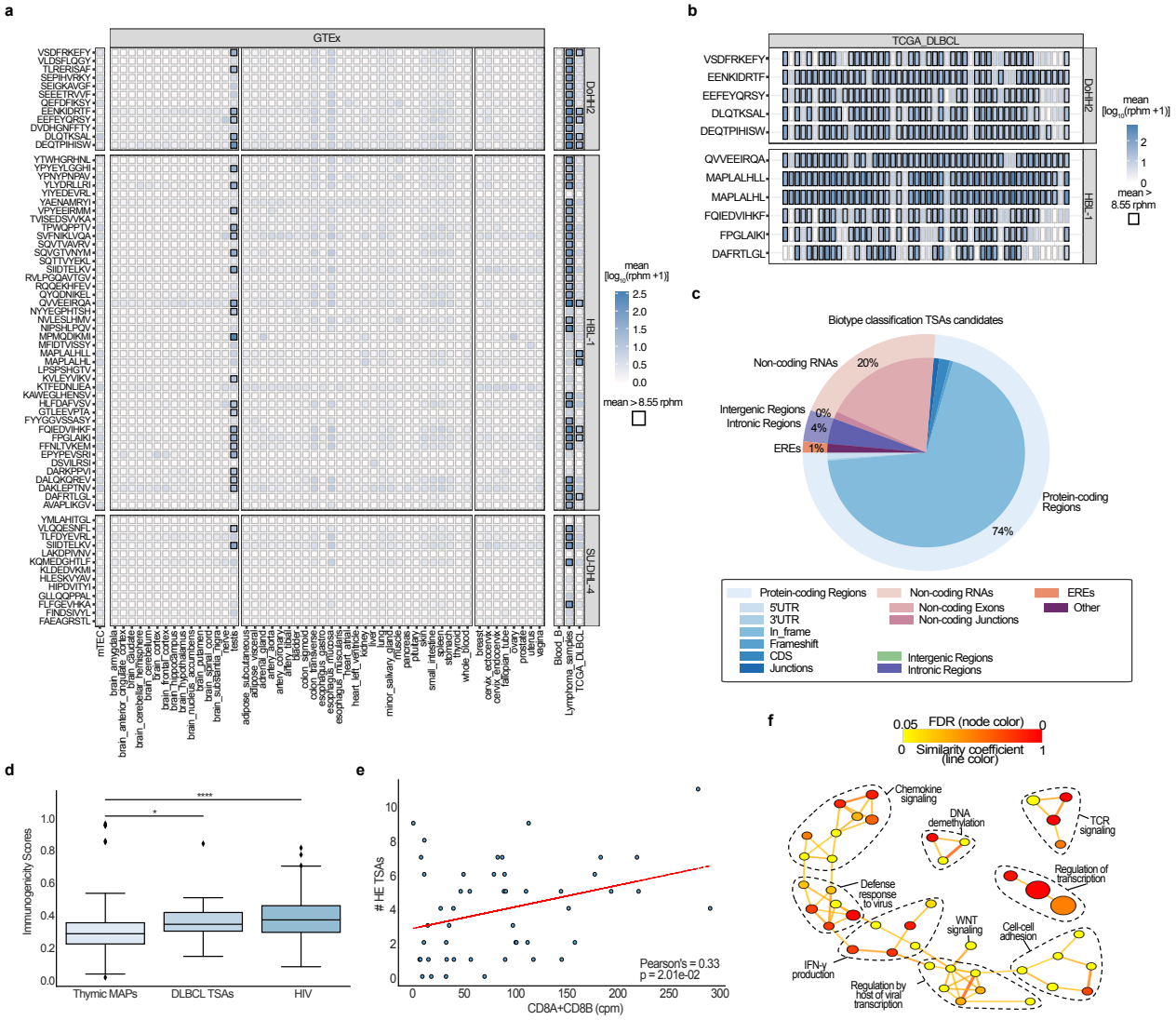


Figure 5. – Discrimination of potential immunotherapeutic targets in DLBCL.

a-c DLBCL MAPs, identified through a TSA-discovery proteogenomic approach, were searched with BamQuery in GTEx tissues (n=12–50 / tissue), mTECs (n=11), sorted blood B-cells (n=14), our DLBCL specimens (n=3) and/or TCGA DLBCL (n=48) bam files in unstranded mode with genome version GRCh38.104 and dbSNP version 155.

a, Heatmap of average RNA expression of 67 TSA candidates in indicated tissues. Boxes in which a peptide has an rphm>8.55 are highlighted in black.

b, Heatmap of average RNA expression of the highest shared and expressed TSA candidates (11) in cancer samples DLBCL from TCGA (n=48). Boxes in which MAPs expression (rphm) is >8.55 are highlighted in black.

c, Percentage of the most likely biotype attributed by BamQuery for TSA candidates (n=67).

d, Repitope immunogenic scores calculated for negative control thymic MAPs (n=158), highly expressed DLBCL TSAs (n=18, 25% of TSAs most upregulated by DLBCL TCGA versus normal blood in GTEx and sorted B cells), and positive control HIV MAPs (n=450). Mann-Whitney U test was used for comparisons (*p<0.05, ****p<0.0001).
e, Pearson's correlation in TCGA DLBCL patients (n=48) between the count of highly expressed (HE) TSAs expressed by each patient and the expression of cytotoxic T cells markers (CD8A+CD8B, in counts per million (cpm)). The red line is a linear regression.
f, Network analysis of GO term enrichment among genes overexpressed by patients expressing an above-median number of HE-TSAs. Line color reflects the similarity coefficient between connected nodes. Node color reflects the false discovery rate (FDR) of the enrichment. Node size is proportional to gene set size.

1.23.6BamQuery: an online tool to facilitate TA prioritization

We implemented an online portal to perform analyses on user-defined lists of MAPs. As we could not enable searches on GTEx (due to restricted use of these data), we included queries of MAPs in mTECs and DCs[39, 40] (Supplementary Table 2) as a proxy of tumor-specificity and immunogenicity. While expression in mTECs is considered a good proxy for normal cell expression[52, 53], we showed previously that mTECs share more transcriptomic features with epithelial than hematopoietic cells[9]. Prioritizing TAs based only on mTECs would therefore not be sufficient and we included DCs as they exert a non-redundant role in central tolerance establishment with mTECs[54].

To validate this choice, we randomly selected 10% of hematopoietic-specific (2,429) and 10% of epithelium-specific (3,237) MAPs from the HLA ligand atlas (Extended Data Fig. 6a, b). We queried their expression in mTECs, DCs, GTEx epithelial tissues, and a set of hematopoietic cells (Supplementary Table 2). At the RNA level, DCs and mTECs presented the highest hematopoietic and epithelial MAPs expression levels, respectively (Extended Data Fig. 6c, d). We refined our analysis by focusing on hematopoietic and epithelial MAPs being lowly (<8.55 RPHM) expressed in mTECs and DCs, respectively. This revealed dramatically higher expression of hematopoietic and epithelial MAPs in hematopoietic (highest in DCs) and epithelial tissues (highest in mTECs), respectively (Fig. 6a, b). We conclude that MAPs lowly expressed in mTECs are highly expressed in DCs, and vice-versa.

Next, we tested whether MAPs expression in mTECs and DCs would predict their immunogenicity. We queried in mTECs and DCs 1,180 and 4,917 non-mutated human MAPs verified experimentally as immunogenic and non-immunogenic, respectively, and curated in Ogishi et al.[50]. Immunogenic MAPs presented a lower expression than non-immunogenic MAPs in both types of samples (Fig. 6c). On this dataset, we trained a logistic regression model to classify immunogenic and non-immunogenic MAPs using the RPHM values of mTECs and DCs as features. Measurements of model performance and robustness using the cross-validation method (area under the ROC curve (AUC) = ~0.75, Extended Data Fig. 6e) showed that the RPHM values of MAPs in mTECs and DCs are predictors of MAP immunogenicity.

Finally, we evaluated whether MAP expression in both mTECs and DCs reflects the probability of presentation in benign tissues. Using 10% of random MAPs from the HLA Ligand Atlas (8,694), we found that MAPs lowly expressed in both mTECs and DCs were less presented (Fig. 6d) and expressed (Fig. 6e) in tissues of the HLA Ligand Atlas and GTEx, respectively. Upon examination of these MAPs features, we found that the probability of being highly expressed in mTECs and DCs increased exponentially with the number of possible genomic regions (Fig. 6f). Altogether, these results show that concomitant expression in mTECs and DCs expression is a reliable proxy of the presentation/expression in benign tissues and that MAPs having fewer possible regions of origin have a greater probability of being safe-to-target TAs.

The BamQuery public interface is accessible through http://bamquery.iric.ca/ and incorporates the logistic regression predictor model to report the conferred probability that a MAP is immunogenic. BamQuery is also available as a standalone version that can be configured to work with proprietary bam files. We believe that BamQuery will greatly help researchers in their attempts to identify specific and immunogenic TAs.

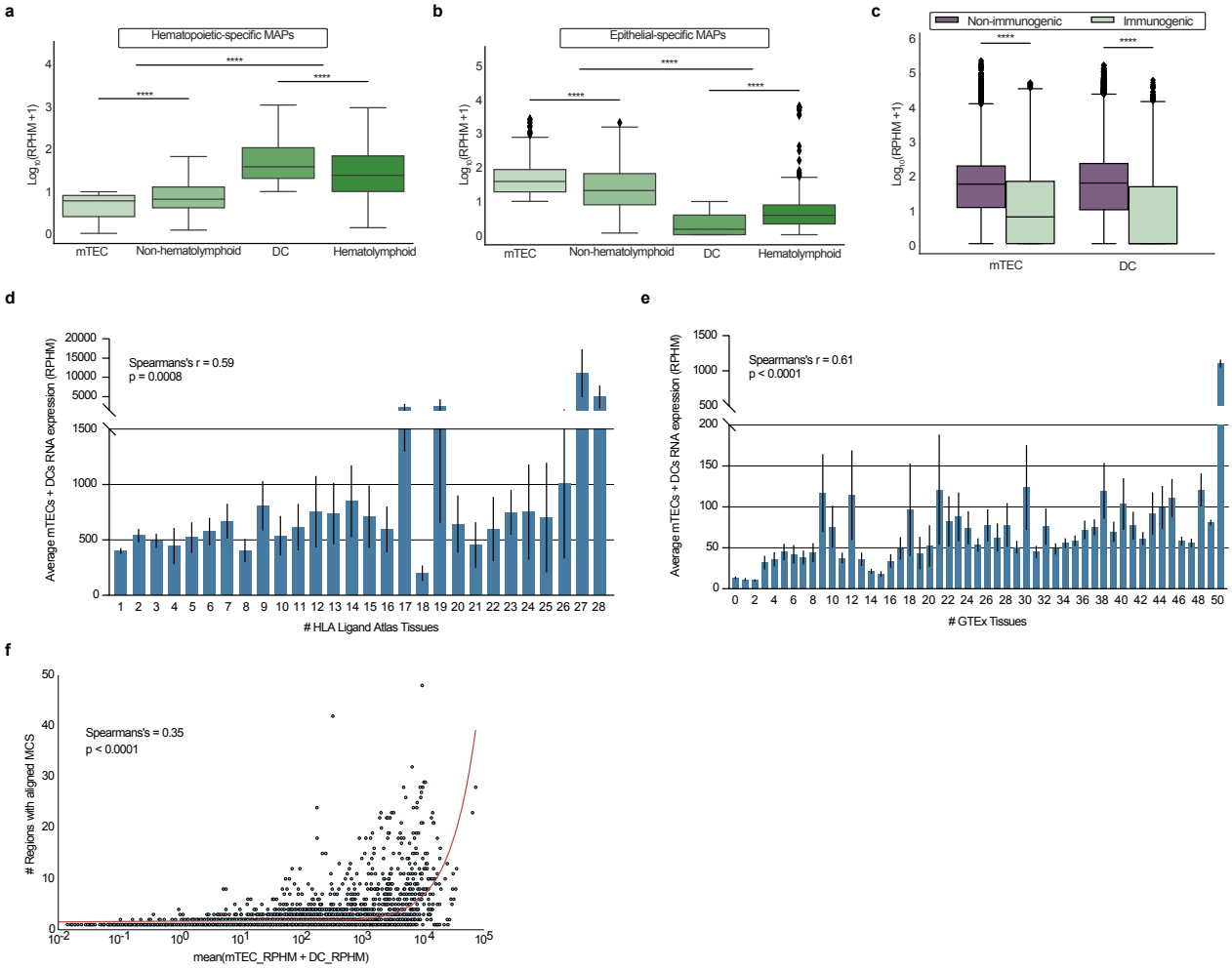


Figure 6. – BamQuery: an online tool to facilitate TAs prioritization.

a-b, Average RNA expression of hematopoietic-specific (a) and epithelial-specific (b) MAPs in mTECs (n = 11), non-hematolymphoid GTEx tissues (n = 2,389), DCs (n =19) and hematolymphoid GTEx tissues (n=196). Wilcoxon rank-sum test two-sided was used for comparisons (****p<0.0001).

c, Average RNA expression of non-mutated human immunogenic (n=1180) and non-immunogenic (n=4917) MAPs in mTECs (n = 11) and DCs (n =19). Mann-Whitney U test was used for comparisons (****p<0.0001).

d-e, Average mTECs+DCs RNA expression of a random selection of MAPs from the HLA Ligand Atlas (n=8621, 10% of the Atlas) as a function of the number of the HLA Ligand Atlas tissues presenting them (d) or as a function of the number of GTEx tissues in which the MAPs are expressed above an average of 8.55 RPHM (e). The average expression was correlated (Spearman) with the number of tissues. Error bar, SEM.

f, Spearman's correlation between the number of expressed genomic locations and the average expression in mTECs and DCs of the same MAPs used in (d). The red line is a linear regression (distorted by the log transformation of the x-axis).

1.24Discussion

Fuelled by studies focused on TAs, the immunopeptidomics field is expanding rapidly[3, 55, 56]. This expansion comes with an impressive diversity of homemade methodological approaches addressing the challenges raised by the characterization of non-canonical and mutated MAPs. Specifically, the fact that ~75% of the human genome can be transcribed[57] (and therefore possibly translated) evidenced the necessity of examining the expression of each region able to code for a presumed TA. BamQuery was designed not only to enable such examination but also to enable a uniformization of TA validation approaches across laboratories.

The recent discovery that a significant fraction of the immunopeptidome derives from non-coding regions has brought the contribution of the "dark genome" into the spotlight[2]. Since then, multiple studies have attempted to characterize cryptic MAPs, most often by using mass spectrometry informed by databases dedicated to the identification of specific classes of ncMAPs (intron-derived, ERE-derived, etc.)[8, 20, 58]. However, these approaches suffer from their dedication as the identified MAPs could also derive from other transcripts, absent from these databases. Accordingly, based on evidence showing that greater RNA expression confers a greater probability of MAPs generation[7, 13], we implemented a biotype annotation tool in BamQuery and showed that many presumed ncMAPs could be coded with greater probability by regions annotated with different biotypes. Strikingly, an important fraction of the canonical MAPs (~30%) could also be translated, with a greater probability, from non-canonical regions. While this result requires more in-depth analyses to elucidate the true origin of these MAPs, this possible dramatic contribution of the non-coding genome to the immunopeptidome is a sobering thought given that cryptic proteins are translated as efficiently as canonical proteins and generate MAPs 5-fold more efficiently per translation event[5].

Therapies targeting truly tumor-specific antigens can be highly effective[59], while those targeting antigens unsuspectedly expressed by normal cells can be lethal for patients[60]. Notably, BamQuery evidenced a high expression of many TAs, including mutated and ERE MAPs, in normal tissues, resulting from previously unreported coding regions and suggesting that targeting them would be unsafe. Here, we acknowledge that our approach can be considered very cautious.

While it is true that evaluating the RNA expression from all possible regions in the genome of a given MAP can be very stringent; we believe that the quality of the TAs is more important than the quantity when it comes to developing cancer treatments. Careful selection of TAs should prioritize those with a single genomic location and cancer-specific expression, to avoid undesirable effects. By using BamQuery in a rigorous manner, we can help to ensure that the TAs selected for immunotherapy development are of high quality and have the potential to be effective and safe treatments. Eventually, the availability of RIBO-seq data (which can be analyzed with BamQuery as well) could help to address this question. Meanwhile, in the absence of tools robustly predicting the translational origin of MAPs, the approach reported herein is the most circumspect for TA selection. Ideally, we recommend prioritizing TAs with a single possible region of origin (with cancer-specific expression) because other regions cannot code for such TAs in normal tissues.

Thanks to its exhaustivity, speed, ease of use, and versatility (bulk & single-cell RNA-seq + RIBO-seq, usable with mouse or human genome, on any kind of wild-type or mutated MAPs), BamQuery enables for the first time a uniformization of proteogenomic analyses in MHC-I immunopeptidomics.

1.25Methods

1.25.1Data and Code Availability

The Python and R scripts generated during this study are available on GitHub, https://github.com/lemieux-lab/BamQuery. The standalone version of BamQuery can be downloaded at http://bamquery.iric.ca/installation.html. Details regarding all samples used in this study are listed in Supplementary Table 2.

1.25.2Datasets

The eight human mTEC samples have been prepared and sequenced for previous studies of our team (GEO:GSE127825 and GEO:GSE127826) (Larouche et al., 2020[20]; Laumont et al., 2018[10]). Three additional mTEC samples were published (ArrayExpress:E-MTAB-7383) by Fergurson et al.[25].

1.25.3BamQuery

BamQuery is designed to analyze MAPs ranging in length from 8 to 11 amino acids (aa). As peptide input, BamQuery supports three different formats that can be pulled into a single input file.

A) Peptide mode: only the amino acid sequence of the MAP is provided, hence BamQuery performs a comprehensive search for its RNA-seq expression. All results reported in the present article were obtained with this mode.

B) MAP coding sequence (MCS) mode: the amino acid sequence of the MAP is provided, hence BamQuery performs the search for the expression of the given MCS only.

C) Manual mode: the amino acid sequence of the MAP is provided followed by an MCS, the corresponding location in the genome of the given MCS, and the strand (+ forward or - reverse), whereby BamQuery performs the expression search at the given location for the given MCS at the given genomic location and strand (useful to evaluate the expression of mutated MAPs whose genomic location is known but which cannot be located by BamQuery due to unavailable annotations in dbSNP or STAR failure).

BamQuery performs five important steps for each peptide queried.

1. **Reverse translation of MAPs.**

Each input MAP in peptide mode is reverse-translated into all possible MCS. The MCS are compiled into a fastq file.

2. **Identification of genomic locations.**

MCS are then mapped to the reference genome (user-defined, meaning that several genome versions are supported (GENCODE 26, 33 or 38)) using STAR v2.7.9.a[17] running with default parameters except for –seedSearchStartLmax, --winAnchorMultimapNmax, --outFilterMultimapNmax, --limitOutSJcollapsed, --limitOutSAMoneReadBytes, --alignTranscriptsPerWindowNmax, --seedNoneLociPerWindow, --seedPerWindowNmax, --alignTranscriptsPerReadNmax that were replaced by 20, 10.000, 10.000, 5.000.000 , 2.660.000, 1.000, 1.000, 1.000, 20.000, respectively. MCS genomic locations (perfect alignments) are selected from the output STAR file Aligned.out.sam. Perfect alignments are defined as MCS matching exactly the reference genomic sequence or as MCS bearing mismatches annotated as known polymorphisms in the dbSNP database (user-selected dbSNP 149, 151, or 155 releases). Therefore, each alignment included in Aligned.out.sam is exanimated to compare the read sequence nucleotide by nucleotide against the reference genomic sequence at that position (assessed using the samtools fetch command within python via the pysam (https://github.com/pysam-developers/pysam) library at the genomic location of the given alignment). If a difference is detected between a nucleotide of the aligned read sequence and the nucleotide of the reference genomic sequence at a given position, the position is queried in the python dictionary containing the SNVs of the dbSNP database selected by the user. If all discrepancies in the current alignment are known (supported by the SNVs in the dbSNP database) the alignment is retained as it is considered perfect, otherwise, the alignment is discarded. To reduce the complexity of tracing perfect STAR alignments, only single nucleotide variants (SNVs) of dbSNP annotations were considered to define perfect alignments.

3. MAP RNA-seq reads counting.

Next, the expression of each MCS is queried in each BAM file (CRAM files are also supported) using the samtools view[61] command within python via the pysam library (only primary alignment reads (pysam option -F0X100), originally present in fastq files, are queried) at their respective genomic location. BamQuery supports RNA-seq unstrandedness / strandedness libraries (user-defined parameter, default: strandedness). To collect reads in unstranded libraries, the -F0X100 option is used in the pysam view command. In stranded libraries, depending on the sequencing reads type (single-end, paired-end), library preparation (forward or backward) and sense of the MCS genomic location (forward or backward), the options in the pysam view command are: -F0X100 & -f0X50 for R1 mate and -F0X100 & -f0XA0 for R2 mate in paired-end, forward library and reverse genomic location; -F0X100 & -f0X60 for R1 mate and -F0X100 & -f0X90 for R2 mate in paired-end, forward library and forward genomic location; -F0X110 for R1 mate in single-end, forward library and forward genomic location; -F0X100 & -f10 for R1 mate in single-end, forward library and reverse genomic location; -F0X100 & -f0X60 for R1 mate and -F0X100 & -f0X90 for R2 mate in paired-end, reverse library and reverse genomic location; -F0X100 & -f0X50 for R1 mate and -F0X100 & -f0XA0 for R2 mate in paired-end, reverse library and forward genomic location; -F0X110 for R1 mate in single-end, reverse library and reverse genomic location; -F0X100 & -f10 for R1 mate in single-end, reverse library and forward genomic location. The retrieved reads are examined one by one and counted if they exactly span the queried MCS at the genomic location. Therefore, each retrieved read is transformed into a list in Python and its alignment location is transformed into an array containing the location of each amino acid in the read. The indices of the array locations corresponding to the first and last amino acid locations in the MCS at a given genomic location are used to extract from the read list the subsequence that is compared to the MCS. If both the MCS and the subsequence of a retrieved read are the same, the read count for the current MCS increases by one. Finally, the total read count ($\mathrm{tr}_{\mathrm{MAP}}$) for a given MAP is computed by summing all RNA-seq reads from all MCS genomic locations.

4. Normalization

The $\mathrm{tr}_{\mathrm{MAP}}$ count is transformed into “reads per hundred million” values (RPHM) by normalizing them with the total number of primary reads sequenced (corresponding to the total read number

present in fastq files) according to the formula: $\mathrm{RPHM} = \frac{\mathrm{tr_{MAP}}}{\mathrm{R_t}} * 10^8$ where $\mathrm{R_t}$ represents the total number of primary RNA-seq reads of the sample. These final values are log-transformed $\log_{10}(\mathrm{RPHM} + 1)$ to allow comparison and averaging between samples, thus removing the bias of large values.

5. Biotype classification

All genomic locations identified for each MAP are compiled into a bed file and their biotypes are obtained using BEDtools[62] intersect with the following options -a (annotation file), -b (genomic locations), -wao (writes the original annotation and genomic location entries along with the number of base pairs of overlap between the two features), and the following annotations: RepeatMasker (GRCh38/hg38 assembly, to annotate the EREs) and GENCODE (for all other biotypes, gene set annotations releases v26_88, v33_99, v38_104). The complete list of biotypes annotated by BamQuery based on RepeatMasker and GENCODE can be consulted at http://bamquery.iric.ca/biotype_classification.html.

Given that MAPs may have alignments in regions where several different biotypes overlap (such as protein-coding transcripts overlapping with non-coding RNAs, see the example shown in Extended Data Fig. 7a), we used the expectation-maximization (EM) statistical model to estimate, for each biotype, the read distribution coefficient. In this model, reads at each genomic location are weighted for each biotype at the given location according to their coefficients and consequently, the biotype of each MAP is scored according to the percentage of reads corresponding to each biotype (in-frame, introns, ncRNA, ERE, etc.). The EM algorithm iterates between the expectation (E) and maximization (M) step until the parameter set of the last iteration is unchanged, therefore finding the parameter set that maximizes the posterior probability of the observed data, in our case the reads that overlap with one or more biotypes. To train the EM algorithm, we first collected canonical and ncMAPs (Supplementary Table 1) and ran BamQuery on normal and cancer datasets (normal: GTEx and mTECs, cancer: TCGA) to obtain the total reads covering each MAP at each MCS genome location. We then computed the probability of each biotype as follows:

Let $\emptyset = (\emptyset_A, \emptyset_B, \emptyset_c \dots)$, be the set of parameters to estimate, where $\emptyset_A, \emptyset_B, \emptyset_c \dots$ are the probabilities that the read belongs to the In_frame (A), non_coding_exon (B), intron (C), etc. biotypes. EM starts with an arbitrary initial estimation of 0.1 for each biotype's probability. In the E-step, the distribution of the total number of reads for each MAP is computed using the current biotype's parameters, as follows:

Let $R_i =$ total reads of MAP_i

$$Z(\emptyset_j^t, L_i) = \frac{\sum_{k=1}^{L} \left(r_k * \frac{\emptyset_j^t}{\sum_{b=1}^{B} \emptyset_b^k} \right)}{R_i}$$

Where $\emptyset_j^t$ is the current probability for biotype j in MAP_i. L_i is the MCS genome locations for MAP_i. r_k is the number of reads overlapping location k and B is the set of biotypes overlapping the location k.

In the M-step, the new set of parameters is determined using the current computations, as follows:

$$\emptyset_j^{t+1} = \frac{\sum_{i=1}^{MAPs} \emptyset_j^t}{\text{total MAPs}}$$

Where $\emptyset_j^{t+1}$ is the new probability for biotype j obtained after summing all the probabilities distributions of all MAPs computed in the last E step and normalizing by the total number of MAPs. The iterative process concludes if the following condition is met for all biotypes: $\emptyset_j^t = \emptyset_j^{t+1}$ and the last set of estimated parameters is used to assign the proportion of reads assigned to each biotype at any genomic location.

Therefore, BamQuery scores for each MAP the biotype as the percentage of reads assigned to each biotype class (in-frame, introns, ncRNA, ERE, etc.). For example, a canonical MAP with alignments in non-canonical regions could be indicated as follows In_frame: 84.09% - Intronic: 15.91%, meaning that ~84% of the total reads overlap with a known transcript and that the MAP is within the known protein frame, while ~16% of the reads overlap with transcripts in an intronic region.

BamQuery informs the biotype of each MAP in three different settings, as follows:

1. Biotype computed for each MCS genome location: BamQuery reports the percentage contribution of the biotypes overlapping the given location. The percentage of each biotype is calculated as the coefficient of each biotype normalized by the sum of the coefficients of all biotypes in the location, as follows:

$$\emptyset_{i,j}^{k} = \frac{\emptyset_{i,j}^{k}}{\sum_{b=1}^{B} \emptyset_{b}^{k}}$$

Where $\emptyset_{i,j}^{k}$ is the coefficient assigned to the biotype j for the MAP_i at the location k.

2. Biotype computed from all MCS genome locations found in the set of queried samples: the biotype of each MAP is assigned based on the total read count in the sample set. This calculation follows three steps:
    a) The total number of reads in each MCS genome location is distributed according to the biotype percentages assigned to the location in the previous step.
    b) Normalization of the distributed count of reads by the total number of reads in the entire set of samples.
    c) The final biotype of each MAP is obtained by summing all normalized reads distributions across its MCS genomic location.
3. Biotype for each subset of samples (e.g., GTEx, TCGA, mTEC samples): the biotype of each peptide is assigned following the same steps as before but according to the total count of reads in each subset of samples.
4. Best guess biotype: BamQuery also reports the most likely biotype for each MAP (Best Guess) following the rules below:
    a) Since a MAP is most likely to be generated from a known canonical protein if the MAP ever appears in-frame of a protein the best guess assigned is In-frame with the certainty given in the biotype classification.
    b) Otherwise, the best guess biotype is assigned according to the biotype with the highest percentage of the biotype ranking.

Full documentation of supported options, examples of use, and descriptions of BamQuery reports can be found at http://bamquery.iric.ca/

1.25.4K-mer databases

K-mer databases were generated by retrieving the primary mapped reads from the bam files of each mTEC sample with samtools view[61] (-F260 option) followed by SamToFastq from Picard tools to recover R1 and R2 fastq files (https://broadinstitute.github.io/picard/index.html). Next, R1 reads were reverse complemented using the fastx_reverse_complement function of the FASTX-Toolkit v0.0.14. and fastq files of all mTEC samples were concatenated. Finally, Jellyfish count (v2.2.3, options -m = 27 and -s =1G)[21] was used to generate the database from the fastq file, and jellyfish query was used to query the MCS in the database.

1.25.5Kallisto quantification

Transcript expression quantifications of mTEC samples were performed with kallisto[23] v0.43.0 quant with default parameters except for --rf-stranded. The expression of each HLA atlas peptide was obtained from the mean TPM expression value of all transcripts associated with the peptide source genes.

1.25.6BamQuery Accuracy

For each MCS of the canonical nine-mer MAPs, we defined the BamQuery accuracy, as follows:

$$\text{Accurracy} = 100\% - \text{error rate}$$

$$\text{error rate} = \frac{|\text{BamQuery read count} - \text{Jellyfish read count}|}{\text{Jellyfish read count}} * 100$$

Therefore, the accuracy is the difference in the error rate with respect to 100%, the error rate being the percentage value of the difference in the observed MCS read count in BamQuery and the actual MCS read count in Jellyfish.

1.25.7Single cell RNA-seq analyses

Previously published single-cell RNA-seq data from the healthy and cancerous lungs were downloaded from the NCBI BIOPROJECT (accession number PRJEB31843) and Array Express (accession number E-MTAB-6653), respectively. Reads were aligned on the human reference genome (GRCh38) using STAR version 2.7.9a[17]. Cell population annotations were performed using gene lists from Madissoon et al.[34] and Lambrechts et al.[33] for the healthy and cancerous lung datasets, respectively. For the subsequent profiling of MAP expression with BamQuery, the HCATisStab7509734 and the BT1375 samples were subsampled from the healthy and cancerous lung datasets, respectively. For the subsequent profiling of MAP expression with BamQuery, the HCATisStab7509734, and the BT1375 samples were subsampled from the healthy and cancerous lung datasets, respectively. For both genes and MAPs expression, read counts were normalized based on the total number of reads detected in each cell (size factor) with the computeSumFactors function of the scran v1.18.7 R package. Normalized read counts were log-transformed with the logNormCounts function of the scuttle package (v1.8.4), and dimensionality reduction was performed with scran (v1.18.7). The differential expression analyses of MAPs between the cell populations of the healthy and cancerous lungs were performed with the FindAllMarkers function of Seurat with the MAST model. Cells of the healthy lungs were also re-clustered based on their MAP expression using the runUMAP and runTSNE functions of the scater package (v1.18.6), and cell lineages and populations previously annotated based on gene expression were represented on the resulting UMAP and TSNE graphs. Co-expression of MAPs in the tumor cells of the lung was also assessed. To do so, we selected the MAPs identified as overexpressed in lung cancer cells by the differential expression analysis and computed spearman correlations between the expression of each possible pair of MAPs. Finally, MAP expression in the cell populations of the healthy lung was visualized with violin plots using the VlnPlot function of the Seurat package (v.4.1.0)[63].

1.25.8Immunogenicity predictions

Immunogenicity predictions of HE-TSAs were performed with Repitope[50]. Feature computation was performed with the predefined MHCI_Human_MinimumFeatureSet variable and updated (July 12, 2019) FeatureDF_MHCI and FragmentLibrary files provided on the Mendeley repository

of the package (https://data.mendeley.com/datasets/sydw5xnxpt/1). HIV MAPs (positive control) were obtained from https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html .

1.25.9Differential gene expression analysis

Transcript expression quantifications were performed on TCGA DLBCL bulk RNA-seq samples with kallisto v0.43.0 with default parameters. Then, with BamQuery, we attributed to each patient a count of highly expressed TSA transcripts (HE-TSA), i.e., the number of TSAs whose expression was above their median RNA expression across all patients having a non-null expression of the given TSAs. Patients having an above-median number of HE-TSAs (n=26) were compared to those below-median (n=22) through a differential gene expression analysis. This analysis was conducted in R3.6.1 as reported previously[64]. In brief, raw read counts were converted to counts per million (cpm), normalized relative to the library size, and lowly expressed genes were filtered out by keeping genes with cpm >1 in at least 2 samples using edgeR 3.26.8 and limma 3.40.6. This was followed by voom transformations and linear modeling using limma's lmfit. Finally, moderated t-statistics were computed with eBayes. Genes with p-values < 0.05 and $-1 \geq \log_2(FC) \geq 1$ were considered significantly differentially expressed (386 genes upregulated and 1304 downregulated).

1.25.10 GO term and enrichment map analyses

Biological-process gene-ontology (GO) term over-representation was performed with DAVID (https://david.ncifcrf.gov) on genes upregulated by DLBCL patients expressing high levels of HE-TSAs. Functional annotations with p-value < 0.05 were considered significant. The GO-term list was then imported in Cytoscape v3.7.2 and used to cluster redundant GO terms and visualize the results with EnrichmentMap v3.2.1 and default parameters. The network was visualized using the default "Prefuse Force-Directed Layout" in Cytoscape. Groups of similar GO terms were manually circled.

1.25.11 Other bioinformatic analyses

Amino acid compositions were assessed with the ProtParam module of Biopython. Read coverage in scRNA-seq data was evaluated with the geneBody_coverage module of RSeQC on the bam file generated by CellRanger. Codon frequencies were obtained from the codon usage database (http://www.kazusa.or.jp/codon/).

1.25.12 Logistic regression model

The cross-validation procedure was employed to divide the training dataset into training and validation subsets. This was done using the StratifiedShuffleSplit function from the sklearn Python library, with 10 splits and a test size of 0.2. Subsequently, the logistic regression model from the sklearn Python library was utilized to classify MAPs as either immunogenic (1,180) or non-immunogenic (4,917). The model was trained using two highly correlated features, mTEC_expression and DC_expression, with a Pearson's correlation coefficient of 0.84 ($p < 0.0001$). The target variable was binary, indicating immunogenic (1) or non-immunogenic (0). The model was trained using default parameters, except for the solver, which was set to "liblinear", and the inclusion of sample_weights to account for class imbalance. The weights were calculated as the ratio of positive to negative samples in each split.

1.25.13 Construction of MS database for TSA identification

We used RNA-seq data from 3 published datasets of diffuse large B-cell lymphoma samples (DLBCL)[5]. Cancer-specific proteomes were built using k-mer profiling as described previously[10]. RNA-Seq reads were chopped into 33-nucleotide k-mers and only those present <2 in mTECs were kept. Overlapping k-mers were assembled into contigs, which were then three-frame translated and linked using “JJ” as separators. This database was concatenated with each sample’s canonical proteome for MAP identification.

1.25.14 Quantification and Statistical Analysis

All statistical tests used are mentioned in the respective figure legends. For all statistical tests, *, **, ***, *** and **** refers to $p< 0.05$, $p< 0.01$, $p< 0.001$ and $p< 0.0001$, respectively, and are reported in the figures. Correlations were assessed with the Pearson or Spearman correlation

coefficient, a red line in the correlation plots represents the linear regression. Plots and statistical tests were performed using scipy.stats and seaborn packages of Python v3.6.8. Unless mentioned otherwise, all boxes in box plots show the third (75th) and first quartiles (25th) and the box band shows the median (second quartile) of the distribution; whiskers extend to 1.5 times the interquartile distance from the box. Unless mentioned otherwise, all bar plots show the average with error bars: 95% confidence interval (CI).

1.26Acknowledgments

We are grateful to Qingchuan Zhao, Assya Trofimov, Nandita Noronha, and Caroline Labelle for useful biological insights, suggestions, and testing BamQuery. We also thank all other members of our laboratories for their thoughtful recommendations. We thank Eric Audemard and Geneviève Boucher of the IRIC bioinformatic platform for assistance with bioinformatics tools. This study was supported by grants from the Canadian Cancer Society (707264), and the Canadian Institutes of Health Research (FDN 148400). GE is supported by post-doctoral fellowships from the IRIC, FRQS, The Cole Foundation, and the FNRS. We thank the Genotype-Tissue Expression (GTEx) Project for providing RNA-seq data. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS

1.27References

1. Pishesha, N., Harmand, T.J. & Ploegh, H.L. A guide to antigen processing and presentation. *Nat Rev Immunol* (2022).
2. Lang, F., Schrörs, B., Löwer, M., Türeci, Ö. & Sahin, U. Identification of neoantigens for individualized therapeutic cancer vaccines. *Nat Rev Drug Discov* **21**, 261-282 (2022).
3. Chong, C., Coukos, G. & Bassani-Sternberg, M. Identification of tumor antigens with immunopeptidomics. *Nat Biotechnol* **40**, 175-188 (2022).
4. Laumont, C.M. et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun* **7**, 10238 (2016).
5. Ruiz Cuevas, M.V. et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep* **34**, 108815 (2021).
6. Erhard, F. et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**, 363-366 (2018).
7. Khurana, E. et al. Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108 (2016).
8. Chong, C. et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* **11**, 1293 (2020).
9. Ehx, G. et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity* **54**, 737-752.e710 (2021).
10. Laumont, C.M. et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med* **10** (2018).
11. Zhao, Q. et al. Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer. *Cancer Immunol Res* (2020).
12. Gee, M.H. et al. Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. *Cell* **172**, 549-563.e516 (2018).
13. Probst, P. et al. Sarcoma Eradication by Doxorubicin and Targeted TNF Relies upon CD8(+) T-cell Recognition of a Retroviral Antigen. *Cancer Res* **77**, 3644-3654 (2017).

14. Ehx, G. & Perreault, C. Discovery and characterization of actionable tumor antigens. *Genome Medicine* **11**, 29 (2019).
15. Pearson, H. et al. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest* **126**, 4690-4701 (2016).
16. Takahama, Y., Ohigashi, I., Baik, S. & Anderson, G. Generation of diversity in thymic epithelial cells. *Nat Rev Immunol* **17**, 295-305 (2017).
17. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
18. Smigielski, E.M., Sirotkin, K., Ward, M. & Sherry, S.T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**, 352-355 (2000).
19. Marcu, A. et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer* **9** (2021).
20. Larouche, J.D. et al. Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Med* **12**, 40 (2020).
21. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
22. Ballouz, S., Dobin, A., Gingeras, T.R. & Gillis, J. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Res* **46**, 5125-5138 (2018).
23. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527 (2016).
24. Liu, X. et al. A comparison of transcriptome analysis methods with reference genome. *BMC Genomics* **23**, 232 (2022).
25. Fergusson, J.R. et al. Maturing Human CD127+ CCR7+ PDL1+ Dendritic Cells Express AIRE in the Absence of Tissue Restricted Antigens. *Front Immunol* **9**, 2902 (2018).
26. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
27. Varenne, S., Buc, J., Lloubes, R. & Lazdunski, C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol* **180**, 549-576 (1984).
28. Sørensen, M.A., Kurland, C.G. & Pedersen, S. Codon usage determines translation rate in Escherichia coli. *J Mol Biol* **207**, 365-377 (1989).

29. Yu, C.H. et al. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol Cell* **59**, 744-754 (2015).
30. Yewdell, J.W. & Holly, J. DRiPs get molecular. *Curr Opin Immunol* **64**, 130-136 (2020).
31. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
32. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098 (2013).
33. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med* **24**, 1277-1289 (2018).
34. Madissoon, E. et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol* **21**, 1 (2019).
35. Petti, A.A. et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun* **10**, 3660 (2019).
36. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
37. Patrick, R. et al. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biol* **21**, 167 (2020).
38. Hirama, T. et al. Proteogenomic identification of an immunogenic HLA class I neoantigen in mismatch repair-deficient colorectal cancer tissue. *JCI Insight* **6** (2021).
39. Choi, J. et al. Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Res* **47**, D780-D785 (2019).
40. Silvin, A. et al. Constitutive resistance to viral infection in human CD141(+) dendritic cells. *Sci Immunol* **2** (2017).
41. Rivero-Hinojosa, S. et al. Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat Commun* **12**, 6689 (2021).
42. Tan, X. et al. dbPepNeo: a manually curated database for human tumor neoantigen peptides. *Database (Oxford)* **2020** (2020).

43. Xia, J. et al. NEPdb: A Database of T-Cell Experimentally-Validated Neoantigens and Pan-Cancer Predicted Neoepitopes for Cancer Immunotherapy. *Front Immunol* **12**, 644637 (2021).
44. Bonaventura, P. et al. Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy. *Sci Adv* **8**, eabj3671 (2022).
45. Liepe, J. et al. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* **354**, 354-358 (2016).
46. Bjornevik, K. et al. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* **375**, 296-301 (2022).
47. Tsang, M.L. & Münz, C. Cytolytic T lymphocytes from HLA-B8+ donors frequently recognize the Hodgkin's lymphoma associated latent membrane protein 2 of Epstein Barr virus. *Herpesviridae* **2**, 4 (2011).
48. Pabst, C. et al. GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood* **127**, 2018-2027 (2016).
49. Almeida, L.G. et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res* **37**, D816-819 (2009).
50. Ogishi, M. & Yotsuyanagi, H. Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space. *Front Immunol* **10**, 827 (2019).
51. Adamopoulou, E. et al. Exploring the MHC-peptide matrix of central tolerance in the human thymus. *Nat Commun* **4**, 2039 (2013).
52. Rattay, K., Meyer, H.V., Herrmann, C., Brors, B. & Kyewski, B. Evolutionary conserved gene co-expression drives generation of self-antigen diversity in medullary thymic epithelial cells. *J Autoimmun* **67**, 65-75 (2016).
53. Kadouri, N., Nevo, S., Goldfarb, Y. & Abramson, J. Thymic epithelial cell heterogeneity: TEC by TEC. *Nat Rev Immunol* **20**, 239-253 (2020).
54. Audiger, C., Rahman, M.J., Yun, T.J., Tarbell, K.V. & Lesage, S. The Importance of Dendritic Cells in Maintaining Immune Tolerance. *J Immunol* **198**, 2223-2231 (2017).

55. Arnaud, M. et al. Sensitive identification of neoantigens and cognate TCRs in human solid tumors. *Nat Biotechnol* **40**, 656-660 (2022).
56. Ouspenskaia, T. et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol* **40**, 209-217 (2022).
57. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
58. Smart, A.C. et al. Intron retention is a source of neoepitopes in cancer. *Nature Biotechnology* **36**, 1056-1058 (2018).
59. Welters, M.J. et al. Induction of tumor-specific CD4+ and CD8+ T-cell immunity in cervical cancer patients by a human papillomavirus type 16 E6 and E7 long peptides vaccine. *Clin Cancer Res* **14**, 178-187 (2008).
60. Morgan, R.A. et al. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J Immunother* **36**, 133-151 (2013).
61. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
62. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
63. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).
64. Noronha, N. et al. Major multilevel molecular divergence between THP-1 cells from different biorepositories. *Int J Cancer* **147**, 2000-2006 (2020).

1.28 Supplemental Information

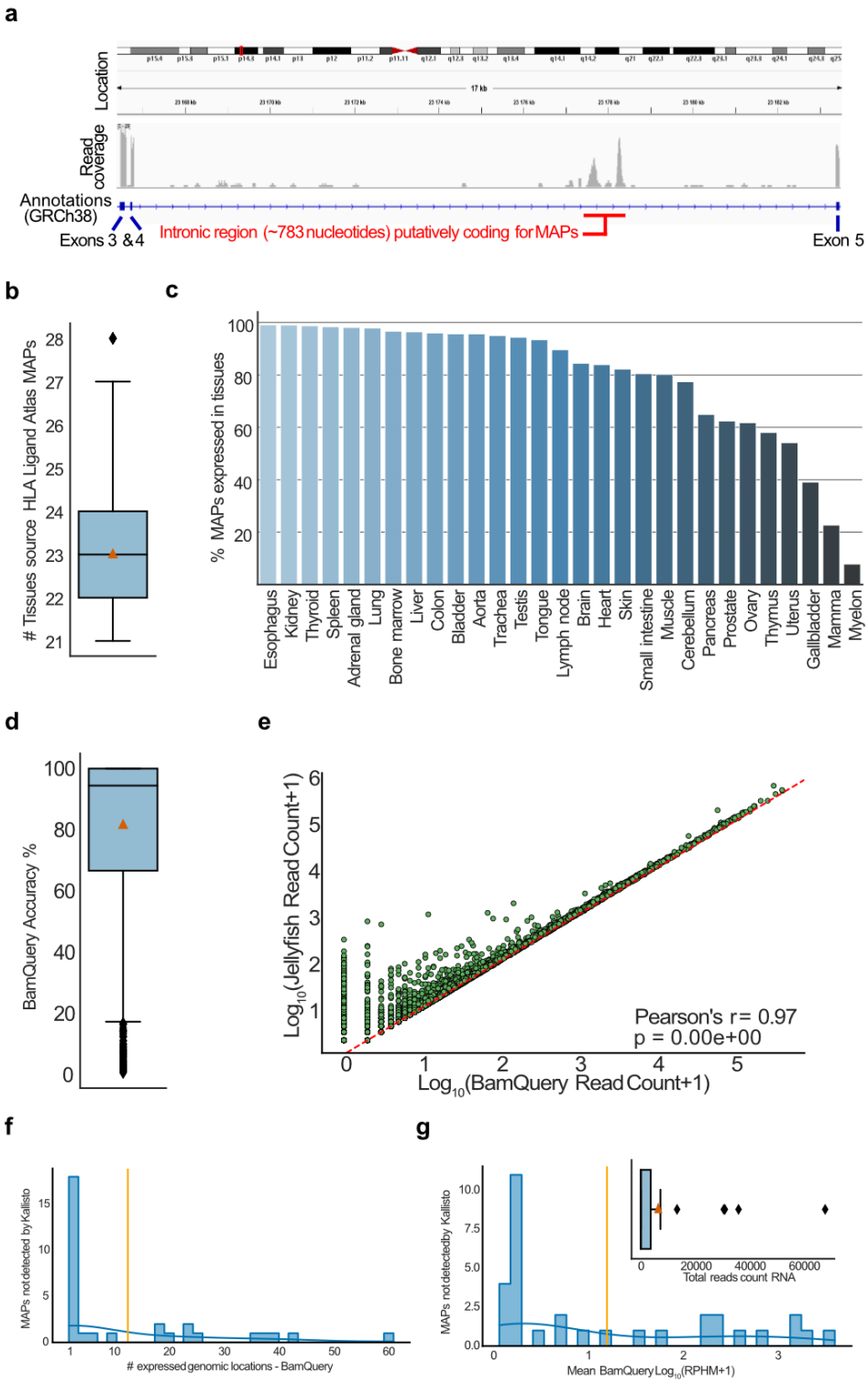


Figure 7. – Origin canonical MAPs and BamQuery's quality control

d-**f**, Published MAPs reported as canonical (n=1,702) were searched with BamQuery in mTEC bam files in stranded with genome version GRCh38.p13, gene set annotations release v38_104, dbSNP release 151, keeping variants alignments, and allowing higher levels of MCS alignments by STAR.

a, Genome browser (IGV) illustration for the gene LINC02718 (chr11:23,166,352-23,183,625) in a sample of acute myeloid leukemia (GSM4432540 on GEO) of the heterogeneity of read coverage observed in a typical intronic region (between exon 4 and 5). This would make the usage of genomic annotations irrelevant to quantify the expression of the small region putatively coding for MAPs as most of the annotated intron is not or lowly covered by reads (depth of coverage represented in grey).

b, Number of tissues at the origin of the canonical MAPs from the HLA ligand atlas shared in at least 20 tissues (n=1,702). Orange triangle represents the average (23).

c, Percentage of MAPs (n=1,702) presented by the indicated tissues.

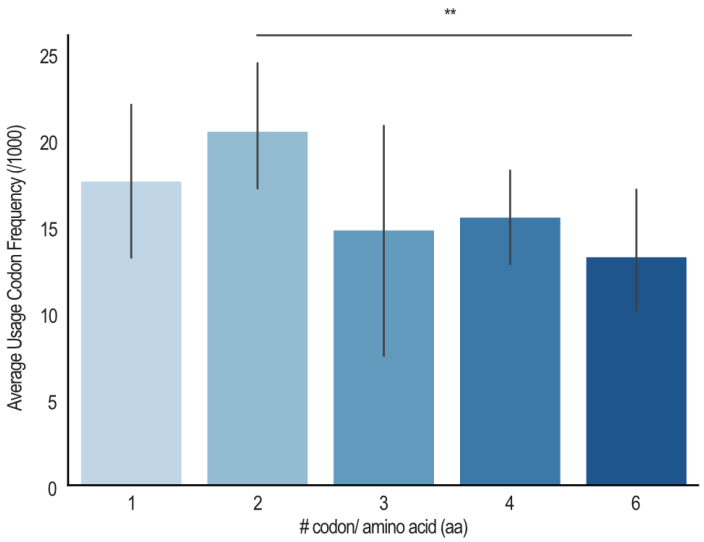
d, Percentage accuracy measured between BamQuery-acquired read counts and Jellyfish's K-mer counts for nine-mer MAPs (n=1,211). Orange triangle represents the average percentage accuracy (82%).

e, Pearson's correlation between BamQuery-acquired read counts and Jellyfish's K-mer counts for MCS of canonical nine-mer MAPs (n=1,211) from the HLA Ligand Atlas (present in at least 20 different tissues) and 8 mTEC samples

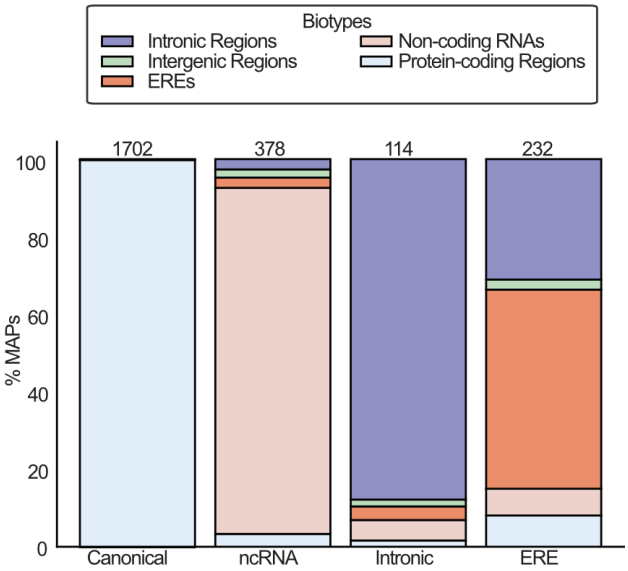
f, Number of genomic locations detected by BamQuery for the 32 MAPs undetected in Kallisto's TPM quantification. Orange line represents the average number of genomic locations (11).

g, BamQuery-acquired RPHM expression for the 32 MAPs undetected in Kallisto's TPM quantification. Orange line represents RPHM expression average (1.1). Inside panel: total RNA-seq BamQuery-acquired reads for the 32 MAPs undetected by Kallisto. Orange triangle represents the average total RNA-seq BamQuery-acquired reads (n=6,474).

a


**
Average Usage Codon Frequency (/1000)
25
20
15
10
5
0
1
2
3
4
6
codon/ amino acid (aa)

b


Biotypes
Intronic Regions
Intergenic Regions
EREs
Non-coding RNAs
Protein-coding Regions
1702
378
114
232
100
80
60
40
20
0
% MAPs
Canonical
ncRNA
Intronic
ERE

c

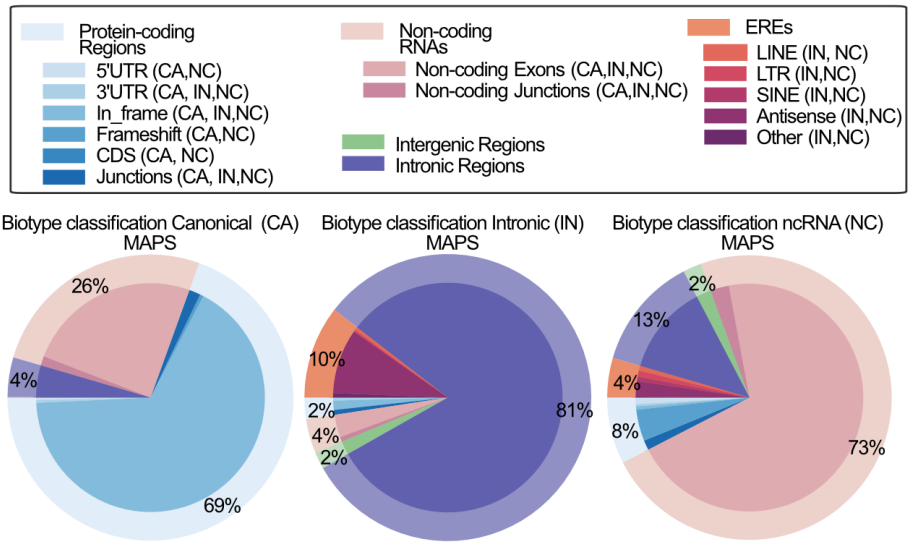

Protein-coding Regions
5'UTR (CA,NC)
3'UTR (CA, IN,NC)
In_frame (CA, IN,NC)
Frameshift (CA,NC)
CDS (CA, NC)
Junctions (CA, IN,NC)
Non-coding RNAs
Non-coding Exons (CA,IN,NC)
Non-coding Junctions (CA,IN,NC)
Intergenic Regions
Intronic Regions
EREs
LINE (IN, NC)
LTR (IN,NC)
SINE (IN,NC)
Antisense (IN,NC)
Other (IN,NC)
Biotype classification Canonical (CA) MAPS
26%
4%
69%
Biotype classification Intronic (IN) MAPS
10%
2%
4%
2%
81%
Biotype classification ncRNA (NC) MAPS
2%
13%
4%
8%
73%

Figure 8. – Immunopeptidome properties of canonical and noncanonical MAPs.

a, Average frequency of codons (among 1000 codons located in human reference protein coding sequences) encoding each of the 20 amino acids. Codons of amino acids encoded by the same number of different synonymous codons were grouped together (x axis).

b, Percentage of MAPs attributed to indicated biotypes by BamQuery based on the best guess biotype origin and on the genomic regions expressed in GTEx tissues and mTECs. X-axis indicates the biotype reported in the original study (groups). For clarity, BamQuery biotypes were summarized into five general categories: protein coding regions, non-coding RNAs, EREs, intronic and intergenic.

c, Percentage of the most likely biotype attributed by BamQuery to canonical, intronic and ncRNA MAPs.

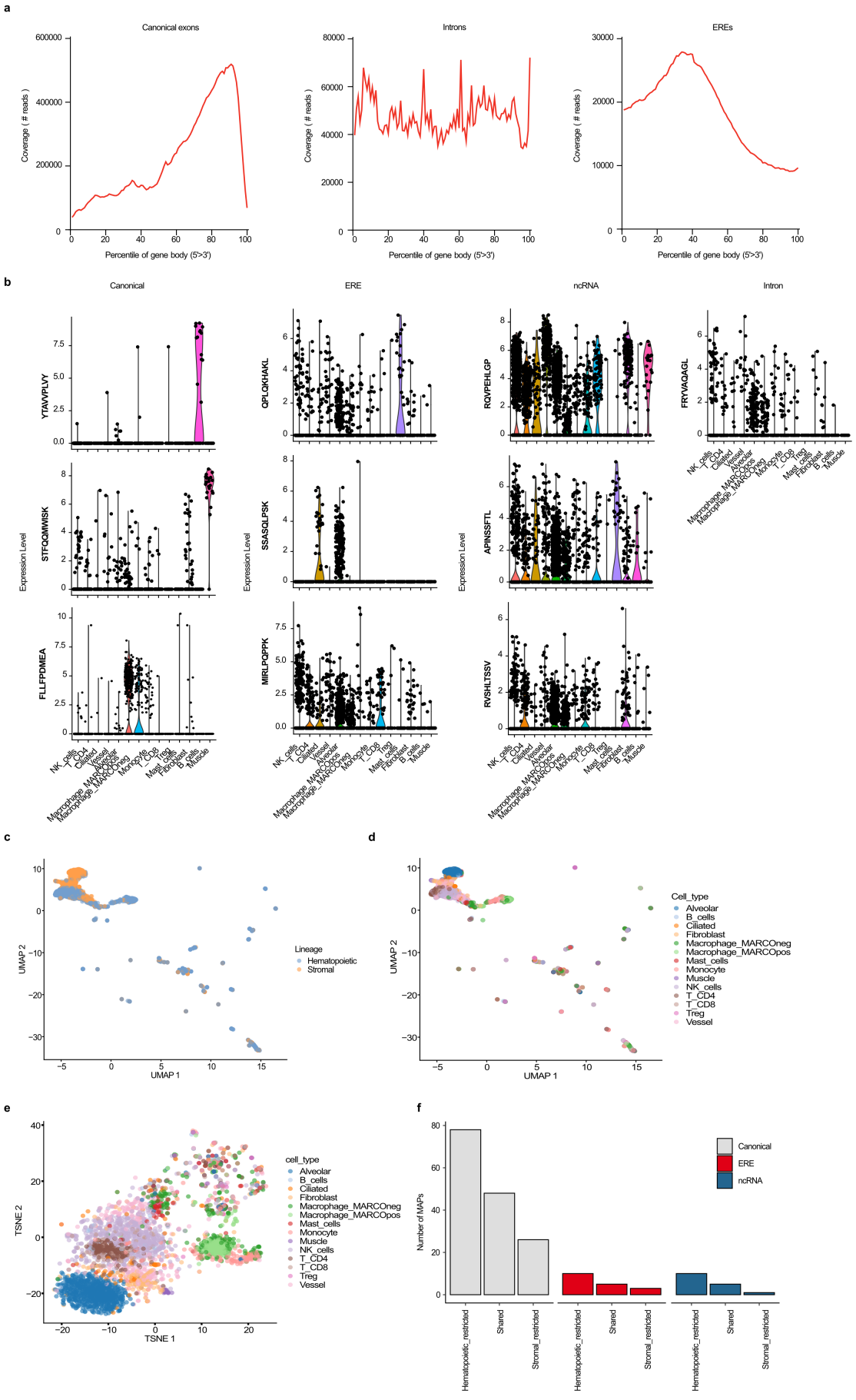

a
Canonical exons
Introns
EREs
Coverage (# reads)
Percentile of gene body (5'>3')
b
Canonical
ERE
ncRNA
Intron
YTAVVPLVY
STFQQMWISK
FLLFPDMEA
QPLQKHAKL
SSASQLPSK
MIRLPQPPK
RQVPEHLGP
APINSSFTL
RVSHLTSSV
FRYVAQAGL
Expression Level
NK_cells
T_CD4
Ciliated
Vessel
Alveolar
Macrophage_MARCOneg
Macrophage_MARCOpos
Monocyte
T_CD8
Treg
Mast_cells
Fibroblast
B_cells
Muscle
c
UMAP 1
UMAP 2
Lineage
Hematopoietic
Stromal
d
Cell_type
Alveolar
B_cells
Ciliated
Fibroblast
Macrophage_MARCOneg
Macrophage_MARCOpos
Mast_cells
Monocyte
Muscle
NK_cells
T_CD4
T_CD8
Treg
Vessel
e
TSNE 1
TSNE 2
cell_type
f
Number of MAPs
Canonical
ERE
ncRNA
Hematopoietic_restricted
Shared
Stromal_restricted

Figure 9. – BamQuery analysis of normal and cancer lung single cell datasets.

a, Number of lung scRNA-seq reads covering canonical genes, Intronic regions and EREs.

b, Expression of canonical, ERE, ncRNA, or intronic MAPs identified as differentially expressed in the normal lung dataset.

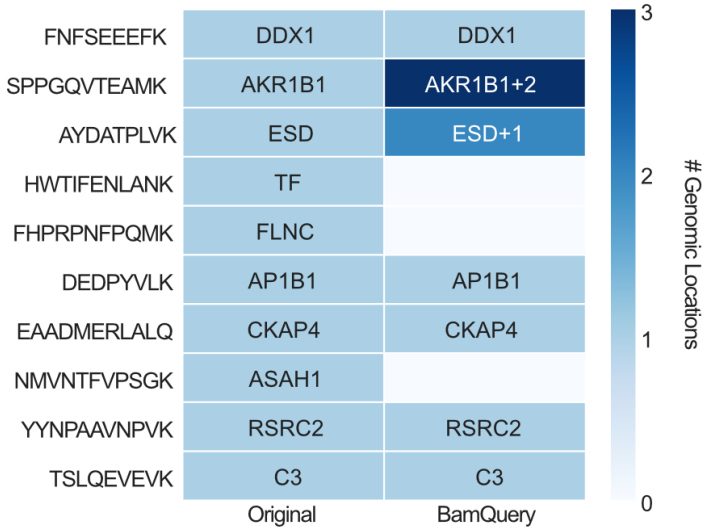
c, UMAP depicting the clustering of the hematopoietic and stromal cells from the normal lung based on their MAP expression.

d, UMAP showing the clustering of the cell populations from the normal lung based on their MAP expression.

e, TSNE showing the clustering of the cell populations from the normal lung based on their MAP expression.

f, Number of canonical, ncRNA, or ERE MAPs identified by the differential expression analysis as restricted to the hematopoietic or stromal compartments or shared by cells of both lineages.

a


FNFSEEEFK
SPPGQVTEAMK
AYDATPLVK
HWTIFENLANK
FHPRPNFPQMK
DEDPYVLK
EAADMERLALQ
NMVNTFVPSGK
YYNPAAVNPVK
TSLQEVEVK
DDX1
AKR1B1
ESD
TF
FLNC
AP1B1
CKAP4
ASAH1
RSRC2
C3
DDX1
AKR1B1+2
ESD+1
AP1B1
CKAP4
RSRC2
C3
Original
BamQuery
3
2
1
0
Genomic Locations

b

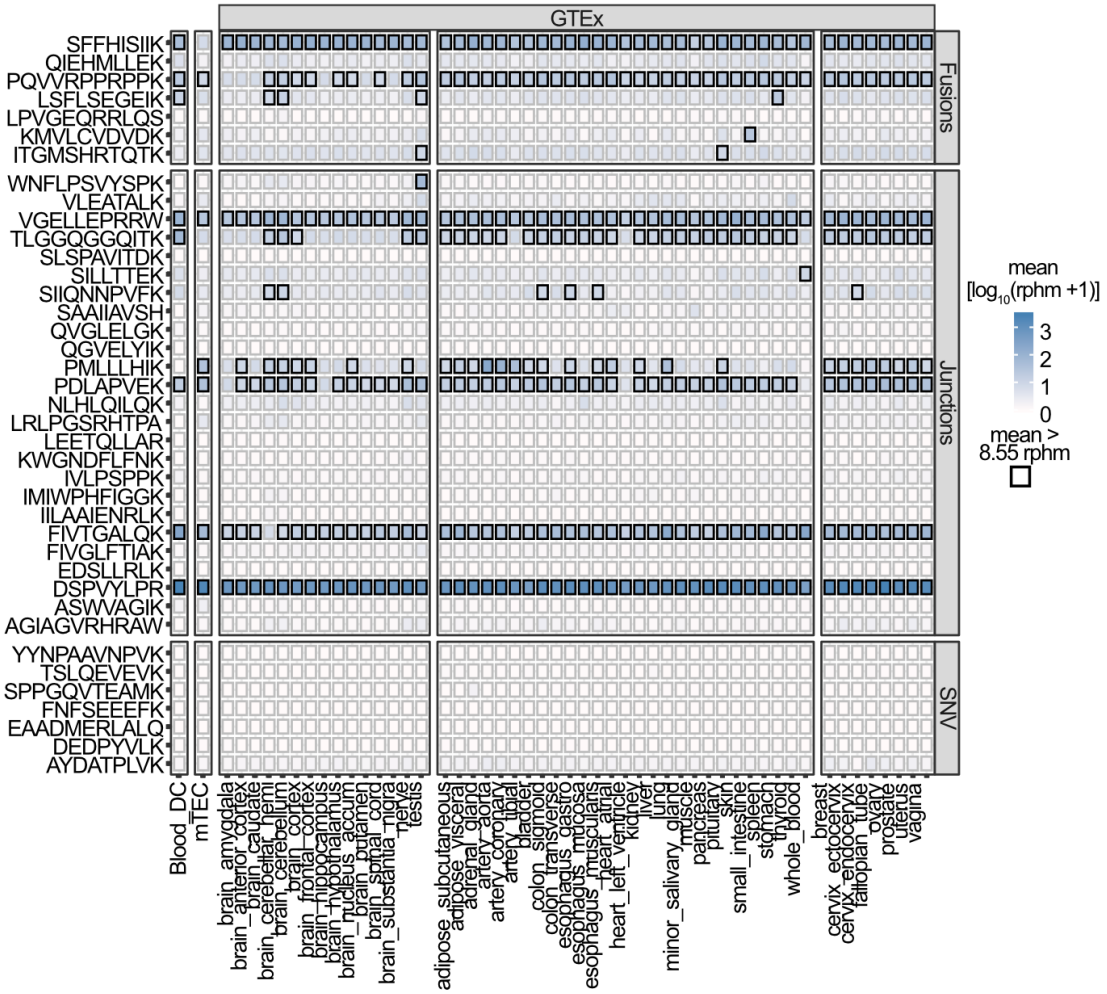

GTEx
Fusions
Junctions
SNV
SFFHISIIK
QIEHMLLEK
PQVVRPPRPPK
LSFLSEGEIK
LPVGEQRRLQS
KMVLCVDVDK
ITGMSHRTQTK
WNFLPSVYSPK
VLEATALK
VGELLEPRRW
TLGGQGGQITK
SLSPAVITDK
SILLTTEK
SIIQNNPVFK
SAAIIAVSH
QVGLELGK
QGVELYIK
PMLLLHIK
PDLAPVEK
NLHLQILQK
LRLPGSRHTPA
LEETQLLAR
KWGNDFLFNK
IVLPSPPK
IMIWPHFIGGK
IILAAIENRLK
FIVTGALQK
FIVGLFTIAK
EDSLLRLK
DSPVYLPR
ASWVAGIK
AGIAGVRHRAW
YYNPAAVNPVK
TSLQEVEVK
SPPGQVTEAMK
FNFSEEEFK
EAADMERLALQ
DEDPYVLK
AYDATPLVK
mean
[log10(rphm +1)]
3
2
1
0
mean >
8.55 rphm
Blood_DC
mTEC
brain_amygdala
brain_anterior_cortex
brain_caudate
brain_cerebellar_hemi
brain_cerebellum
brain_cortex
brain_frontal_cortex
brain_hippocampus
brain_hypothalamus
brain_nucleus_accum
brain_putamen
brain_spinal_cord
brain_substantia_nigra
testis
adipose_subcutaneous
adipose_visceral
adrenal_gland
artery_aorta
artery_coronary
artery_tibial
bladder
colon_sigmoid
colon_transverse
esophagus_gastro
esophagus_mucosa
esophagus_muscularis
heart_atrial
heart_left_ventricle
kidney
liver
lung
minor_salivary_gland
muscle
pancreas
pituitary
skin
small_intestine
spleen
stomach
thyroid
whole_blood
breast
cervix_ectocervix
cervix_endocervix
fallopian_tube
ovary
prostate
uterus
vagina

Figure 10. – BamQuery elucidates safer immunotherapeutic targets.
a, Heatmap of number of genomic locations at which the expression of the SNVs-derived TAs was assessed by BamQuery vs by the original study.
b, Heatmap of average RNA expression of published fusions, junctions, and SNVs-derived TAs in indicated tissues. Boxes in which a peptide has an average rphm>8.55 are highlighted in black.

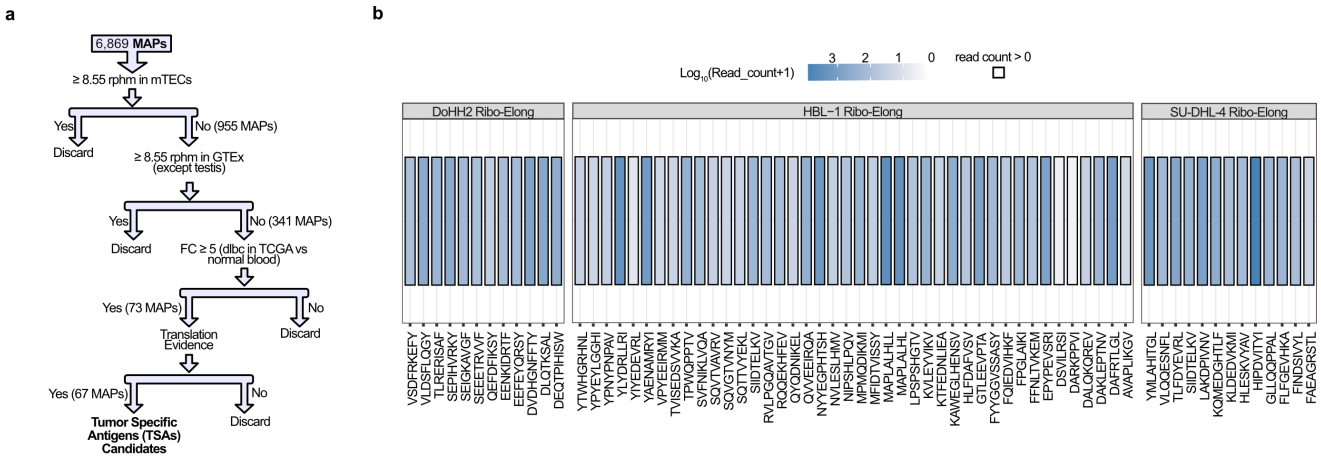


Figure 11. – Discrimination of potential immunotherapeutic targets in DLBCL.

a, Decision tree to discriminate TSAs from DLBCL.

b, Heatmap of average BamQuery-acquired read count of the 67 TSA candidates in indicated samples. Boxes in which a peptide has a rphm>8.55 are highlighted in black.

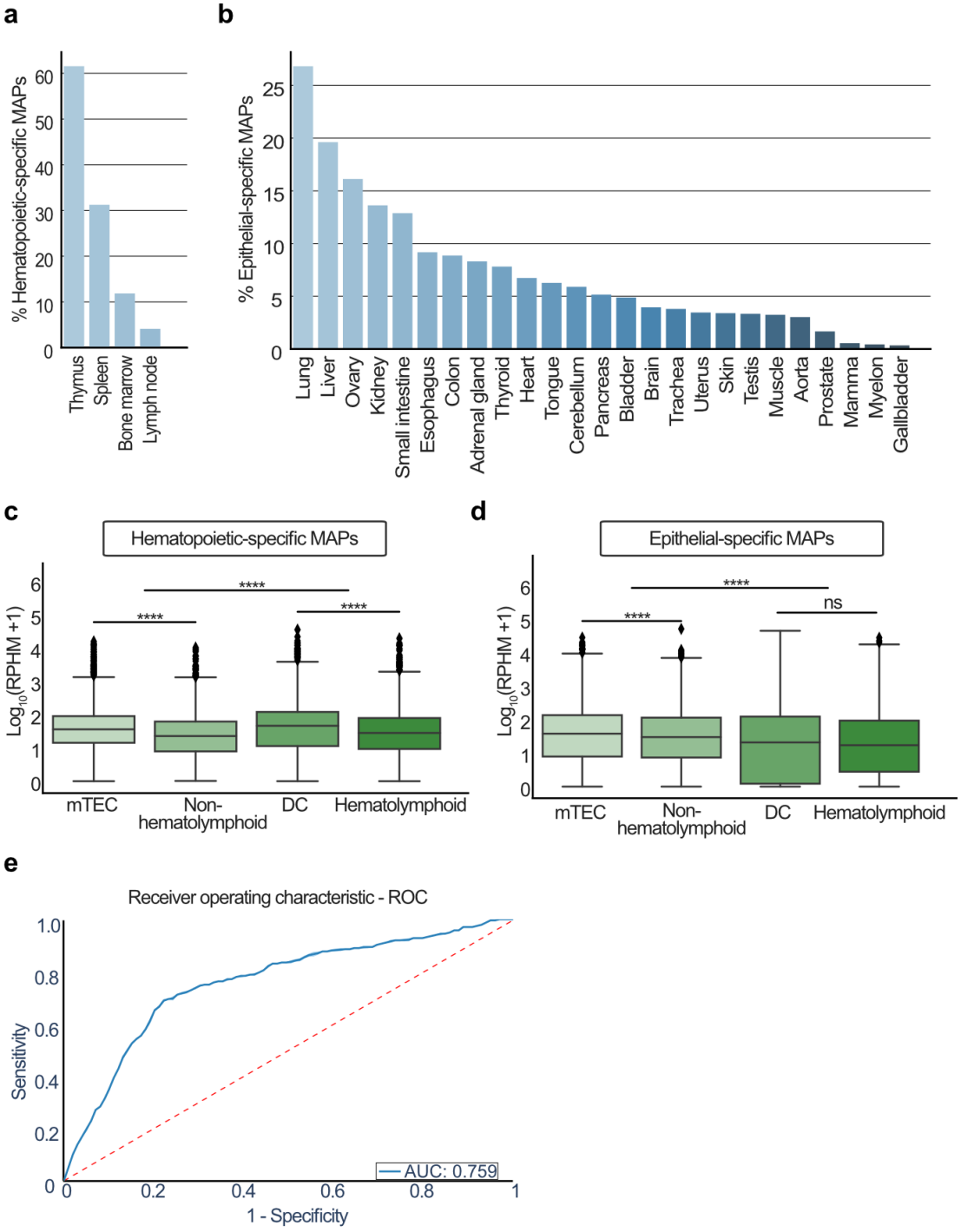


Figure 12. – mTECs and Blood_DC for TAs proritization.

a-b, Percentage of hematopoietic-specific MAPs (n=2,429) (a) and epithelial-specific (n=3,237) (d) presented by the indicated tissues.

c-d, Average RNA expression of hematopoietic-specific (c) and epithelial-specific (d) MAPs in mTECs (n = 11), non-hematolymphoid GTEx tissues (n = 2,389), DCs (n =19) and hematolymphoid GTEx tissues (n=196). Wilcoxon rank-sum test two-sided was used for comparisons (****p<0.0001).

e, Receiver operating characteristic curve (ROC) for prediction of immunogenic based on RNA expression (RPHM) of mTEC and DC samples. AUC== ~0.75 with a 95% confidence interval (CI): 0.7588 - 0.7591.

a

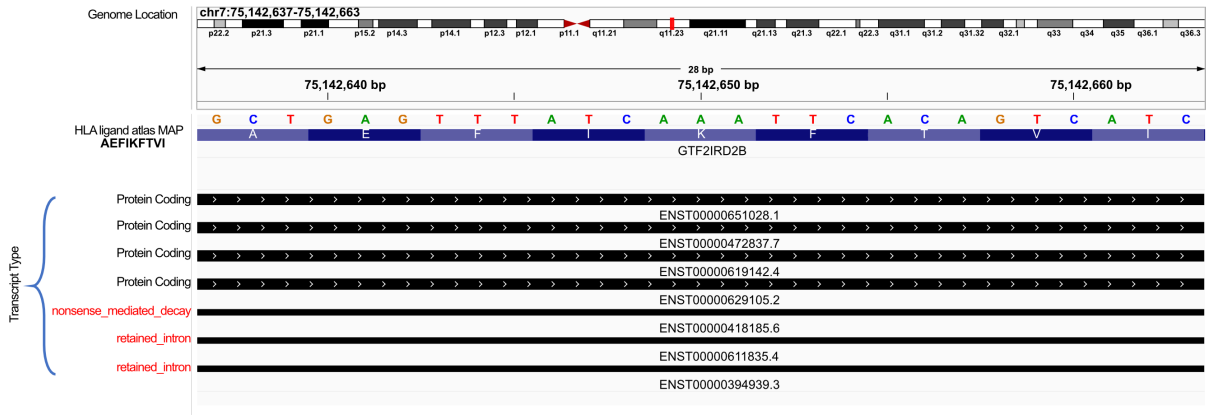


Figure 13. – Different biotypes overlap at the same genomic location.

a, The AEFIKFTVI peptide (HLA ligand atlas) at the indicated genomic location overlaps with protein-coding and non-coding RNAs transcripts.

1.29 Supplementary tables

All Supplementary Tables are available online in '.xlsx' format:

(https://www.biorxiv.org/content/10.1101/2022.10.07.510944v1.supplementary-material)

The full list is provided below:

- Supplementary Table 1[supplements/510944_file04.xlsx]
- Supplementary Table 2[supplements/510944_file05.csv]
- Supplementary Table 3[supplements/510944_file06.txt]
- Supplementary Table 4[supplements/510944_file07.txt]
- Supplementary Table 5[supplements/510944_file08.txt]
- Supplementary Table 6[supplements/510944_file09.xlsx]
- Supplementary Table 7[supplements/510944_file10.csv]
- Supplementary Table 8[supplements/510944_file11.xlsx]
- Supplementary Table 9[supplements/510944_file12.xlsx]

Chapter 4 – Discussion

The purpose of this thesis was to investigate the extent to which non-canonical proteins contribute to the repertoire of MAPs, and to determine the potential usefulness of the non-canonical MAPs as targets in cancer vaccines. In **Chapter 2**, a proteogenomic approach was developed and used to examine the contribution of non-canonical proteins to the immunopeptidome and proteome in diffuse large B-cell lymphoma (DLBCL). The main results of this study showed that most non-canonical MAP source proteins were found only in the immunopeptidome and not in the proteome, indicating their preferential access to the MHC class I pathway. In **Chapter 3**, we introduced BamQuery, a computational tool that attributes comprehensive RNA expression to each MAP candidate and allows comparing their expression in healthy and cancerous tissues to define potential tumor antigen (TA). Using BamQuery, we discovered that some MAPs previously reported as TAs might have been misclassified as such and thus pose a threat to healthy tissues. In addition, facilitated by BamQuery we found promising highly shared TAs among DLBCL patients.

Overall, our work has contributed to broadening our understanding of the biogenesis of non-canonical proteins and their involvement in the immunopeptidome and the entire proteome. We provided the research community with an easy-to-use tool that facilitates the identification of safe and actionable TAs for cancer vaccine design. In the next sections, we will highlight the strengths and weaknesses of our two approaches and discuss improvements that could be made to both strategies.

1.29.1Ribo-db approach to identify non-canonical translation products

In **Chapter 2**, we presented a novel proteogenomic approach called Ribo-db to analyze the immunopeptidome of diffuse large B-cell lymphoma (DLBCL). Considering that MS relies on searching for the best theoretical spectra matching the experimental spectra of the sequenced peptides in protein reference database; we set out to build comprehensive databases containing canonical and non-canonical translations. Our approach used a combination of ribosome profiling (Ribo-seq), RNA sequencing (RNA-seq) and mass spectrometry (MS) to identify and characterize

proteins. We use Ribo-seq data to provide an accurate view of the active translation events and to give context to the transcription observed through RNA-seq. Briefly, for each cell line, we collected Ribo-seq translation initiation sites (TIS) and elongation sites. The latter, refers to the ribosome process of adding new amino acids to a growing polypeptide chain during protein synthesis. Using the TIS data, we identified the positions in the genome that were most likely to be true start codons. Using StringTie[1] in the BAM files of the aligned Ribo-seq and RNA-seq elongation reads, we generated the full-transcriptome assembly that allows the identification of known and novel transcripts. Next, we intersected the positive start codons with the full-transcriptome assembly to detect those transcripts that have active translations. Assembled transcripts representing annotated transcripts that crossed with positive start codons representing annotated start codons were considered canonical, otherwise non-canonical. Finally, high-quality single nucleotide polymorphisms (SNPs) identified from the RNA-seq data were integrated into the retained assembled transcripts that were then translated *in silico* to build the custom protein databases. Thus, Ribo-db approach was designed to circumvent the risk of false identifications produced by the limitation associated with the size of the database when MS identification is performed[2]. Ribo-db databases should contain only transcripts susceptible to be truly translated which avoids the population of irrelevant sequences as in databases built from 3 or 6-frame translations of RNA-seq data[3].

By combining Ribo-seq elongation and RNA-seq reads to capture the complete transcriptome allowed the inclusion of annotated and unannotated transcripts that may have otherwise been missed. This is because Ribo-seq elongation fragments are not restricted to polyadenylated transcripts, as poly(A) RNA-seq sequencing is. Instead, Ribo-seq elongation allows the non-polyadenylated transcripts detection that can be the source of putative non-coding RNAs[4] and therefore non-canonical MAPs[5, 6]. By including these transcripts, our approach provided a more comprehensive inclusion of the actual translations than using only RNA-seq data to build the custom protein database.

Processing of Ribo-seq fragments present some challenges that may prevent the correct identification of an open reading frame (ORF). These challenges include the short length of the ribosome-protected fragments which leads to multiple mapping positions in the genome; the

difficulty in differentiating ORFs when they overlap; and the absence of ribosome protected fragments along the entire transcript. A previously published method called PRICE[7] addresses some of these issues, but it also excludes ORFs that may be affected by these limitations resulting in incomplete databases. In contrast, our approach applied a more conservative strategy for start codon identification by considering multimap reads and by considering ORFs assembled from both Ribo-seq elongation and RNA-seq data. The latter, for instance, allowed to include full-length transcripts that may otherwise go undetected if only Ribo-seq elongation was used, as Price does. Thus, Ribo-db, not only identified 99.7% of the MAPs identified with PRICE, but also identified an additional 5-6% of MAPs that would have been missed using PRICE alone (Figure 1C, Chapter 2).

Although Ribo-db has several advantages for the identification of non-canonical proteins, it relies on several external tools such as: STAR[8], StringTie[1], freeBayes (arXiv:1207.3907), BedTools[9], which can impact the speed and simplicity of the approach. The main drawback of our approach, which should be addressed as a priority, is that Ribo-db does not currently accurately capture complex rearrangements, including chromosomal and transcriptional fusions. Current mappers present difficulty in mapping RNA-seq reads bearing such variations as they use reference genomes to guide the mapping process which are devoid of the tumoral somatic changes. Thus, such RNA-seq reads may remain unaligned to the genome[10]. To include these rearrangements, an effective strategy would be to first predict the rearrangements using appropriate tools based on RNA-seq and Ribo-seq[11] properties, and then add their translations into the custom databases.

Overall, from the combination of Ribo-seq and RNA-seq we prepared customized databases that included only translated sequences from the 3 DLBCL cell lines. The use of these databases facilitated the identification by MS of canonical and non-canonical proteins in the whole proteome and immunopeptidome.

1.29.2Identification of non-canonical proteins in DLBCL cell lines

Using Ribo-db, we primarily analyzed the immunopeptidome of DLBCL cell lines and identified 4,944 protein sources of MAPs. Of these, 451 (9%) were classified as non-canonical proteins, while the remaining were considered canonical proteins. Among the non-canonical proteins, half were

further classified as cryptic proteins, while the other half was considered novel isoforms. The novel isoforms resulted from ORFs translated in the same frame as a canonical protein annotated in the GENCODE database[12] (Figure 2B, Chapter 2). However, these ORFs mostly initiated from non-AUG codons up or downstream of the annotated start site of the parental canonical protein (Figure 3A, Chapter 2), extending previous observations that under stress conditions initiation favor non-AUG start codons[13]. Furthermore, the translation of these novel isoforms demonstrates the ability of the translation process to generate new proteins from a transcript, highlighting the polycistronic nature of human protein-coding genes[14].

Regarding cryptic proteins, these were derived from annotated non-coding transcripts, 5'UTRs, 3'UTRs, intergenic regions, introns, and frameshifts of canonical genes. These were found to have distinctive properties compared to canonical proteins, extending previous findings[15-18] (Figure 2F, Chapter 2). For instance, cryptic MAPs source proteins derived mostly from non-AUG codons, which is associated with response to stressful conditions (Figure 3A, Chapter 2)[13]. They were also much shorter, as most were composed of two exons compared to the average number of 11 exons of canonical MAP source proteins (Figure 3C, Chapter 2); and had slightly but significantly lower expression compared to canonical MAP source proteins. However, consistent with previous reports[19, 20], the transcriptional expression of such proteins was higher than that of the non-MAP-generating transcripts.

To further investigate the biogenesis of cryptic proteins and aware of their short size, we used our Ribo-db approach to identify proteins in the whole proteome extracts of the three DLBCL lines. Aware that MS favors the detection of the most abundant proteins[21], we separated the short and larger proteins based on their molecular weight (low-molecular-weight proteins $\leq$ 10 kDa > high-molecular-weight). A salient observation from this analysis was that the cryptic proteome identified in whole proteome extracts was distinct from the cryptic proteome identified as a source of MAPs. Only 6% of the cryptic MAPs source proteome was also identified in the whole proteome. This unexpected finding suggested the existence of two distinct cryptic protein repertoires: one populating the whole proteome and another populating the immunopeptidome (Figure 5A, Chapter 2). The detection of two different cryptic proteomes was possible since MAPs have long lifetime (12 h)[22, 23] so their identification was possible even after their parent protein

had possibly undergone rapid degradation. To confirm this hypothesis of such MAPs deriving from rapidly degraded proteins (RDPs), we evaluated whether their source proteins were short-lived in the cytoplasm preventing their identification in the whole proteome.

The stable conformation assessment of cryptic MAP source proteins predicted that these proteins were short-lived in the cytoplasm. Indeed, they turned out to be less stable in vivo and highly disordered, meaning that they may not reach a stable conformation (Figure 5F, G, Chapter 2). Additionally, despite their small size, their rapid turnover predicted that cryptic proteins were ~5-fold more efficient compared to canonical proteins in generating MAPs (Figure 3F, Chapter 2). Their rapid degradation and ability to generate MAPs, supported the hypothesis that these MAPs may have originated from RDPs. Therefore, they could be considered as a prototype of DRiPs, explaining why most of them were not seen in whole proteome extracts. These results further demonstrate that immunopeptidome studies can be used as a sink for peptide identification and, consequently, proteins that would otherwise be invisible in conventional whole proteome MS studies.

While our approach has demonstrated the existence of two distinct non-canonical protein repertoires, there is still potential venues to further explore the immunopeptidome and proteome using diverse MS approaches. First, to improve the detection of non-canonical proteins, a more efficient approach could involve data independent acquisition (DIA) to acquire MS data. This strategy would provide a broader view of the immunopeptidome and the entire proteome, although analysis of the data may be more complex. However, recent *de novo* sequencing algorithms can be leveraged for the deconvolution of MS2 data to identify peptides[24]. Furthermore, ribosomal profiling data could then be investigated to validate the translation of such peptide identifications. Second, as cryptic MAPs source proteins were expected to have few tryptic sites (Figure 5D, Chapter 2), the use of trypsin led to the collection of fewer fragments per protein while the MS whole proteome analysis was performed. A significant improvement to improve the identification of smaller proteins in whole proteome MS analysis could be to explore the use of alternative proteases to complement trypsin digestion. Proteases such as Glu-C, LysN, Lys-C, Asp-N or chymotrypsin target different specific sites on proteins[25], so the frequency of such amino acids in non-canonical proteins could be evaluated first to identify the most suitable

alternative. Third, employing diverse techniques for ionization and separation, alongside adjusting gradient time for separation, could be beneficial to maximize peptide identification. Indeed, different settings would facilitate the identification and analysis of distinct subsets of peptides based on their physical and chemical properties, thereby revealing a more comprehensive view of the proteome and immunopeptidome. For instance, performing polarity switching during the ionization process in MS can improve the signal intensity by considering both positive and negative ions, which may result in a higher yield of charged molecules[26]. Also, considering the use of different separation ions as gas chromatography (GC), or capillary electrophoresis (CE) and use of longer gradients for the separation process could improve the resolution and sensitivity of the analysis, allowing for the detection and identification of a greater number of peptides. In this regard, high field asymmetric waveform ion mobility spectrometry (FAIMS) for gas phase separation has already demonstrated to increase peptide and protein identifications[27, 28]. For instance, FAIMS has demonstrated the ability to identify 50% more MAPs when used as a front-end separation technique in MS[29]. Fourth, it should be noted that fragmentation methods have been optimized for tryptic peptides. However, due to the endogenous nature of MAPs, fragmentation can result in noisy scans with internal ion series and neutral losses[30]. In our study, we used higher energy collisional dissociation (HCD) to fragment MAPs, which yields good fragmentation quality for MAPs[30]. Yet, a dual fragmentation approach using both electron transfer dissociation (ETD) (c/z) and higher-energy collisional dissociation (HCD) (b/y) in a single scan has been shown to produce more informative scans for MAPs[31]. This approach is expected to allow a highly reliable identification of the peptide sequence, thus facilitating the localization of PTMs. Therefore, a comparative analysis between fragmentation methods would be beneficial to identify the one that maximizes MAPs identification. Fifth, in our analysis we missed the identification of PTM-bearing MAPs as we considered very few post-translational modifications (oxidation (M) and deamidation (NQ)). However, cancer cells are characterized by the deregulation of cellular processes that can induce modifications in proteins such as phosphorylation. As a result, the degradation of these proteins can generate highly immunogenic MAPs[32]. Indeed, Zarling et al. employed an MS-based approach to identify several phosphopeptides associated with cytoplasmic signaling and cellular transformation pathways,

indicating a potential association between cellular function and their display on tumor cells[32]. To expand the identification of PTMs, current software can be parametrized to detect them through database searches but also through scan analysis using de novo sequencing (PEAKS PTM)[33]. However, because PTM-bearing MAPs, such as phosphorylated peptides, only constitute a small fraction of the immunopeptidome (~1%)[34], the use of enrichment strategies could greatly facilitate their isolation and identification[35]. This, in turn, can provide a broader view of both canonical and non-canonical modified proteins in the samples. Lastly, the recently published Multi-Omic Native Tissue Enrichment (MONTE) workflow could be employed to facilitate serial multiomics analysis of tissue samples to provide a more complete understanding of the non-canonical proteome[36]. As a proof of concept, MONTE enabled the identification of non-canonical proteins by analyzing the immunopeptidome (MHC-I and MHC-II), ubiquitylome, proteome, phosphoproteome, and acetylome from the same tissue sample of primary patient lung adenocarcinoma (LUAD) tumors. Thus, this innovative approach offers a significant improvement in the detection of non-canonical proteins from various angles, providing new insights into disease pathology and potential treatment strategies. Overall, the application of any of these approaches, whether used individually or in combination, has great potential to significantly improve the sensitivity of immunopeptidome and proteome identification. Particularly concerning the immunopeptidome, these methods offer a promising solution to address the limited number of peptides identified in our analysis. Specifically, in the case of the HBL1 cell line, the number of identified MAPs($3.2x10^3$ MAPs, Figure 6D,E, Chapter 2) was found to be below the expected count of distinct MAPs which is estimated to be around ~$1x10^4$ for cells containing approximately ~$2x10^5$ MHC I molecules[37]. Consequently, the identification of a significantly larger number of MAPs would complement and enrich our current findings.

Together, cryptic proteins make up a significant portion of the immunopeptidome and cytoplasmic proteome in DLBCL cell lines, at approximately 5% and 13%, respectively. Our findings, as well as those of others, suggest that these proteins are not simply the result of translational noise[38, 39]. Instead, non-canonical proteins may constitute a specialized repertoire for communicating cellular state to the immune system in addition to acting in metabolism and cellular regulation functions previously observed[38]. Until now, most proteogenomic studies on

the non-canonical proteome have been conducted in cancer cells[17, 40-42] to perform identification of actionable targets for immunotherapy rather than to investigate their characterization. Therefore, it would be valuable to extend the results observed here in cancer cells to normal cell lines. The Ribo-db approach could be used to perform the same analysis that was carried out in this study of cancer cells, but this time with normal human B-lymphoblastoid cell lines (B-LCLs) for example. This would allow to answer questions such as: What is the overlap between the cryptic proteome identified in normal and that of cancer cells? Is the translation of cancer-specific cryptic proteins the result of the neoplastic transformation? Is the observation of two repertoires of non-canonical proteins unique to cancer cells? Does the contribution of canonical proteins depend on the cell type of origin? A previous study has already analyzed the cryptic MAP repertoire of B-LCLs[15] and reported that cryptic proteins make up approximately 10% of the immunopeptidome. However, the database used in this study for MAP identification was flooded with irrelevant sequences that may have affected the accuracy of the identifications. In fact, the database was built from 6-frame translations of the samples RNA-seq reads and a very high false discovery rate (FDR) threshold of 5% was used, which is normally advised to be kept at 1%[43]. Therefore, it may be necessary to revisit the identifications and observations made in normal cells using Ribo-db to properly compare them to the non-canonical proteome detected in cancer.

1.29.3Non-canonical proteins result and origin of the DLBCL oncogenic program

The large number of non-canonical proteins detected in the immunopeptidome and whole proteome in DLBCL cells (2,503 proteins), suggests that the oncogenic program inherent to DLBCL may alter gene expression thereby increasing non-canonical translations. Our results showed that non-canonical proteins were translated from all chromosomes, but chromosomes 12 and 16 appeared to be particularly rich sources of novel isoform and cryptic proteins, respectively (Figure 6A, Figure 10B, Chapter 2). Both of these chromosomes have been documented to be involved in cytogenetic abnormalities in DLBCL[44, 45]. In addition to potentially being the result of genetic alterations, non-canonical proteins may also contribute to perturbations in important signaling pathways causing further genetic alterations. Specifically, novel isoforms may alter the translation balance of canonical proteins from the genes that produce them. Some of these genes being involved in signaling pathways such as NOTCH (Figure 6C, Chapter 2), which is often deregulated

in cancer[46] and known to regulate cell proliferation, fate, differentiation and death[47]. Similarly, cryptic 5'UTR proteins may inhibit the translation of canonical proteins from genes involved in transcription, translation, and antiviral response signaling pathways (Figure 6E, Chapter 2), extending previous reports[48]. These findings indicate that the non-canonical proteome in DLBCL appeared to be involved in cell growth and stress response, potentially contributing to impaired DNA repair and cancer progression. Given these important findings, an important next step would be to determine the precise function of cryptic proteins in cancer cells and their impact in the mentioned signaling pathways. One approach to address these questions would be to consider the ORFs of the detected cryptic proteins to design RNA guides for CRISPR/Cas9 knockout libraries, thereby assessing the impact on the phenotype of DLBCL cell lines. To assess the phenotype in a CRISPR/Cas9 knockout library, it would be necessary to perform a variety of experiments to measure how the loss of the specific cryptic protein gene affects the function or behavior of the cells. For example, it could be desirable to measure changes in gene expression, protein levels, cell growth or survival, or the ability of the cells to respond to stimuli. By comparing the phenotype of the specific cryptic protein gene knocked out cells to those with an intact copy of the gene; it would be possible to determine the function of the gene and its role in the biological processes previously described.

Given that non-canonical proteins appeared to be the result of the DLBCL oncogenic program and that most of the cryptic MAP source proteins were observed exclusively in the immunopeptidome, it was next important to consider whether such MAPs could be suitable targets for the development of vaccines. Previous research have shown that non-canonical MAPs are a major source of targetable TSA[42] and that RNA expression could be an indicator of the MAP presentation likelihood[49]. With this in mind, we decided to delve deeper into how to facilitate the validation and prioritization of candidate MAPs for TAs by estimating the probability of MAP presentation using RNA-seq data. Thus, we developed BamQuery, a tool presented in **Chapter 3**, with the objective of standardizing the prioritization of MAPs based on their expression in healthy versus cancerous tissues.

1.29.4BamQuery : Exhaustive capture of MAPs RNA expression

BamQuery is a tool that helps to assess whether a MAP is highly expressed in cancer compared to healthy tissues, using its RNA expression to validate it as a therapeutic target (TA) and predict its immunogenicity. BamQuery operates on previously identified MAPs and BAM files corresponding to samples, both of which must be provided in form of lists. The MAP list provides the amino acid sequence of candidate TAs for evaluation of their RNA expression in the BAM files listed. These BAM files could include normal and cancer samples from different sources, such as GTEx for normal tissues, TCGA for cancer tissues and own sources.

BamQuery works in five main steps to facilitate the assessment of MAP expression. First, it collects the potential MAP coding sequences (MCS) of each MAP by performing reverse translation of the amino acid sequence. Second, it maps the MCS to the genome using the STAR aligner[8] and collects all MCS locations mapped to the genome for each MAP. Third, it counts the total RNA-seq reads for each MAP by examining the overlapping reads at each MCS location in the BAM files provided by the user. Fourth, the sum of the total RNA-seq reads overlapping the MCS location is normalized by the total RNA-seq sample read count (primary read alignment of the BAM file reads). Finally, the biotype of each MAP is calculated as a function of all expressed locations and the number of reads at each location. As a result, BamQuery provides the RNA expression of each MAP for each BAM file and a MAP biotype classification based on the MAP's genomic locations in relation to reference annotations.

One key feature of BamQuery is the assignment of exhaustive RNA expression to any MAP. Given the degeneracy of the genetic code, BamQuery was designed to assess the exhaustive RNA expression of any MAP by examining the local expression of all potential genomic locations of MAP coding sequences (MCS). Thus, RNA expression could be used as an indicator of MAP presentation, rather than relying on MS quantification of the normal tissues immunopeptidome after immunoprecipitation of MHC-I complex[50]. To evaluate the accuracy of BamQuery's quantification, we compared the RNA-seq expression of 1,211 canonical MAPs from the HLA Ligand Atlas[51] in 8 mTECs, as these cells express most of the canonical genes for inducing central tolerance[52]. Using the Jellyfish[53] tool, we compared the exact occurrence of each MCS encoding any canonical MAP in mTECs' RNA-seq data. We observed a strong correlation between

BamQuery's total read count and Jellyfish occurrences, indicating the high accuracy of BamQuery (94% mean accuracy, Figure 1B and Figure 7D, Chapter 3). However, BamQuery assigned a read count of 0 to 4,194 MCS for which Jellyfish reported a mean number of occurrences equal to 1 (data not shown). These discrepancies were attributed to possible sequencing errors and limitations of the STAR aligner to find all possible MCS locations in the genome. Although, BamQuery is very efficient at counting RNA-seq reads in a BAM file (step 3) with a speed of approximately 0.0005 minutes/MCS/location compared to the grep command (~1 minute/MCS/location); its speed and accuracy are limited by the STAR aligner. First, STAR may take quite some time to align the MCS, depending on the number of MAPs queried and their amino acid sequences, and second it may not align MCS to all possible genomic locations.

We have previously shown that MAPs derived from ERE sequences were enriched in amino acids encoded by 6 synonymous codons (Figure 2B-C, Chapter 3). The high number of synonymous codons in ERE-derived MAPs can lead to the generation of many MCS upon reverse translation, up to $6x10^7$, which can negatively impact the performance of aligner. To address this issue, we compared the performance of STAR to two other aligners (GSNAP and BBMap) in finding alignments for a set of both canonical and non-canonical MAPs, including ERE-derived MAPs. We found that STAR significantly outperformed the other two aligners in alignment time and the number of locations found per MAP, except for one ERE-derived MAP (data not shown). It has been previously reported that STAR can have performance issues when aligning sequences in complex areas of the genome[54], which may prevent the collection of the MAPs locations full set. To improve speed, we limited the length of each peptide sequence to 11 amino acids, which is the maximum length usually considered for MAPs, and recommend filtering uninteresting MAPs, such as those derived from canonical proteins. To improve EREs MCS alignment, one strategy that could be used, advised by the author of STAR, is to mask all but one copy of the exact repeats from the genome index used for mapping. This allows the MCS to be aligned with a single copy of each repeat, and from there the reconstruction of all MCS alignments can be done, since it is known which masked loci correspond to each unmasked repeat. Implementation of this strategy would be helpful to capture all possible MCSs in ERE regions.

However, we could completely avoid the limitations of the STAR aligner by using an aligner-free method to map all MCS locations in the genome. A potential solution is to implement the Aho-Corasick algorithm, a string search algorithm that can locate strings within an input text and output all possible matching locations. In our case, the strings would be the list of candidates MAPs (amino acid sequences) for which a keyword trie would be built. A trie is a data structure that stores the sequences and enables parallel searching in a text through the Aho-Corasick algorithm. The Aho-Corasick algorithm would be used to scan the 3-frame translations of the genome annotation, ERE sequences, and standardized ORFs from Ribo-seq data[55] in the search for the keyword trie. After collection of all possible locations for each MAP, MCS could be backtracked from such locations in the genome as they are key to detect and count only the RNA-seq reads from BAM files that fully span a given MCS at a given location. The use of this strategy would bring several benefits to our approach. First, the complexity of the Aho-Corasick algorithm is linear which makes it a fast solution. The linear search time is defined by O(n + m + z), where n is the input text length (n = sum of the length of the three 3-frame translations), m is the string lengths (m = sum of the MAPs amino acid sequence lengths) and z is the total MAPs occurrences in the text. Therefore, the current first and second steps in BamQuery that performs reverse translation and mapping with STAR, will be omitted because with this strategy we would be looking for pseudoalignments of the amino acid sequence directly. Second, it will allow the collection of all possible MAP locations, including those of ERE-derived MAPs, as the 3-frame translation of those regions would be considered in the input text. Third, it would be possible for BamQuery to query longer peptides, up to 25 amino acids, for which RNA-seq reads could overlap their entire MCS (RNA-seq reads are typically 75 base pairs long). For example, MHC class II molecules associated peptides usually present longer peptides (13-17 amino acids in length) than MHC class I molecules (8-11 amino acids in length). Or, tryptic peptides that can be up to 25 amino acids long. In fact, BamQuery was recently tuned and used to confirm the presence of specific mutations in tryptic peptides by evaluating them in colorectal cancer RNA-seq Bam files[56]. However, this analysis with BamQuery required preprocessing of the tryptic peptides by trimming them (up to 16 amino acids) taking care that the observed mutation was included in the final sequence.

Overall, BamQuery is a useful tool for collecting RNA expression data for short peptides in contexts other than the immunopeptidome. However, in the context of immunotherapy, BamQuery is particularly useful for carefully selecting TAs. While it is true that evaluating the RNA expression from all possible regions in the genome of a given MAP can be very stringent; we believe that the quality of the TAs is more important than the quantity when it comes to developing cancer treatments. Careful selection of TAs should prioritize those with a single genomic location and cancer-specific expression, to avoid undesirable effects. By using BamQuery in a rigorous manner, we can help to ensure that the TAs selected for immunotherapy development are of high quality and have the potential to be effective and safe treatments.

1.29.5Assessment of the RNA expression of canonical and non-canonical MAPs

In **Chapter 2**, we demonstrated the flexibility of the translation process, as proteins can be translated from non-coding regions of the genome and from frameshift variations of canonical proteins. To account for this, BamQuery aims to collect the RNA expression of each MAP by considering the multiple regions of the genome that could potentially translate the MAP. Therefore, the biotype classification of each MAP should also reflect the regions expressed in the samples according with their expression in RNA-seq reads terms for each location. With this in mind, we implemented the expectation maximization (EM) algorithm, which aims to estimate the parameters ϑ that maximize the log likelihood $logP(x;\vartheta)$ of the observed data[57]. By using EM, we could estimate the RNA-seq reads likelihood to be associated with each biotype (In frame, 3'UTR, 5'UTR, frameshift, etc.) from observations of both canonical and non-canonical MAP locations. This is especially useful for distributing the RNA-seq reads in a location where several biotypes overlap, as multiple ORFs may be annotated at the same location. The final biotype classification of a given MAP should reflect all biotypes that overlap with the MCS locations expressed in the samples. An example of this can be seen in Figure 13, chapter 3, where an MCS of the canonical peptide AEFIKFTVI from the HLA ligand atlas, is located on chromosome 7, position that overlaps protein-coding and non-coding RNA transcripts. The AEFIKFTVI biotype classification then was computed independently for the queried samples (GTEx and mTEC), but the biotype was also aggregated for all the samples and presented in the Total reads count RNA column (Figure 1).

Peptide Type	Peptide	GTEX	mTEC	Total reads count RNA
HLAligandAtlas	AEFIKFTVI	Non_coding Exons: 77.52% - In_frame: 22.48%	Non_coding Exons: 80.74% - In_frame: 19.26%	Non_coding Exons: 77.57% - In_frame: 22.43%

Figure 1. – AEFIKFTFTVI biotype as shown in the BamQuery biotype classification output

The final BamQuery biotype classification for AEFIKFTVI reports the transcript distribution in the queried samples (GTEx and mTEC) considering all locations and the number of overlapping reads at those locations.

The final AEFIKFTVI biotype classification showed that 77.57% of RNA-seq reads contributed to Non-coding exons while the 22.43% to In-frame biotype. This classification was consistent if we observed in detail the numerous RNA-seq reads found in certain genomic locations (chr7:73200604-73200630, chr7:73249892-73249918, chr7:75192049-75192075), where the biotype of the underlying transcripts is Non-coding exons. It was also in line with remaining regions where Non-coding exons and In-frame transcripts overlap in which the distribution of reads for such biotypes was affected by the parameters estimated with EM (Figure 2). This example illustrates the importance of considering all genomic locations when assigning MAP expression, rather than individual locations, to accurately assess the transcriptional capacity of a MAP. Therefore, by using BamQuery to classify both canonical and non-canonical MAPs (Supplementary Table 1[supplements/510944_file04.xlsx]), we observed that some of these MAPs can be encoded by a variety of genomic regions. Some of these locations presented elevated expression in normal samples, which may result in a different biotype classification than the one previously reported (Figure 2G, Chapter 3). For example, some canonical MAPs (26%) could be translated from putative non-coding transcripts, and in the case of ERE-derived MAPs, only 56% were estimated to be derived from ERE regions. Interestingly, for remaining ERE-derived MAPs they could be derived from canonical regions in particular from In-Frame translations (Figure 2H, Chapter 3). Thus, considering these observations, we concluded that MAPs can be encoded by numerous genomic regions which can present elevated expression leading to a different classification than those previously reported.

Peptide	Alignment	Strand	Transcript	gene_level_biotype	transcript_level_biotype	genomic_position_biotype	Total reads count RNA
AEFIKFTVI	chr21:23863126-23863152	-	No Annotation	Intergenic	Intergenic	Intergenic	0
	chr7:73200604-73200630	+	ENST00000453092.3	transcribed_processed_pseudogene	transcribed_processed_pseudogene	Non_coding Exons	245589
		+	ENST00000544802.5	transcribed_processed_pseudogene	processed_transcript	Non_coding Exons	245589
	chr7:73249892-73249918	-	ENST00000449689.2	transcribed_unprocessed_pseudogene	transcribed_unprocessed_pseudogene	Non_coding Exons	13570
		-	ENST00000618962.4	transcribed_unprocessed_pseudogene	retained_intron	Non_coding Exons	13570
		-	ENST00000620147.4	transcribed_unprocessed_pseudogene	processed_transcript	Non_coding Exons	13570
	chr7:74754001-74754027	+	ENST00000464471.1	protein_coding	retained_intron	Non_coding Exons	299014
		+	ENST00000482232.1	protein_coding	retained_intron	Non_coding Exons	299014
		+	ENST00000573035.5	protein_coding	protein_coding	In_frame	299014
		+	ENST00000614986.4	protein_coding	protein_coding	In_frame	299014
		+	ENST00000620879.4	protein_coding	protein_coding	In_frame	299014
		+	ENST00000621734.4	protein_coding	protein_coding	In_frame	299014
	chr7:74803300-74803326	-	ENST00000451013.6	protein_coding	protein_coding	In_frame	18381
		-	ENST00000610955.1	protein_coding	retained_intron	Non_coding Exons	18381
		-	ENST00000625377.2	protein_coding	protein_coding	In_frame	18381
	chr7:75142637-75142663	+	ENST00000394939.3	protein_coding	retained_intron	Non_coding Exons	50623
		+	ENST00000418185.6	protein_coding	nonsense_mediated_decay	Non_coding Exons	50623
		+	ENST00000472837.5	protein_coding	protein_coding	In_frame	50623
		+	ENST00000611835.4	protein_coding	retained_intron	Non_coding Exons	50623
		+	ENST00000619142.4	protein_coding	protein_coding	In_frame	50623
		+	ENST00000629105.2	protein_coding	protein_coding	In_frame	50623
	chr7:75192049-75192075	-	ENST00000614592.4	transcribed_unprocessed_pseudogene	transcribed_unprocessed_pseudogene	Non_coding Exons	348165
		-	ENST00000616377.4	transcribed_unprocessed_pseudogene	processed_transcript	Non_coding Exons	348165
		-	ENST00000622829.4	transcribed_unprocessed_pseudogene	processed_transcript	Non_coding Exons	348165

Figure 2. – Detailed RNA-seq reads count for all genomic locations of the AEFIKFTVI peptide

Breakdown of the RNA-seq reads total count underlying AEFIKFTVI MCS localizations. Biotype classification shows the biotype in the gene, the transcript and the position in the transcript level overlapping with the localization.

However, it is important to note that the EM algorithm was trained using a limited set of canonical and non-canonical MAPs, and their expression evaluation was based on RNA-seq data rather than direct translation assessment. Therefore, the attribution of biotypes to MAPs coded by overlapping regions should be considered as predictions. Achieving an unambiguous attribution of biotypes while considering all the genomic regions presents a significant challenge and would require a dedicated study to obtain more accurate results. To improve the accuracy of biotype attribution, potential avenues include retraining the EM algorithm using MAPs that have been evaluated based on Ribosome profiling data or exploring advanced methods such as neural networks. These advanced approaches can incorporate additional features such as codon usage of the aligned MCS region, expression on Ribosome profiling, codon periodicity, and RNA-seq expression. By considering a broader range of features, we could enhance the predictive capabilities and refine the biotype attribution process.

It is crucial to carefully evaluate the actual expression of the candidate antigen transcript to select the safest immunotherapeutic targets. Therefore, we examined the RNA expression of previously reported TAs in triple-negative breast cancer[58], medulloblastoma[59] and colorectal cancer (CRC)[60] to evaluate their probability of being presented by normal cells: GTEx, mTECs, and

sorted dendritic cells (DCs)[61, 62]. We found that some of these TAs could be encoded by a variety of genomic sources highly expressed in such healthy tissues, in addition to the sources initially reported. This suggests that such TAs highly expressed in healthy tissues are likely to be poor safe targets for immunotherapy. Thus, special caution should be exercised when considering them as they could be presented by MHC I molecules from healthy tissues.

To make it easier for researchers to use BamQuery, we have created an online portal (https://bamquery.iric.ca/search) allowing users to analyze MAPs in mTECs and Blood DCs[61, 62]; as these normal samples are often used as proxies for tumor specificity and immunogenicity. The portal can be used as a first step to filter out MAPs that are not of interest for immunotherapy as those that are highly expressed in mTECs or DCs, are probably also expressed in other normal tissues. Furthermore, users can also download and install BamQuery (https://bamquery.iric.ca/documentation/installation.html) for use a standalone version allowing them to investigate the MAPs expression in GTEx, TCGA, or their own RNA-seq samples.

Overall, BamQuery is a powerful tool that provides extensive information about the transcriptional landscape of TA candidates, enabling users to make informed decisions. This may include conducting immunogenicity assays or systematic screening in tumor-organoid models. In recent years, three-dimensional (3D) in vitro cell culture models that recapitulate some of the characteristics of the original tumor tissue emerged as a promising tool for testing novel potential therapeutic applications[63]. Tumor organoids offer a remarkable representation of the complex diversity and physical architecture found in actual tumor tissues. This feature is crucial for the assessment of potential drug efficacy and renders them as promising alternative models for the screening of immunotherapeutic drugs[64]. Hence, the more realistic mimicry of the tumor microenvironment makes tumor organoids a more relevant and reliable model for studying human diseases compared to traditional animal models or cell lines. Because human model systems can provide accurate predictions of how patients will respond to a given treatment, large pharmaceutical companies such as Roche have begun to embrace and invest in these technologies. In fact, very recently (May 2023), Roche announced the creation of an organoid research institute called the Institute for Human Biology (IHB) in Basel, Switzerland. With interesting TA candidates identified using BamQuery, researchers can then use 3D in vitro models

to confirm their specificity and evaluate the viability of T cells in targeting such antigens[65]. By using tumor organoids, researchers can better assess the potential efficacy and safety of immunotherapeutic targets in a controlled and representative microenvironment.

1.29.6Further characterization of canonical and non-canonical MAPs

Using BamQuery, we analyzed previously reported 9-mers canonical and non-canonical MAPs and compared the number of possible MCS that could be at their origin. Notably, upon reverse translation, non-canonical MAPs presented a higher number of MCS compared to canonical MAPs (Figure 2A, Chapter 3). This was attributed to their amino acid composition, as they were richer in amino acids encoded by numerous codons (Figure 2B, Chapter 3). Notably, intronic and ERE MAPs were enriched in arginine (R), serine (S) and leucine (L), being amino acids encoded by 6 synonymous codons (Figure 2C, Chapter 3). These amino acids appear to be less frequent on average in the protein-coding sequences than those encoded by a smaller number of synonymous codons (Figure 8A, Chapter 3); suggesting that non-canonical MAPs would use more rare codons than canonical MAPs. In fact, we observed that the codons used in the MCS of non-canonical MAPs had a significantly lower genomic frequency than the MCS corresponding to canonical MAPs (Figure 2D, Chapter 3). This observation was consistent with previous observations where MAP source transcripts, whether canonical or non-canonical, appeared to use rare codons more frequently than non-MAP source transcripts[15]. However, our data suggest an inadvertent codon bias in non-canonical versus canonical MAP source transcripts. Although codon usage was not analyzed in our first study (**Chapter 1**) these findings suggest that rare codons may also have regulated the generation of the non-canonical MAP source proteins in DLBCL predicted to degrade rapidly.

Thus, these results lead us to hypothesize that use of rare codons for non-canonical proteins could lead to perturbations in their synthesis and thus to the generation of rapidly degradable proteins such as DRiPs[66]. For this reason, non-canonical proteins would have a preferential entry to proteasomal degradation and consequently a preferential access to the class I pathway. To confirm this hypothesis, it would be desirable to perform an in-depth study on codon composition of the canonical and non-canonical proteome detected in DLBCL cell lines.

This could be carried out to, firstly, consolidate the difference observed here with respect to amino acid enrichment between canonical and non-canonical proteins; but also, to assess the implication of codon usage in the biogenesis of non-canonical proteins and thus in the formation of DRiPs.

1.29.7Identification and Validation of Tumor Specific Antigens in DLBCL

Our ultimately goal was to identify and prioritize potential targets for immunotherapy in DLBCL samples. To do this, we used a series of searches with BamQuery to filter out MAPs that were not specific to tumors. First, we removed MAPs that were highly expressed in mTEC samples. Then, on the remaining MAPs, we searched for their expression in GTEx and eliminated those that were highly expressed in at least one normal tissue (excluding the testis). From the remaining MAPs, we retained only those that had at least a 5-fold change in expression between cancer (DLBCL samples from TCGA) and normal blood. Finally, we kept only those MAPs that overlapped with Ribo-seq elongation reads and labeled them as tumor-specific antigens. We applied this process to the 525 non-canonical MAPs identified in **Chapter 2** and labeled 61 of them as TSAs (data not shown). As expected, these 61 TSA candidates showed overlap with Ribo-seq elongation reads in their corresponding samples, as these peptides were identified using the Ribo-db approach, which utilizes Ribo-seq information to build customized databases.

While our Ribo-db approach was intended to identify non-canonical MAPs, it was not specifically designed to find tumor-specific antigens (TSAs) in samples. In fact, different approaches to create custom databases may not necessarily result in the identification of the same MAPs. Thus, to identify and prioritize potential targets for immunotherapy in DLBCL samples, we used a well-established method that leverages RNA-seq data to build customized, cancer-specific databases to identify TSAs[42]. This approach initially identified 6,869 MAPs in the three DLBCL samples, of which 67 were categorized as tumor-specific antigen candidates after applying BamQuery searches to filter out MAPs that were highly expressed in normal tissues (Figure 12A, Chapter 3). Our analysis showed that the BamQuery search on the DLBCL Ribo-seq elongation data was useful in identifying MAPs (6) that did not have any evidence of translation (as indicated by a Ribo-seq read count of 0). Therefore, the use of Ribo-seq data should be

pursued to provide additional evidence for the translation and presentation of such targets by the MHC I molecules. It is important to note that ribosome-protected fragments often result in short reads (typically 28-33 nucleotides in length) that may not fully overlap with the MCS. To account for this, we designed BamQuery to count partially overlapping reads that have a minimum overlap of 60% with the MCS at a given location. To accurately assess the total read count for a given location, we count reads according to the fraction of overlap with the MCS. However, it is important to note that the presence of Ribo-seq reads alone does not necessarily indicate the translation of a peptide. To improve the accuracy of BamQuery analyzing Ribo-seq data, we could consider evaluating the 3-nucleotide (3-nt) periodicity of overlapping MAP reads. The 3-nt periodicity in Ribo-seq data depicts the ORF of translation. The periodicity therefore refers to the repeating pattern in the reads that occurs every three nucleotides, as ribosomes move along the mRNA molecule at a rate of one codon (a sequence of three nucleotides) per time unit[67]. Thus, by determining the ORF of the Ribo-seq reads, only those that match with the MAP's ORF should count. To obtain this information, we could use methods such as RibORF[68] or PROTEOFORMER[69] to analyze the Ribo-seq data and reveal the 3-nt periodicity for each length of the Ribo-seq reads . We could then evaluate the Ribo-seq reads overlapping the MCS and count only those that their periodicity coincides with the MAP.

Altogether, BamQuery is a versatile tool that can be used to identify potential candidates for cancer vaccine treatment from various sequencing data sources, including bulk, single-cell, and Ribo-seq data from both human and mouse genomes. To the best of our knowledge, BamQuery is the first tool to be particularly useful for analyzing MAPs expression in RNA-seq data. Analysis of single-cell RNA data with BamQuery provides information on tumor heterogeneity and microenvironment that could affect MAP expression. By implementing improvements to the tool, such as evaluating the 3-nt periodicity of overlapping MAP reads from Ribo-seq, we can increase the confidence in the identification of valid targets for vaccine treatment with BamQuery. Overall, BamQuery represents a valuable resource for researchers and clinicians working on the development of cancer vaccines.

1.29.8Conclusions

Through this thesis, we aimed to investigate the extent to which non-canonical proteins contribute to the proteome in cancer cells, and to develop a method for identifying and prioritizing non-canonical MAPs as potential targets for cancer immunotherapy. Our results showed that non-canonical proteins contribute independently to the immunopeptidome or overall proteome, and have distinct characteristics compared to canonical proteins. Therefore, non-canonical MAPs may be promising targets for immunotherapy, but their expression in normal tissues should be carefully validated before they can be prioritized for cancer vaccine design.

In our immunopeptidome study of DLBCL samples, we found that about 10% of the MAPs were derived from non-canonical proteins. Half of these non-canonical MAP-source proteins were labeled as novel isoforms of canonical proteins, while the other half was labeled as cryptic proteins. Our analysis showed that cryptic MAP-source proteins tended to be less abundant and shorter but were about 5 times more efficient at generating MAPs compared to canonical proteins. These cryptic MAP-source proteins appeared to be highly disordered and unstable, which makes them prototypical DRiPs and gives them preferential access to the MHC I pathway. We also observed low overlap between the non-canonical proteins detected in the immunopeptidome and those detected in the whole proteome, suggesting the existence of two distinct protein repertoires. Additionally, we found that the expression of non-canonical proteins might be favored by DLBCL neoplastic transformation process, which is characterized by cytogenetic abnormalities. At the same time, we found that the expression of these non-canonical proteins may serve as a regulatory mechanism that prevents the correct translation of canonical proteins. This dual role of non-canonical protein expression may contribute to the oncogenic program in DLBCL cells.

To further evaluate non-canonical MAPs as potential targets for vaccine design, we developed a tool called BamQuery to assess their RNA-seq expression in various tissues. BamQuery was designed to comprehensively collect RNA-seq expression for any MAP by considering all possible genomic regions that could be the source of the MAP, using this information as a proxy for MAP presentation. Using BamQuery, we found that non-canonical MAPs can be translated from rarer codons than canonical ones, suggesting that their source

proteins may have a deficient 3D conformation that gives them DRiPs appearance. We also discovered that previously published TAs were highly expressed in healthy tissues, which would make them poor targets for immunotherapy. Finally, we demonstrated the utility of BamQuery in identifying potential safe immunotherapeutic targets in DLBCL that are derived from non-canonical translations.

Overall, our findings suggest that non-canonical regions of the genome broaden the range of MHC I peptides that can be presented to T cells. However, it is important to carefully evaluate the expression of these non-canonical MHC I peptides in healthy tissues to determine their suitability as targets for immunotherapies. We have developed BamQuery to facilitate this process by quantifying the RNA expression of non-canonical MHC I peptides and using it as a proxy for their MHC I presentation. Thus, BamQuery can be used as the first step in evaluating the safety of targets for immunotherapy.

1.30References

1. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
2. Li, H. et al. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **17**, 1031 (2016).
3. Blakeley, P., Overton, I.M. & Hubbard, S.J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**, 5221-5234 (2012).
4. Livyatan, I. et al. Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res* **41**, 6300-6315 (2013).
5. Apcher, S., Daskalogianni, C. & Fahraeus, R. Pioneer translation products as an alternative source for MHC-I antigenic peptides. *Mol Immunol* **68**, 68-71 (2015).
6. 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* **110**, 17951-17956 (2013).
7. Erhard, F. et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**, 363–366 (2018).
8. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
9. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
10. Degner, J.F. et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-3212 (2009).
11. Kumar, S., Vo, A.D., Qin, F. & Li, H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep* **6**, 21597 (2016).
12. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
13. Starck, S.R. et al. Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351**, aad3867 (2016).

14. Brunet, M.A., Levesque, S.A., Hunting, D.J., Cohen, A.A. & Roucou, X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res* **28**, 609-624 (2018).
15. Laumont, C.M. et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun* **7**, 10238 (2016).
16. Lu, S. et al. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Research* **47**, 8111-8125 (2019).
17. Ouspenskaia, T. et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol* (2021).
18. van Heesch, S. et al. The Translational Landscape of the Human Heart. *Cell* **178**, 242-260 e229 (2019).
19. Pearson, H. et al. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest* **126**, 4690-4701 (2016).
20. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics* **14**, 658-673 (2015).
21. Lubec, G. & Afjehi-Sadat, L. Limitations and pitfalls in protein identification by mass spectrometry. *Chem Rev* **107**, 3568-3584 (2007).
22. Blaha, D.T. et al. High-Throughput Stability Screening of Neoantigen/HLA Complexes Improves Immunogenicity Predictions. *Cancer Immunol Res* **7**, 50-61 (2019).
23. Prevosto, C. et al. Allele-Independent Turnover of Human Leukocyte Antigen (HLA) Class Ia Molecules. *PLoS One* **11**, e0161011 (2016).
24. Tran, N.H. et al. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* **16**, 63-66 (2019).
25. Giansanti, P., Tsiatsiani, L., Low, T.Y. & Heck, A.J. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc* **11**, 993-1006 (2016).
26. Gong, X., Xiong, X., Zhao, Y., Ye, S. & Fang, X. Boosting the Signal Intensity of Nanoelectrospray Ionization by Using a Polarity-Reversing High-Voltage Strategy. *Anal Chem* **89**, 7009-7016 (2017).

27. Pfammatter, S. et al. A Novel Differential Ion Mobility Device Expands the Depth of Proteome Coverage and the Sensitivity of Multiplex Proteomic Measurements. *Mol Cell Proteomics* **17**, 2051-2067 (2018).
28. Pfammatter, S. et al. Extending the Comprehensiveness of Immunopeptidome Analyses Using Isobaric Peptide Labeling. *Anal Chem* **92**, 9194-9204 (2020).
29. Klaeger, S. et al. Optimized Liquid and Gas Phase Fractionation Increases HLA-Peptidome Coverage for Primary Cell and Tissue Samples. *Mol Cell Proteomics* **20**, 100133 (2021).
30. Wilhelm, M. et al. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat Commun* **12**, 3346 (2021).
31. 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* **111**, 4507-4512 (2014).
32. Zarling, A.L. et al. Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *Proc Natl Acad Sci U S A* **103**, 14889-14894 (2006).
33. Zhang, J. et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* **11**, M111 010587 (2012).
34. Solleder, M. et al. Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands. *Mol Cell Proteomics* **19**, 390-404 (2020).
35. Abelin, J.G. et al. Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry. *Nat Protoc* **10**, 1308-1318 (2015).
36. Abelin, J.G. et al. Workflow enabling deepscale immunopeptidome, proteome, ubiquitylome, phosphoproteome, and acetylome analyses of sample-limited tissues. *Nat Commun* **14**, 1851 (2023).
37. Engelhard, V.H. Structure of peptides associated with class I and class II MHC molecules. *Annu Rev Immunol* **12**, 181-207 (1994).
38. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140-1146 (2020).
39. Samandi, S. et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife* **6**, e27860 (2017).

40. Chong, C. et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* **11**, 1293 (2020).
41. Erhard, F., Dolken, L., Schilling, B. & Schlosser, A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunol Res* **8**, 1018-1026 (2020).
42. Laumont, C.M. et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med* **10**, eaau5516 (2018).
43. Nesvizhskii, A.I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114-1125 (2014).
44. Vick, E.J. et al. Age-Related Chromosomal Aberrations in Patients with Diffuse Large B-Cell Lymphoma: An In Silico Approach. *World J Oncol* **9**, 97-103 (2018).
45. Chan, W.Y., Wong, N., Chan, A.B., Chow, J.H. & Lee, J.C. Consistent copy number gain in chromosome 12 in primary diffuse large cell lymphomas of the stomach. *Am J Pathol* **152**, 11-16 (1998).
46. Yuan, X. et al. Notch signaling: an emerging therapeutic target for cancer treatment. *Cancer Lett* **369**, 20-27 (2015).
47. Guruharsha, K.G., Kankel, M.W. & Artavanis-Tsakonas, S. The Notch signalling system: recent insights into the complexity of a conserved pathway. *Nat Rev Genet* **13**, 654-666 (2012).
48. Young, S.K. & Wek, R.C. Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *J Biol Chem* **291**, 16927-16935 (2016).
49. Ehx, G. et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity* **54**, 737-752 e710 (2021).
50. Kuznetsov, A., Voronina, A., Govorun, V. & Arapidi, G. Critical Review of Existing MHC I Immunopeptidome Isolation Methods. *Molecules* **25** (2020).
51. Marcu, A. et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer* **9** (2021).

52. Rattay, K., Meyer, H.V., Herrmann, C., Brors, B. & Kyewski, B. Evolutionary conserved gene co-expression drives generation of self-antigen diversity in medullary thymic epithelial cells. *J Autoimmun* **67**, 65-75 (2016).
53. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
54. Ballouz, S., Dobin, A., Gingeras, T.R. & Gillis, J. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Res* **46**, 5125-5138 (2018).
55. Mudge, J.M. et al. Standardized annotation of translated open reading frames. *Nat Biotechnol* **40**, 994-999 (2022).
56. Wu, Z. et al. Proteogenomics and Differential Ion Mobility Enable the Exploration of the Mutational Landscape in Colon Cancer Cells. *Anal Chem* **94**, 12086-12094 (2022).
57. Do, C.B. & Batzoglou, S. What is the expectation maximization algorithm? *Nature Biotechnology* **26**, 897-899 (2008).
58. Bonaventura, P. et al. Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy. *Sci Adv* **8**, eabj3671 (2022).
59. Rivero-Hinojosa, S. et al. Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat Commun* **12**, 6689 (2021).
60. Hirama, T. et al. Proteogenomic identification of an immunogenic HLA class I neoantigen in mismatch repair-deficient colorectal cancer tissue. *JCI Insight* **6** (2021).
61. Choi, J. et al. Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Res* **47**, D780-D785 (2019).
62. Silvin, A. et al. Constitutive resistance to viral infection in human CD141(+) dendritic cells. *Sci Immunol* **2** (2017).
63. Feder-Mengus, C., Ghosh, S., Reschner, A., Martin, I. & Spagnoli, G.C. New dimensions in tumor immunology: what does 3D culture reveal? *Trends Mol Med* **14**, 333-340 (2008).
64. Grassi, L. et al. Organoids as a new model for improving regenerative medicine and cancer personalized therapy in renal diseases. *Cell Death Dis* **10**, 201 (2019).

65. Dijkstra, K.K. et al. Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. *Cell* **174**, 1586-1598 e1512 (2018).
66. Yewdell, J.W. & Holly, J. DRiPs get molecular. *Curr Opin Immunol* **64**, 130-136 (2020).
67. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
68. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4**, e08890 (2015).
69. Verbruggen, S. et al. PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-assisted Proteogenomic Hunt for New Proteoforms. *Mol Cell Proteomics* **18**, S126-S140 (2019).